DEVELOPMENT OF A CROSS-CULTURAL HAZARD PERCEPTION TEST

PETYA VENTSISLAVOVA PETROVA

A thesis submitted in partial fulfilment of the requirements
of Nottingham Trent University for the degree of

Doctor of Philosophy

January 2019

Abstract

This thesis is concerned with the development of a culturally-agnostic version of the UK hazard perception (HP) test, with the ultimate hope of providing a road safety assessment tool that can be exported to different countries whilst retaining its diagnostic validity. This aim was inspired by the United Nations Decade of Action on Road Safety (2010-2020), which was instigated to reduce the global burden of road deaths.

The first three experiments explored the various design elements of a future global blueprint for hazard perception testing. Unfortunately, none of these studies revealed a difference between UK experienced and novice drivers, considered important for test validity. When the test was compared across three countries (using clips and participants from China, Spain and the UK; experiment 4), the experience difference was still elusive, though cultural differences were evident.

To address the lack of validity, a new HP test-variant was developed: a hazard prediction test. This test presents the same driving clips to viewers, but the screen is occluded at the point of hazard onset and participants are asked "What happens next?". In a UK sample, experiment 5 finally showed an experiential difference. When taken back to China and Spain (experiment 6), the overall experiential difference remained, while cultural differences were ameliorated.

Following further development of the hazard prediction protocol (experiment 7), the test was applied to a brand new cultural context in Israel. Once again, the hazard prediction test was successful in differentiating between safe and less-safe drivers on the basis of experience.

In conclusion, the hazard prediction test appears to be a more robust methodology for international export, ostensibly reducing problems of criterion bias, subjective judgements on scoring windows, and language difficulties in explaining what a hazard is. The final chapter summarises this blueprint for a culturally-agnostic hazard test, and recommends the protocol be adopted globally.

Acknowledgements

First, I would like to thank my supervisor Prof. David Crundall. I have been extremely privileged to have you as a supervisor. Your motivation and passion for Transport Psychology, and research in general, are tremendously contagious. Thank you for giving me the confidence and freedom to develop my own ideas. Thank you for sharing with me your valuable knowledge and for being so patient every time I did not follow your suggestions (even though I was wrong most of the time). Thank you for joining me in this cross-cultural adventure and for your enthusiasm every time I appeared on your door with some crazy idea. Not many PhDs involve travelling the world! Thank you for our research discussions and those about everyday life over coffee (and beer). But most of all, thank you for your friendship (and for staying with me almost the whole night in the hospital in Montreal, which involved missing the conference reception drinks).

I would also like to thank Dr. Candi Castro. I somehow convinced her back in 2013 that I was the right person for the job. Thank you for believing in me and for your support during all these years. Thank you for introducing me into the exciting world of Transport Psychology and for giving me the opportunity to develop myself as a researcher. Thank you for being by my side during this journey (my time in the University of Granada was amazing) and for your friendship!

Special thanks to Prof. Frank McKenna and Dr. Damian Poulter for agreeing to be my PhD examiners. I appreciate it with all my heart!

I would like to thank Nottingham Trent University for awarding me with this PhD scholarship and Santander Bank for funding my cross-cultural studies.

The cross-cultural studies would have not been possible without the help of both the Spanish (University of Granada) and Chinese (Tsinghua University) research teams. I would like to thank Dr. Yutao Ba and Prof. Wei Zhang for making my research visit to Beijing such a wonderful experience. Special thanks to everybody in Prof. Zhang's research team for helping with data collection and for accepting me as one of them during my time in Beijing.

Thanks to Dr. Candi Castro and Dr. Pedro Garcia-Fernandez for helping with data collection and filming the Spanish driving footage for the cross-cultural study. Also, thanks to my friend and colleague Andres Gugliotta. We started this "transport" journey together back in 2013. I still remember coding "What happens next" videos until 3 am in the morning. I will never forget our time in Granada (thanks for our endless conversations on the phone and for being there when I most needed it).

I would also like to thank Prof. Tova Rosenbloom (Bar Ilan Univerity) who hosted me in Tel Aviv while filming the Israeli footage. Thank you for the wonderful Sabbath with you and your family.

I would also like to thank my two co-supervisors Alex Sumich and Antonio Castro.

I am especially grateful to my colleagues in Nottingham Trent University:
-Andrew Mackenzie for his continuous support during all these years. Without your valuable knowledge in eye movements, Experiment 1 would not have been possible.
-Angie Young for her support not only with my PhD but also during my first months of lectureship. Always there to listen and help!
-Vikki Kroll for her support and for sharing with me one of the most amazing research trips to Australia.
-Thom Baguley for introducing me into the exciting world of multilevel modelling and for assisting with the analyses for the published version of the cross-cultural study.
-Our technicians Roy Barson, Ben Thompson and Josh Baker (who picked up the phone at 11pm to answer my E-prime questions) for recovering my data not once, not twice but three times (I know…). I don't know what I would have done without you!

My biggest thanks goes to my family. I would not be where I am today without my mother Fidanka, my grandparents Vasil and Liliana, my aunt Biliana and my cousins. And without my brother Enzo (if you read this when you are a bit older, please follow your dreams). They have always believed in me no matter what. I just cannot say thank

you enough for being such an example and for being such a lovely family. I am extremely blessed to have you in my life!

Finally, thank you V.V.! You have always been the motivating force behind all this!

During my interview for the PhD scholarship, I was asked what phrase or piece of research has most influenced or inspired my work.

I answered:

“What Happens Next?” …

To my Grandparents – Vasil Neshev and Liliana Nesheva

Contents

Abstract **3**

Acknowledgments **4**

Contents **8**

List of Tables **17**

List of Figures **18**

1 Chapter 1: General introduction **23**

1.1 Chapter overview 23

1.2 Road traffic injuries are a global health problem 23

1.3 The United Kingdom 27

1.4 The hazard perception test 29

1.4.1 What is hazard perception? 29

1.4.2 What is the official UK hazard perception test? 30

1.4.3 The history of the hazard perception test 33

1.4.4 Why is the evidence mixed? 38

1.4.4.1 Problems with using simple response times 38

1.4.4.2 Differences in response mode 39

1.4.4.3 Differences in analysis 40

1.4.4.4 Differences in stimuli 41

1.4.4.5 Criterion bias 41

1.4.4.6 Common vocabulary and the lack of theory underlying HP 42

1.5 A new framework 43

1.6 Conclusion 49

1.7 Outline of this thesis 51

2 Chapter 2: Methodology **53**

2.1 Driving research 53

2.1.1 Methods used in this thesis 54

2.2 Video-based studies 54

2.1.1 Hazard Perception clips 55

2.3 Filming of hazard perception clips 59
2.3.1 Camera placement 59
2.3.2 Filming UK clips 59
2.3.3 Filming Chinese clips 61
2.3.4 Filming Spanish clips 61
2.3.5 Filming Israeli clips 62
2.3.6 Which clips were used in which studies 63
2.3.7 Editing, synchronisation and clip selection 63
2.4 Apparatus 66
2.4.1 E-prime 2.0 67
2.4.2 SMI 500Hz, Eye-tracking equipment 67
2.4.3 Qualtrics 69
2.5 Forward-Backward translation 70
2.6 Instructions 71
2.7 Ethics 72
3 Chapter 3: DESIGN OF THE HAZARD PERCEPTION VIDEO CLIPS 81
Experiment 1
Does adding mirror information to the traditional hazard perception clips improve driving performance?
3.1 Introduction 81
3.1.1 The impact of mirror information 83
3.1.2 The impact of adding a car interior 85
3.1.3 The role of eye movements in hazard perception 85
3.1.4 Overtaking and undertaking hazards 87
3.1.5 Overview of the study 87
3.2 Method 88
3.2.1 Participants 88
3.2.2 Design 89
3.2.3 Materials 91
3.2.4 Procedure 91

3.3 Results 94
3.3.1 Eye movements measures 94
3.3.1.1 Fixations on hazards 94
3.3.1.2 Fixations on mirrors 95
3.3.1.3 Spread of search along the horizontal and vertical meridians 96
3.3.1.4 Overtaking hazards and mirrors 96
3.3.2 Behavioural responses to hazards 97
3.3.2.1 Response time to predefined hazards 97
3.3.2.2 Accuracy 97
3.3.2.3 Extra responses to hazards 98
3.3.2.4 Response times after first fixation 98
3.3.2.5 Response times to overtaking hazards 98
3.4 Discussion 100
4 Chapter 4: DESIGN OF THE HAZARD PERCEPTION CLIPS 104
Experiments 2 and 3
Does the design of the car overlay impact hazard perception performance? Comparing transparent vs. opaque overlay
4.3 Summary of previous results 104
4.2 Introduction 106
4.2.1 The importance of "see-through" design 106
4.2.2 Overview of the study 109
4.3 Opaque vs. Transparent car layout: hazardous and immersion ratings (Experiment 2) 109
4.3.1 Method 109
4.3.1.1 Participants 109
4.3.1.2 Design and materials 110
4.3.1.3 Procedure 110
4.3.2 Results and discussion 110
4.3.2.1 Hazard ratings 111
4.3.2.2 How realistic? 111

4.3.2.3 How close to real driving? 111
4.3.2.4 Preferences 112
4.4 Opaque vs. Transparent car layout: Behavioural measures (Experiment 3) 113
4.4.1 Method 113
4.4.1.1 Participants 113
4.4.1.2 Design 113
4.4.1.3 Materials and apparatus 114
4.4.1.4 Procedure 114
4.4.2 Results 116
4.4.2.1 Eye movements measures 116
4.4.2.1.1 Fixations on hazards 117
4.4.2.1.2 Fixations on mirrors and frames 117
4.4.2.1.3 Spread of search along the horizontal and vertical meridians 120
4.4.2.1.4 Precursor analysis 121
4.4.2.2 Behavioural responses to hazards 121
4.4.2.2.1 Response time to predefined hazards 121
4.4.2.2.2 Accuracy 122
4.4.2.2.3 Extra responses to hazards 122
4.4.2.3 Ratings 122
4.5 General discussion 123
5 Chapter 5: AN INTERNATIONAL PERSPECTIVE ON HAZARD PERCEPTION 127
Experiment 4
Testing the Hazard perception test in three different countries: China, Spain and the UK
5.1 Summary of previous findings 127
5.2 Introduction 129
5.2.1 An international perspective on hazard perception testing 129
5.2.2 Are hazards culturally specific? 130
5.2.3 Comparison of HP performance between UK and Malaysian drivers 131
5.2.4 Overview of the study 132

5.3 Method 134
5.3.1 Participants 134
5.3.2 Materials and apparatus 135
5.3.3 Design 140
5.3.4 Procedure 141
5.4 Results 141
5.4.1 Response accuracy to hazards 142
5.4.2 Response times to hazards 143
5.4.3 Extra hazard responses 145
5.4.4 Hazardousness ratings for all hazard clips 147
5.5 Discussion 149
6 Chapter 6: THE HAZARD PREDICTION TEST 157
Experiment 5
Free-response hazard prediction test: occlusion points
6.1 Summary of previous findings 157
6.2 Introduction 159
6.2.1 The hazard prediction test 159
6.2.2 Overview of the study 161
6.3 Method 161
6.3.1 Participants 161
6.3.2 Design 162
6.3.3 Materials 165
6.3.4 Procedure 165
6.4 Results 167
6.5 Discussion 169
7 Chapter 7: CROSS-CULTURAL HAZARD PREDICTION 172
Experiment 6
Testing the Hazard Prediction test in three different countries: China, Spain and the UK
7.1 Summary of previous findings 172

7.2 Introduction 173

7.2.1 Advantages of the hazard prediction test 175

7.2.2 Overview of the study 176

7.3 Method 176

7.3.1 Participants 176

7.3.2 Materials and apparatus 178

7.3.3 Design and procedure 180

7.4 Results 181

7.4.1 Accuracy 181

7.4.2 Hazardousness ratings for hazard prediction 183

7.5 Discussion 184

7.6 Comparison of the two tests 185

7.7 General discussion 188

7.7.1 Do different countries produce different hazards? 189

7.7.2 The limitations of hazard prediction tests 190

8 Chapter 8: REFINING THE HAZARD PREDICTION TEST 193

Experiment 7

Validation of a multiple-choice hazard prediction test

8.1 Summary of previous findings 193

8.2 Introduction 194

8.3 Method 197

8.3.1 Participants 197

8.3.2 Design, materials and procedure 197

8.4 Results 200

8.5 Comparison of the multiple-choice test with the free-response test 202

8.6 Discussion 204

9 Chapter 9: THE HAZARD PREDICTION TEST 208

Experiment 8

The Hazard Prediction test applied to the Israeli driving context

9.1 Summary of previous findings 208

9.2 Introduction 209
9.2.1 Features of the hazard prediction test and evidence of its effectiveness 209
9.2.2 Guidelines followed to create the test 211
9.2.2.1 Mirrors 211
9.2.2.2 Car overlay 212
9.2.2.3 Non-hazardous trials 212
9.2.2.4 Occlusion points 214
9.2.2.5 Multiple-choice format 214
9.2.3 Israel 214
9.2.4 Hazard Perception in Israel 216
9.2.5 Overview of the study 218
9.3 Method 219
9.3.1 Participants 219
9.3.2 Design, apparatus and materials 219
9.3.3 Test development 220
9.4 Results 224
9.4.1 Is there a hazard? 224
9.4.2 Accuracy on the Multiple-Choice Prediction Question 226
9.5 Discussion 227
10 Chapter 10: GENERAL DISCUSSION 233
10.1 Brief overview of the thesis 233
10.2 Summary of the main finding of each study 235
10.2.1 Experiment 1: Design of the hazard perception video clips. Does adding mirror information to the traditional hazard perception clips improve driving performance? 235
10.2.2 Experiments 2 and 3: Does the design of the car overlay impact hazard perception performance? Comparing transparent vs. opaque overlays 238
10.2.3 Experiment 4: Testing the Hazard perception test in three different countries: China, Spain and the UK 241

10.2.4 Experiment 5: Free-response hazard prediction test: occlusion points 243
10.2.5 Experiment 6: Testing the Hazard Prediction test in three different countries: China, Spain and the UK 245
10.2.6 Experiment 7: Validation of a multiple-choice hazard prediction test 248
10.2.7 Experiment 8: The Hazard Prediction test applied to the Israeli driving context 249
10.3 Overview of the results 252
10.3.1 Areas in which this thesis was successful 252
10.3.2 Areas that need further research 254
10.4 Theoretical implications 256
10.5 Practical implications 261
10.5.1 International perspective 263
10.6 The challenges of cross-cultural research 264
10.6.1 How does the differences in traffic culture impact driving behaviour? 264
10.6.2 Validation of a methodology instead of a theoretical construct 266
10.6.3 The challenges of cultural equivalence 268
10.6.3.1 Content equivalence 268
10.6.3.2 Semantic equivalence 269
10.6.3.3 Technical equivalence and conceptual equivalence 270
10.6.3.4 Criterion equivalence 270
10.6.3.5 Sample equivalence 271
10.7 Limitations 273
10.7.1 Stimuli design 273
10.7.2 Sampling limitations 274
10.3.7 Qualtrics software 275
10.8 Reflections on the research process 276
10.8.1 Was the hazard perception test a straw-man hypothesis? 277
10.8.2 Alternative explanations for the null effects associated with the hazard perception methodology 278

10.8.3 Does HP practice mean that learners are better at HP tests than experienced drivers? 280
10.8.4 Size effect and marginal differences 281
10.8.5 Contribution to theory 282
10.9 Conclusion 284
References **285**
Appendix A **312**
Appendix B **317**
Appendix C **319**
Appendix D **338**
Appendix E **348**

List of tables

1.1 Definitions suggested by Pradhan and Crundall (2017) for the processes involved in hazard avoidance (taken from Pradhan and Crundall, 2017) 48

2.1 Description of the UK clips 74

2.2 Description of the Chinese clips 76

2.3 Description of the Spanish clips 78

2.4 Description of the Israeli clips 79

4.1 Counterbalancing order across groups 115

4.2 Means of the ratings (from 1 to 7) for each question: How hazardous do you think that particular clips was? (Q1), How close to real driving did that clip look? (Q2), How much did you feel 'you were there' in that clip?(Q3) across the experience groups 123

5.1 Mean demographic values for participants 135

5.2 A description of the a priori hazards selected within each clip 138

6.1 A description of the a priori hazards selected within each clip 166

7.1 Mean demographic values for participants 177

7.2 A description of the occlusion points of the a priori hazards selected within each clip 179

8.1 Mean age, mileage and driving experience for the different driving groups 197

8.2 Mean mileage, driving experience and mean difference between experience and mileage for the matched groups 202

8.3 Pearson correlations for accuracy vs. years of driving and accuracy vs. miles for both types of tests (multiple-choice and free-response type) 203

9.1 A description of the a priori hazardous and non-hazardous traffic situations selected for each clip 222

List of figures

1.1 Percentage of road traffic fatality per 100000 population within the main continental areas 25

1.2 World map illustrating the zones that are most and least affected by road deaths per 100000 population. The areas coloured in grey represent countries with the lowest death rate and those coloured in red are the countries with the highest death rate 25

1.3 Fatalities in reported road accidents in the United Kingdom between 2006 and 2016 (taken from DfT Annual Report, 2017) 27

1.4 Graphic representation of a typical time frame of hazard onset and offset. The two top images represent the start of the time window (onset) and bottom right image the end (offset) 31

1.5 A Computer-Generated Imagery (CGI) clip representing typical hazardous situation where a van is approaching from the left, intending to join the lane of the driving car. The yellow circle identifies the hazard for the reader and does not appear in the actual video 32

1.6 The images represent stills from a video along a timeline (the black arrow). The red arrow represents a response that was made to subtle cues that a participant noticed, suggesting that a pedestrian might step out. This falls before the scoring window opens and therefore is counted, unfairly, as a 'miss'. The first green arrow reflects a different participant's concern about the car ahead. This is not the true hazard, yet due to the response occurring in the scoring window, it is counted as a very quick 'hit'. The later green arrow represents a true response to the hazard. Unfortunately there is no way to distinguish between the validity of these two responses, and the participant receives the score associated with the first green arrow rather than the second 39

2.1 Example images of different hazardous video clips developed by several research groups during the last two decades ((a) to (e)) and the DVSA driving test (f) 56

2.2 The graphic represents the positions of the 4 cameras. The green arrows show the placements of the front and rear view camera and the blue arrows the positions of the left and right cameras. The side mirror graphs the amount of sky, road and car side captured by the cameras 60

2.3 HD cameras employed for filming of the driving footage. On figure 2.3a is represented the GO PRO Hero (front footage) and figure 2.3b represents JVC action camera (left, right and rear view footage) 62

2.4 Screenshots of the Chinese (a), Spanish (b), UK (c and d) and Israeli (e) hazardous videos. Images 9 (d) contains an example of opaque graphic overlay applied for experiments 2 and 3 64

2.5 The two screen-shots depict the areas of interest for a typical hazard (top image), and left, right and rear-view mirrors and car frame (bottom image) 69

3.1 Three screenshots of the video clips representing the three different conditions of the experiment. The top panel (3.1a) depicts the video showing only the forward view of the driving scene, the middle panel (3.1b) represents the forward view of the same video, however with a graphic overlay and the bottom panel (3.1c) shows the clip including mirror information 93

3.2 Percentages of time to first fixate the hazard across experience and experimental conditions (with standard error bars added) 95

3.3a Frequency of clicks across conditions for video 2 (FV = forward view, CI = car interior, WM = with mirrors) 99

3.3b Frequency of clicks across conditions for video 6 (FV = forward view, CI = car interior, WM = with mirrors) 99

3.3c Frequency of clicks across conditions for video 11 (FV = forward view, CI = car interior, WM = with mirrors) 100

4.1 The top image represents a clip that contains a transparent car overlay and the bottom image the same clip with opaque car overlay 116

4.2 Total time spend looking at the left, right and back mirror across experience (with standard error bars added) 119

4.3 Percentages of dwells on frames and dashboard across both conditions and experience (with standard error bars added) 119

4.4 Vertical scanning of the driving environment across experience (experienced vs. novices) and conditions (transparent vs. opaque) (with standard error bars added) 120

5.1 Three screen shots taken from hazard perception clips filmed in China (top panel), Spain (middle panel) and the UK (bottom panel) 137

5.2 The percentage of hazards correctly responded to by all participant groups across the three countries of origin for the clips (with standard error bars added) 143

5.3 The mean response time of all three driver groups to hazards across the 3 clips sets recorded in each country (reconverted from square-root and z-score transformed means, with standard error bars added) 144

5.4 The mean number of extra hazard responses per minute for each block of clips across the three groups of participants (with standard error bars added) 146

5.5 The mean number of extra hazard responses per minute for each block of clips across experienced and inexperienced drivers (with standard error bars added) 147

5.6 The hazardousness ratings of participants across three nationalities and the clip sets from each country. Ratings were given on a 1-7 scale, where higher ratings indicate greater perceived hazardousness (standard error bars added) 148

5.7 a RTs of the Chinese participants across all 10 Chinese clips 153

5.7 b RTs of the Spanish participants across all 10 Spanish clips 153

5.7 c RTs of the UK participants across all 10 UK clips 154

6.1 Three frames illustrating the three occlusion points for one clip. While driving along a road, the car in front pulls over. This requires the film car to overtake the stopping vehicle. However, a safe driver should be aware that there is a vehicle visible in the right side mirror that might pose a problem (top panel; distal occlusion). As the clip progresses, the car in the side mirror disappears, as it moves into the blind spot ready to overtake. Thus, the evidence that this vehicle is going to pose a problem increases considerably (middle panel; intermediate occlusion). Finally, the overtaking car becomes visible in the forward view confirming that the car from behind is indeed overtaking and poses a hazard for the film car’s intended action (bottom panel; proximal occlusion) 164

6.2 Percentage of correct responses across all 3 occlusion points and experienced groups (with standard error bars added) 168

7.1 The mean percentage of accuracy of predicting the hazards of all three driver groups across the 3 clips sets recorded in each country (with standard error bars added) 182

7.2 Hazardousness ratings for the hazard prediction test for the experience groups across clip origin (with standard error bars) 184

7.3 The mean percentage of the hazards correctly identified/predicted for both hazard perception and hazard prediction tests across driving experience (with standard error bars) 187

7.4 The mean percentage of hit rate for the hazard perception and the hazard prediction test across driver’s nationality (with standard error bars) 187

7.5 The mean percentage of hit rate for the hazard perception and the hazard prediction test across clip type (with standard error bars) 188

8.1 Three panels represent the end of a clip: the first panel is the final frame (where the impending hazard is an oncoming white van, hidden by parked vehicles). This is followed by an occlusion panel, before four multiple-choice options are displayed on screen 200

8.2 Percentage accuracy (white bars) and ratings (from 1-7; black bars) across all experienced groups: learners, novices, experienced drivers with less than 6 years of driving experience, and experienced drivers with more than 6 years of driving experience. Standard error bars are added 201

8.3 Percentage accuracy of correct responses across experience groups and test type (with standard error bars added) 203

9.1 Percentage of hits for the questions “Is there a hazard?” across clip type with standard error bars 225

9.2 Percentage of prediction accuracy across driving groups with standard error bars. The dashed line represents the mean chance expectancy 226

10.1 a Sub-processes comprising the hazard avoidance process. The processes involved in the hazard perception process are highlighted in red (hazard fixation, processing and appraisal) 260

10.1 b Sub-processes comprising the hazard avoidance process. The processes involved in the hazard prediction process are highlighted in blue (hazard searching and salience, precursor prioritisation) 261

Chapter 1

GENERAL INTRODUCTUION

1.1 Chapter overview

This thesis is concerned with the development of a culturally-agnostic version of the UK hazard perception test, with the ultimate hope of providing a road safety tool that can be exported to different countries whilst retaining its diagnostic validity. This chapter will begin by summarising the urgent need for global road safety initiatives, and will accept the United Nations' suggestion that countries with poor safety records should look to adapt and adopt initiatives from safer countries. However, it will be argued that the safest countries cannot be passive partners in this transference of safety interventions across geographical borders, but should be active in assessing the fitness for export of their initiatives. The chapter will detail one of the most successful initiatives in the UK over the last 15 years: hazard perception testing. Evidence for and against the diagnostic validity of this test will be reviewed, before discussing whether the test is likely to work equally well in all countries.

1.2 Road traffic injuries are a global health problem

Road traffic injuries are responsible for nearly 1.3 million deaths per year worldwide and are the primary cause of death for young people (and in the top three causes of death for people between five and 44 years; WHO, 2015). This is equivalent to one person being killed as the result of a traffic collision every 25 seconds. Beyond the tragedy of lives lost, the consequences of road collisions can be economically devastating. With 50 million people injured per year and potentially disabled for life, taxable income is replaced with health costs. Specifically, for low- and middle- income countries, costs reach around $100 billion per year (WHO, 2013). In order to address this health problem, the UN General Assembly proclaimed the Decade of Action for Road Safety 2011-2020. The Plan of Action proposed by the UN was directed at a broad audience and aimed to support the development of national and local plans of action. The Global Plan includes various initiatives, such as designing safer roads, improving the safety features of vehicles and enforcing internationally-harmonised laws in order to prevent road traffic injuries. The plan also recommended extracting best practice from countries with the best road-safety statistics, such as the UK, where evidence-based policies have contributed to a 45% reduction in fatalities between 2006 and 2015 (UK Department for Transport, 2016). Disparities across countries in terms of collision statistics are apparent even within Europe. A closer look into the European data shows that South East Europe recorded higher-than-average fatality rates in comparison with North West Europe. The European Transport Safety Council estimated that the road deaths per million inhabitants for 2016 differed considerably between the South East and North West of Europe. The countries placed at the top of the table are Sweden, Norway and the UK with less than 30 deaths per million inhabitants, while Bulgaria and Romania still remain at the opposite end of the continuum with more than 80 deaths per million inhabitants.

Some of the South Eastern European countries face significant challenges in implementing the proposed strategies and targets of the UN. Even greater hurdles face the developing countries, especially those with low income (typically countries in Asia, Africa and South America), where there are no measurable safety targets or established national Road Safety Strategy (WHO, 2010). Low-income countries represent the highest accident rates globally (24.1%), followed by the middle-income (18.4%) and high-income countries (9.2%) (see Figure 1.1). Even though low- and middle-income countries have only half of the world's vehicles, they account 90% for the world's road traffic deaths. Lamentably, this suggests that the chance of dying in a road traffic crash depends on where you live (see Figure 1.2).

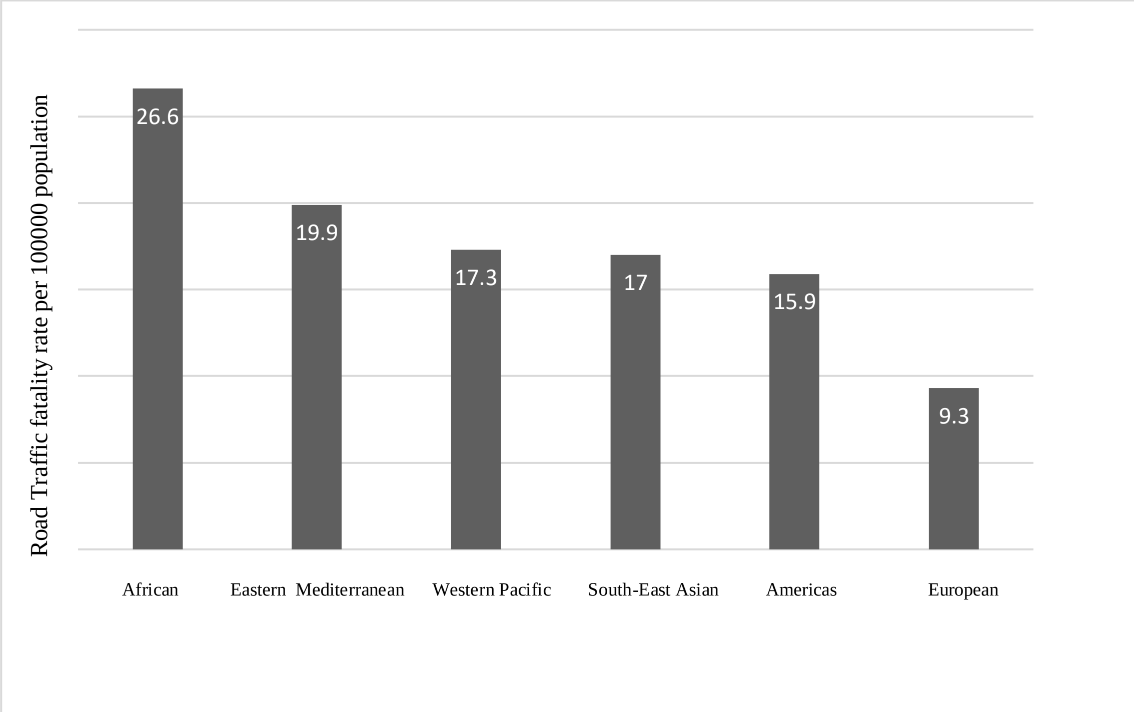


Figure 1.1: Percentage of road traffic fatality per 100000 population within the main continental areas [1]

This image has been removed for copyright reasons

Figure 1.2: World map illustrating the zones that are most and least affected by road deaths per 100000 population. The areas coloured in grey represent countries with the lowest death rate and those coloured in red are the countries with the highest death rate [2]

[1] World Health Organisation (2015). Global Status Report on Road Safety 2015. Retrieved from: https://www.who.int/violence_injury_prevention/road_safety_status/2015/GSRRS2015_Summary_EN_final2.pdf
[2] Retrieved from: http://www.worldlifeexpectancy.com/

Most of the low and middle-income countries are fighting against political, economic and social issues that affect development and challenge the implementation of the proposed initiatives. It is not an easy task to attempt changes and encourage awareness when there is a lack of adequate infrastructure or access to safe vehicles. The WHO Global status report for road safety (2015) showed that 68 countries, of which 84% are low -or middle-income countries, have reported an increase in the number of road traffic deaths since 2010. There is an evident link between successful initiatives to reduce road deaths and the economic situation of a particular country. Europe is clearly the leader in lowering deaths while Africa has the highest road traffic fatality rate. Only 17 countries have made changes to their laws meeting best practices. In all probability, codes of traffic conduct and policy enforcement are one of the most reliable actions for achieving the reduction of road traffic injuries, and this should be the focus of the countries at the bottom of the table. The UN has set an ambitious goal of halving the deaths caused by road traffic accidents by 2020, however it was also predicted that without an immediate action, road traffic crashes will become the seventh leading cause of death by 2030 (WHO, 2018). The expected trend for the high-income countries is more positive (with mortality rate less than nine fatalities per 100,000 population), while the trend for the low- and middle- income countries is not that low (with mortality rate around 20) (Wegman, 2017). Unfortunately, there is only one year left until the deadline set by the UN and many countries have not met the target. Developing countries cannot be expected to simply adopt the safety initiatives of the developed countries, as their circumstances differ and these borrowed initiatives may not be appropriate.

Nonetheless, the UN acknowledges that road safety is a multi-dimensional task that involves all levels of Government and all sectors of society. Given the extent of the problem, small changes in any safety-relevant area can have significant impact even in the safest countries. For this reason, it is important to examine all aspects of the driving task to identify potential gains. Following UN advice, we will now look at what innovations have been successful in the UK, which may offer the potential for export to less-safe countries.

1.3 The United Kingdom

The UK has managed to reach a considerable reduction in its fatality numbers and reported a relatively stable fatality rate in 2016, despite a small increase in fatalities in the last two years (Annual report, DfT, 2017; see Figure 1.3).

This image has been removed for copyright reasons

Figure 1.3: Fatalities in reported road accidents in the United Kingdom between 2006 and 2016 (taken from DfT Annual Report, 2017)

According to the reported road casualties by the Department of Transport (2016) the accident rate went down by 8% and 5% since 2010 and 2014, respectively. During the period between 2010-2015, the British Government employed several actions aiming to improve the safety and prevent death and serious injuries. Since December 2003 a new regulation was implemented prohibiting the use of a held-hand mobile telephone or a hand-held device. The regulation has become even stricter in 2017 (£200 fine and 6 points) as drivers were still being observed to use their mobile phones behind the wheel, and around 35 deaths were recorded in accidents where the mobile phone was a contributory factor in 2016 (DfT, 2017). New speeding fine laws were also introduced in April 2017, motivated by the fact that in 2013 around 3,064 people were killed or seriously injured in crashes where speed was a factor.

In terms of infrastructure there has been an update of the Road Design manual with the first set being released in 2018 and the full update finished by the end of March

2020 (Design Manual for Roads and Bridges, UK Government, 2018). The creation of shared spaces has also provided benefits, including a reduction in the overall traffic flow, and a decrease in both mean speed and the 85th percentile speed of vehicles (Karndacharuk, Wilson & Dunn, 2014; Rudloff, Schoenauernauer & Fellendorf 2013; DfT, 2011). Findings also suggest that more pedestrians make use of the spaces, though collisions typically remain stable or even decrease (e.g. DfT, 2011; Kaparias, et al., 2013). For instance, shared space design in Elwick Square, London, lowered mean speeds of traffic to below 20 mph, though, despite this decrease in speed, 72% of pedestrians expressed concerns over sharing space with vehicles (DfT, 2011).

Although the UK has a good overall road-safety record, young motorcyclists and drivers are still over-represented in the collision statistics. More than 3,800 young people (aged 18-24) were killed on EU roads in 2013 (ETSC, 2016) while in the UK, drivers under 25 account for a third of people who died on British roads. Youngest drivers aged 17-19 only make up 1.5% of UK licence holders, but are involved in 9% of fatal and serious crashes where they are the driver (and more than 200 teenage car passengers were killed or seriously injured when travelling with a young driver; RAC, 2017). Wells, Tong, Sexton, Grayson and Jones (2008) reported that one in five young drivers were involved in a collision in the first six months after passing their practical test. This pattern of results has been replicated in many studies, with novice drivers being overrepresented in crashes in the first 12 months after licensure in countries around the world (Deery, 1999; Foss, Martell, Goodwin, O'Brien & Center, 2011, McCartt Shabanova, & Leaf, 2003; McKnight & McKnight, 2003; Williams & Tefft, 2014).

In order to help young drivers gain the necessary driving experience, the UK government has implemented changes to the driving test and training. The official UK driving test is composed of several parts: a theory test, a computerized hazard perception test, an eye sight test, a vehicle safety test, and a 40-minute driving test with an examiner. A learner is expected to successfully pass all of them. This test is continuously evolving, with new evidence-based innovations included on a regular basis.

For example, a training course called the Pass Plus Scheme was developed in 1999 which after completion allowed novice drivers to apply for an insurance discount. However after examining the results of the scheme in terms of novices' self-report driving behaviour and accidents rates, no differences were found between takers and non-takers (Elliott, 2006).

Another initiative developed for young drivers is the new independent driving section of the on-road test. The learner driver is asked to drive independently for 20 minutes, following road-signs or sat-nav instruction to reach a goal (Department for Transport, 2017).

It was also suggested that Graduated Driver Licensing (ETSC, 2016) would be useful to gradually expose young drivers to increasing risk, better preparing them for independent driving. However, the UK Government declined the program as it considered it highly restrictive (e. g. night driving would not be permitted during the first months, which could be problematic for young shift workers). Other recent innovations include the addition of "Show me, tell me" questions to the driving test in 2017 (where drivers are asked how they would carry out a particular task) and the "Expanded theory test", which is a longer version of the test for those who have had their licence revoked, and are required to undertake the driving test again.

One of the most innovative features of the UK driving test is the Hazard Perception Test (HP test). Since its introduction in 2002, road collisions have ostensibly decreased. Wells et al., (2008) reported a 3% reduction in non-low speed collisions (where blame could be attached) linked to the introduction of the test after surveying over 40,000 learner and novice drivers about learning to drive, and their post-test driving. Scores obtained on the HP test also provided some information, showing that learners who performed better appeared to be 4.5% less responsible for traffic accidents. Due to its role in reducing road collisions, the hazard perception test was awarded with the Prince Michael award for its impact on novice driver collisions.

1.4 The hazard perception test

1.4.1 What is hazard perception?

Hazard perception has been defined as the ability to "read the road" (Mills, Hall, McDonald & Rolls, 1998) or "the process of identifying hazards and quantifying their potential for danger" (Brown & Groeger, 1998). It is often considered as a skill that can be developed with training and experience (Horswill & McKenna, 2004), and is closely linked to where we look, and how we process and understand the visual scene. Any definition of hazard perception skill is, of course, dependent on what we understand to be a hazard. The classical definition proposed by McKenna and Crick (1991) stated that "Hazard perception refers to the ability to detect potentially threatening events by the deployment of a mental model of the road network and how its various elements work. The mental model allows the driver to accumulate information and then run a simulation of what might happen" (McKenna & Crick, 1991).

Although the concept of hazard perception has evolved during the past few years, the above definition contains the main essence of this process, namely the identification of hazards and estimating whether the hazardous situation would pose a danger. Though this definition will be challenged later in this chapter, it encapsulates the early position of many researchers who assisted in the development of the official UK hazard perception test.

1.4.2 What is the official UK hazard perception test?

The Hazard Perception test typically involves 14 short video clips filmed from the driver's perspective, which contain at least one hazardous traffic situation (one clip contains two hazards). Learners are asked to press a button every time they consider they have seen a hazard that would require an evasive manoeuvre in order to avoid a potential collision. The specific definition of a hazard in the UK test is "something that would cause you to take action, such as changing speed or direction".

At the moment a response is produced, a little red flag appears at the bottom of the screen, demonstrating that their button press has been recorded. If one of these responses falls within a pre-specified time window around the appearance of the hazard, then drivers receive a mark out of 5. Each scoring window is divided into five equal segments, with responses in the earliest segment gaining five points, while responses in later segments score fewer marks (see Figure 1.4). The last segment scores one point, and any responses after this (i.e. after the hazard scoring window has closed) score zero points. Equally, any presses before the window score zero points. If multiple presses within the scoring window are recorded, only the earliest receives a mark. Performance on the test is judged on the speed of learner's responses. Faster drivers are typically better at identifying hazards and less likely to be involved in a crash.

This image has been removed for copyright reasons

Figure 1.4: Graphic representation of a typical time frame of hazard onset and offset. The two top images represent the start of the time window (onset) and bottom right image-the end (offset)

Over 100,000 learner drivers take the test every month (with 1.5 million tests taken each year). Since 2015, the old video-based hazard perception clips were replaced with Computer-Generated Imagery (CGI) clips which upgraded the quality of the image (see Figure 1.5). It is possible to access a sample of the official hazard perception clips online, and independently designed tests are also available (though, typically, independent tests are found on commercial sites that require payment for access to the full set of clips. These commercial training aids tend to use video-based clips).

This image has been removed for copyright reasons

Figure 1.5: A Computer-Generated Imagery (CGI) clip representing typical hazardous situation where a van is approaching from the left, intending to join the lane of the driving car. The yellow circle identifies the hazard for the reader and does not appear in the actual video

Academic reports from around the world have supported the success of the typical hazard perception test as a diagnostic tool. A relationship between faster responses on the test and both prospective and retrospective crash risk has been found in several studies (Congdon, 1999; Boufous, Ivers, Senserrick, & Stevenson, 2011; Darby Murray & Raeside 2009; Horswill, Anstey, Hatherly, & Wood 2010; Rosenbloom, Perlman & Pereg, 2011; Horswill, Hill & Wetton, 2015) and hazard perception scores have even been able to predict active crashes in the year following one's driving test (Horswill et al., 2015).

This evidence has also convinced some Australian states to include hazard - licenced drivers in New South Wales and compared the HP results to their crash rate one year after their licencing. It was found that failing the hazard perception test (at least twice) was associated with a significantly higher crash rate one year after the test

(Boufous et al, 2011). Drummond (2000) supported these results by reporting that drivers who obtained higher scores on their HP test had less probability of being involved in a fatal collision one year after sitting the test.

How might it achieve such benefits? McKenna and Horswill (1998) suggested that hazard perception tests provide greater and controlled exposure to diverse driving environments, without placing drivers in any real danger. Second, the need to achieve a pass score in a test encourages drivers to actively seek training, by exposing themselves to hazards in a safe environment, thus reaching a standard of hazard perception skill. Following the success of the hazard perception test in Australia and the UK, the European Transport Safety Council (ETSC, 2016) recommended the implementation of such tests in other European countries. However, within Europe, only the UK and the Netherlands currently employ specific hazard perception testing, and the latter is only a static image test.

1.4.3 The history of the hazard perception test

The first reference to a hazard perception test is found in Pelz and Krupat (1974) who cited a study by Spicer (1964). The study found that drivers, previously involved in road collision, were less able to identify hazardous features when compared to crash-free drivers. According to these results, it was possible to differentiate between safe and unsafe drivers[3]. Pelz and Krupat replicated Spicer's study by showing the participants video clips that contained different traffic situations. While watching the clips, participants were asked to indicate whether the situation was safe or unsafe. Those participants that had not been involved in traffic collisions reacted faster to hazards than those who had been involved. Following a brief lull in hazard perception research, McKenna and Crick (1991) compared hazard perception performance of novice, experienced and expert drivers. The hazard perception test used in their study successfully discriminated between all driving groups, which demonstrated its ability to measure a safety-related skill. Some years later in another study, McKenna and Crick

[3] An equation between safe and experienced drivers has been made in several occasions throughout this thesis. Please, note that being an experienced driver does not necessarily mean being a safe driver but it can still be related and has become an accepted surrogate in the literature.

(1997) compared performance of novices who received training on the hazard perception test to a control group. Young drivers were asked to watch driving clips and generate predictions about the outcomes of different hazardous situations. Response latencies of the training group were significantly faster than those of the control group over time. As a result, researchers quickly raised the possibility that there could be a negative correlation between hazard perception (HP) skill and crash risk, and that a hazard perception test could be feasibly included as part of the official UK driving test (McKenna & Horswill, 1999). In a further study, Horswill and McKenna (2004) defined HP skill as the only higher-order driving ability that correlates with lower crash risk, and Wetton, Hill and Horswill (2011) found the test to be reliable as a part of the official driving test, demonstrating its ability to discriminate between experienced and inexperienced drivers.

In order to further demonstrate the test's ability to discriminate between safe and unsafe drivers, its applicability to a variety of settings and target populations has been investigated. Horswill et al. (2008; 2010) examined the hazard perception ability of older drivers and they found that HP skill declines with age, though this is likely to be due to age-related slowing rather than a loss of skill per se. Hazard perception ability has also been linked to fatigue in novice and experienced drivers, and research suggests that sleepiness in inexperienced drivers significantly slows their responses (Smith, Horswill, Wetton & Chambers, 2009). Increased reaction time to hazards was also related to alcohol consumption (West, Wilding, French, Kemp, & Irving, 1993) and speed choice (Grayson, Groeger, Maycock, Hammond & Field, 2003; McKenna, Horswill & Alexander, 2006).

Diverse methodologies have been used to assess HP skill. Risk appraisal has been measured, asking participants to rate the hazardous situations (Groeger & Chapman, 1996; Pelz & Krupat, 1974; Wallis & Horswill, 2007); scanning behaviour has been measured via eye-movement recording (Chapman & Underwood, 1998; Crundall, Chapman, Phelps & Underwood 2003); drivers have been asked to predict hazards following occlusion (Crundall, 2016; Jackson, Chapman & Crundall, 2009); and location-based responses have been introduced using touch screens (Horswill, Anstey, Hatherly, & Wood, 2010; Scialfa et al., 2011; Wetton et al., 2010). Hazard skill has also been assessed in fully interactive driving simulators where participants are able to drive along a simulated route containing hazards (Garay, Fisher & Hancock 1994; Crundall et al., 2012; Pradhan et al., 2005), or via instrumented cars equipped with control access (Lee, Young & Regan, 2008).

Examining HP skill was even attempted during the practical driving test in New Zealand (NZ Transport Agency, 2012), however it was deemed too unreliable due to variations between tests in terms of traffic and location, and the possibility that a candidate may train on a specific route or area. Furthermore, results were based on the examiner's personal interpretation of the candidate's response of why they carried out a particular manoeuvre. The New South Wales test has also introduced a second level of response, where a judgment of action is required (though it has been criticized for not focusing solely on the hazard perception skill) (Horswill et al., 2015; Wetton et al., 2011).

Although the considerable body of literature provides ostensibly clear evidence of the diagnostic validity of the Hazard Perception test, results have not always been consistent, and several research groups in different countries have failed to replicate the basic experiential effect (Chapman & Underwood, 1998; Crundall, Underwood, & Chapman, 2002; Sagberg & Bjørnskau, 2006; Borowsky, Shinar & Oron-Gilad, 2010; Underwood, Ngai, & Underwood, 2013) and crash-liability effect (e.g., Groeger, Field, Hammond, 1998). Chapman and Underwood (1998) presented their participants with hazardous videos filmed on UK roads and recorded both their eye movements and response latencies to hazards. Surprisingly, they failed to find differences between experienced and novices' response times. However, they found that novices' fixation durations increased during the hazard window more so than those of experienced drivers and their visual spread of search decreased, suggesting that inexperienced drivers were more prone to attentional capture by the hazard. This showed that underlying behavioural differences were present, but that the simple button press was too insensitive to detect them. Crundall, Underwood and Chapman (2002) tested three different driving groups (novices, police drivers and experienced drivers) and failed to find differences in the number of hazards detected. Responses were measured by adjusting a slider to a sliding scale rating how hazardous the situation shown in the clip was, with sudden increases in the slider position reflecting the detection of a specific hazard. Crundall et al. (2003) also recorded eye movements and electrodermal responses (EDRs). They replicated Underwood and Chapman's results in relation to the eye movements with novices reporting longer fixation duration and narrower horizontal spread of search. They also found both police drivers and control group to produce a higher number of discrete EDRs. Again, it appears that there were underlying behavioural differences between the groups but these were not identified in the slider responses.

In Norway, Sagberg and Bjørnskau (2006) examined reaction times to hazardous situation of four different driving groups (holding a driving licence for 1, 5, 9 months and 27.1 years). They also failed to find differences in response latency between the groups, concluding that while some of their hazards might have been useful, the variation in discriminative validity across their clip set rendered their overall score to be non-significant between driver groups.

In Malaysia, Lim, Sheppard and Crundall (2013) compared response latencies and eye movement behaviour of Malaysian and UK drivers. They showed participants both UK and Malaysian footage and asked them to press the mouse every time they thought they had seen a hazard. Participants were also asked to verbally describe the hazard after each video. Lim et al. found that UK participants identified significantly more hazards than Malaysian participants and responded correctly to more hazards from their own country. However, no differences between experience groups were found.

Finally, in Israel, Borowsky, Oron-Gilad and Parmet (2010) also failed to find significant differences between novices and experienced drivers. These studies reflect the difficulty that some research groups have had replicating the basic experiential effect.

The above studies have all applied a variety of methodologies to measure hazard perception performance. It could be argued that the reason behind the mixed evidence is specifically these differences in response mode. It cannot be claimed that some methodologies are better than others, although these inconsistencies could be contributing to the mixed results. Pradhan and Crundall (2017) argued that the hazard perception skill is a complex variable composed of several sub-components (see section 1.5 in this chapter for detailed explanation) and each sub-component should be measured independently. For example, asking participants to rate a hazardous situation measures their subjective risk estimation (which proves very useful in terms of participant's subjective experience but often these measures need to be correlated with other more objective ones), while reacting to hazards or looking at fixations to different parts of the hazardous scene reports a more objective information about where participants direct their attention. Typically, the hazard perception skill is measured by pressing a button every time a participant sees a hazard while watching a short video clip. However, due to its complexity simple reaction times seem to not be able to catch the underlying sub-processes.

For this reason, hazard perception has also been measured by looking at eye movements and RTs have been correlated to fixations. Eye-movement measures have demonstrated their ability to discriminate between experienced and novice drivers, and according to some studies eye movements are more reliable than simple RTs in identifying experiential differences (Crundall et al., 2012). The benefit of measuring fixations to hazards is that it shows whether the participant have spotted the hazard independently of their threshold bias and their reaction/action after spotting it. Although, it cannot be assured that fixation at a certain place reflects adequate processing of information, there are important indicators such as length of fixation and meaningful gaze patterns that show attention was not randomly allocated. The only disadvantage is the need of specific equipment and training which can be costly, although nowadays technology has advanced and it is possible to acquire such equipment for more reasonable prices (including portable devices).

Similar disadvantages are found with driving simulators, and they are much more expensive and the most accurate ones are not easily portable. However, the main advantage of the driving simulators is that wide range of driving behaviour can be recorded in real time under extremely demanding conditions without putting drivers and other road users at risk. A variety of traffic scenarios can be displayed virtually for research and training purposes, although the process requires a large amount of time for data analyses (and trained staff).

Therefore, many studies use video-based stimuli. Not only because of the pragmatics but this methodology is much more cost-effective and, following McKenna and Crick's (1991) argument, simulation and immersion are not required to assess the part-skill of HP. Historically, the hazard perception skill has been assessed using short video clips filmed from the driver's perspective and there is a large body of literature that has successfully applied this methodology. Furthermore, there are debates whether staged or real hazards should be used, as real hazards do not allow full control over the scenario. While filming hazards might also require a certain amount of driving hours, it allows testing of HP skill without putting drivers at risk and it also permits multiple assessments simultaneously.

1.4.4 Why is the evidence mixed?

Why do some studies fail to find clear evidence of differences between experienced and novice drivers, or crash-involved and crash-free drivers? This section will consider possible reasons.

1.4.4.1 Problems with using simple response times

There are a number of potential confounds associated with the traditional method of simply calculating response times within a scoring window. First, a simple button push does not necessarily reflect that the driver was pressing for the actual hazard. The response is considered correct based on when it is made, rather than what it is made for. If there are other potential hazards in the scene, then it is possible that the participant has responded for a different reason than that of the a priori hazard.

Secondly, the need for scoring windows requires hazard onsets to be defined. This is often an arbitrary decision that leads to potentially valid button presses – made just before the window opens – to be given a score of zero. The definition of the hazard onset in the traditional methodology is essentially a balancing of the likelihood of type 1 and type 2 errors (see Figure 1.6). Some researchers may have been more successful in this balancing act than others.

Figure 1.6: The images represent stills from a video along a timeline (the black arrow). The red arrow represents a response that was made to subtle cues that a participant noticed, suggesting that a pedestrian might step out. This falls before the scoring window opens and therefore is counted, unfairly, as a 'miss'. The first green arrow reflects a different participant's concern about the car ahead. This is not the true hazard, yet due to the response occurring in the scoring window, it is counted as a very quick 'hit'. The later green arrow represents a true response to the hazard. Unfortunately there is no way to distinguish between the validity of these two responses, and the participant receives the score associated with the first green arrow rather than the second[4]

1.4.4.2 Differences in response mode

Different research groups design their own tests, and often modify the traditional methodology used in the DVSA test (Pradhan & Crundall, 2017). For instance, other researchers have noted the problems with simple response times and have therefore employed a variety of different response modes. Wetton et al. (2010) replaced the traditional button response with location-based responses via touch screen devices, while Wetton, Hill, & Horswill, (2011, 2013) used a localised mouse click. Both methods provide a measure of accuracy, lacking in the traditional methodology, though they do not get around the need to define scoring windows. Unfortunately, this method may entail other confounds such as individual differences in pointing tasks (e.g. Zhai, Kong & Ren, 2004) and age differences in mouse and touch screen use (e.g. Hertzum

[4] Source: Crundall, D. (2018, June). Understanding hazard perception. Presented at the International Congress of Applied Psychology, Montreal, CA.

& Hornbaek, 2010; Wahlström, Svensson, Hagberg & Wallace-Johnson, 2000; Yamauchi, Seo, Jett, Parks & Bowman, 2015). Also, if experienced drivers spot hazards earlier than inexperienced drivers (e.g. Crundall et al. 2012), then the hazard is likely to be smaller (i.e. further away) than when spotted by inexperienced drivers. According to Fitts' Law (Fitts, 1954), a smaller target will increase demands on accuracy and therefore slow pointing speed, potentially negating the experiential benefit of perceiving the hazard sooner.

Other response modes include sliding levers (Crundall et al., 2003; Pelz & Krupat, 1974), hazardous ratings (Groeger & Chapman, 1996), eye movement measures (Crundall & Underwood, 1998; Crundall et al., 2003; Garay, Fisher & Hancock, 2004) and behavioural responses in simulators (Crundall, Andrews, van Loon & Chapman, 2010). These variations in response mode make it difficult to compare studies that have or have not found hazard perception tests to successfully discriminate between safe and less-safe drivers.

1.4.4.3 Differences in analysis

Response times have also been analysed using different statistical methods. Very few research papers opt for using the five point scoring window, but instead work with the raw response times. McKenna et al. (2006) and Wetton et al. (2010) calculated the average response times across all hazards by standardising raw scores into Z scores thus addressing the differences in duration of each hazard window. However, Parmet Meir & Borrowski (2014) argued that this method of analysis failed to address missing responses and proposed a survival analysis as a more suitable methodology for analysing response times (though this has yet to receive wide-support from the field).

Wallis and Horswill (2007) used a fuzzy signal detection theory arguing that the traditional SDT is not suitable for tasks such as hazard perception as it classifies stimuli into binary categories. They argued that this binary approach (measuring whether a scene is hazardous or not) is not suitable as every scenario is potentially hazardous to some extent. The fuzzy signal detection method calculates a percentage of hit response and assigns a proportion of false alarms and correct rejections, and replacing missing values by mean reaction times.

1.4.4.4 Differences in stimuli

Differences in methodology also exist in relation to the filming of hazardous videos. At the initial stages of HP research, staged hazards used to be the most common method of creating hazardous driving clips (McKenna and Crick, 1991; Catchpole & Leadbeatter, 2000). Sexton (2000) argued that staged events were more efficient for hazard discrimination, however Crundall et al. (2003) argued that this efficiency could have been as a result of biased selection that benefitted experienced drivers (as staged hazards are likely to be set by expert drivers on the basis of their experience). Staged dangerous situations may therefore reduce test validity, as a result of containing potentially unrealistic hazardous situations that may not reflect the subtlety of cues that occur in real world. Despite the staged-vs.-naturalistic hazard debate being acknowledged as an important one, no research has directly addressed this issue (to the knowledge of the author).

Many other differences in stimuli are apparent across research studies. Some researchers use 'latent' hazards (i.e. potential hazards that never occur, Vlakveld, 2014), while other researchers have used abrupt hazards (e.g. Underwood et al., 2013) which are very difficult to anticipate. Some researchers even use still images rather than videos (Tūskė, Šeibokaitė, Endriulaitienė & Lehtonen, 2018).

1.4.4.5 Criterion bias

Lim et al. (2013) raised the possibility of criterion bias influencing the simple push-button response required in the traditional hazard perception methodology. The criterion bias is participant's internal threshold of what is hazardous. Individual thresholds can be influenced by cultural differences in acceptable driving norms (e.g. Lim et al., 2013, 2014), and by driving experience and expertise, with more advanced drivers discounting hazards if they fall within their self-perceived range of skill (e.g. Crundall et al., 2003).

Lim et al. (2013) suggested that perhaps the traditional hazard perception methodology would not be suitable for use in different countries, where environmentally evoked high criterion bias may render the test insensitive to the skills of the safest drivers in those environments. The Malaysian drivers tested in their study reported a higher threshold bias in appraising hazardous situations. They made significantly fewer responses compared to the UK drivers, very likely due to the more hazardous Malaysian driving environment desensitising them to events that would be termed hazardous in the UK.

Such criterion bias is likely to influence response times in a traditional hazard perception test. Highly experienced drivers may spot, and even physiologically react to the presence of hazards sooner than novice drivers, but may withhold a response until the event reaches their threshold for a hazard (e.g. Crundall et al., 2003).

1.4.4.6 Common vocabulary and the lack of theory underlying HP

No consensus of a definition of hazard perception has been reached so far. Different terms such as hazard anticipation (McDonald, Goodwin, Pradhan, Romoser & Williams 2015) and hazard avoidance (Pradhan & Crundall, 2017) have been proposed in order to encapsulate the complexity of the process, however the lack of agreement has created confusion. Authors have made use of different terms to refer to the same concept, most of the time defining other sub-processes of hazard perception that overlap with each other. For example, Vlakveld (2014) refers to hazards as overt and covert latent hazards, while Crundall et al. (2012) refers to the same type of hazards as behavioural prediction (BP) and environmental prediction (EP) hazards. There is even confusion between the concepts of hazard perception and risk perception, referring to the appraisal of a particular situation as “risk perception”, when the correct term would be risk estimation (Egea-Capparos, 2012) or hazard appraisal. One relevant theory in the field that looks specifically at risk appraisal and anticipation has been proposed by Kinnear (2009). Kinnear applied the Interface Model (Fuller, 2005) in conjunction with the Somatic Marker Hypothesis (Damasio, 1994) to assess drivers’ risk appraisal and how this shapes a decision and behavioural response. This theory supports that risk is appraised both emotionally and analytically, where emotions automatically influence decision making and hazard judgement. Kinnear found that novice drivers show less emotional appraisal and lower anticipatory scores to developing hazards in comparison to experienced drivers which suggests that novices might not appreciate the inherent danger and the realistic risk involved. This finding further demonstrated the importance of separating risk appraisal of hazard perception as novices might be able to spot a hazard but not to appraise correctly the danger it poses. This sub-estimation of the risk (most of the time due to lack of experience) leads novices to erroneously think they can cope with the hazardous situation, overestimating their driving skills, adopting a risky driving style, which untimely leads to higher crash rates for this group.

Agreeing on universal definitions and terminology would seem like a logical step to unify research methodologies and remove the mixed evidence from the research record. The theory most often linked to hazard perception test has been the Situation Awareness (SA) theory (Endsley, 1995). The model of SA distinguishes three main levels of global projection or in other words comprehension of other drivers' behaviour on the road (Endsley 1995a). The first level (Level 1) refers to the perception of the elements in the environment; the second level (Level 2) to the comprehension of the situation, meaning that it is not sufficient only to perceive the elements, but to extract the meaning of it; and finally the third level (Level 3) is the projection of all these elements into the near future. These three levels create a mental model which is constantly updating according to the goals of the drivers. However, according to Durso and Gronlund (1999), SA is not based on empirical findings, meaning that it does not meet the basic criteria for a coherent theory. Neither does the model provide a specific explanation of how drivers engage in hazard perception, though its high level descriptive qualities have provided a general framework within which to understand the skill (Horswill & McKenna, 2004; Wetton, Hill & Horswill, 2013). Crundall (2016) argued that the reason behind the specific theoretical lacuna was the motivation to generate a diagnostic test rather than developing a theoretical basis. In their chapter, Pradhan and Crundall (2017) addressed the theoretical understanding of the HP skill and proposed a new framework of hazard avoidance for better understanding the concept.

1.5 A new framework

As noted previously, hazard perception has been defined as the ability to "read the road" (Mills et al. 1998) or "the process of identifying hazards and quantifying their potential for danger” (Brown & Groeger 1998). This definition is too simplistic however and does not unpack the sub-processes that are likely involved. Even the terminology used lacks specificity. For instance, a response time to a hazard does not just reflect perceptual processes, but also involves post-perceptual processes such as appraisal.

Pradhan and Crundall (2017) noted the problems in terminology when they joined together to write a book chapter on hazard perception. In an attempt to simplify the terminology, and bring together US and UK definitions, they started with the concept of *hazard avoidance* to describe the whole process from searching for hazardous precursors to selecting the most appropriate response.

According to Pradhan and Crundall, hazard perception is a broad process which includes a combination of various sub-processes and post perceptive processes, such as hazard searching, precursor prioritisation, hazard prediction, hazard fixation, hazard processing, hazard appraisal and hazard reaction or response (Pradhan and Crundall, 2017). Drivers’ attention could be attracted by the salience of the hazard (e.g. police lights, sudden movement etc.), whereas at other times, the hazard might be less salient and therefore the driver must search for it (*hazard searching*). General visual search patterns of drivers have been widely studied and results have shown that inexperienced drivers have a very narrow spread of visual search focusing on the immediate road ahead. Conversely, experienced drivers are more likely to have a wider spread of search, often fixating to the left and the right of the road ahead (e.g. focusing on parked vehicles, pedestrians, side roads etc.; Chapman & Underwood, 1998; Crundall & Underwood, 1998; Konstantopoulos, Chapman & Crundall, 2010; Lehtonen, Lappi, Koirikivi & Summala 2014). Experienced drivers are also better at identifying *hazard precursors* which provide clues that help anticipate the upcoming hazard. Two types of precursors linked to hazards have been identified so far: behavioural prediction hazard (BP) and environmental prediction hazards (EP) (Crundall et al., 2012). The first type are those hazards and precursors that are directly linked to each other (e.g. a cyclist who is cycling in front of your car suddenly cuts into your path). The second type refers to precursors as parts of the environment from where obscured hazards might occur (e.g. a parked car may obscure a pedestrian who then steps into the road). Whereas the former hazards can be predicted from the behaviour of targets that subsequently become the hazards, the latter hazards can only be inferred from their statistical co-occurrence with specific environment features. Precursors might seem similar to latent hazards (Pollatsek et al., 2006), however latent hazards are those that could potentially materialize (could be hidden)-e.g., a pedestrian behind a van that suddenly emerges into the path of the driver, while precursors are clues from the environment that helps the driver predict a potential future hazardous situation - e.g. a van that could be occluding a pedestrian.

Crundall et al. examined three groups of drivers (learner drivers, experienced drivers and driving instructors), asking them to drive in a simulator through nine hazardous scenarios while their eye movements were recorded. They found that learner drivers had difficulties in spotting BP precursors and EP hazards, with fewer fixations to these elements in comparison to the experienced drivers. It has been suggested that EP hazards are more apparent to drivers that have been exposed to situations featuring similar precursors. Also, less experienced drivers typically find difficulties in identifying the sources of possible hazards in the driving environment due to the insufficient exposure to these types of situations (which one gains with years of driving) (McKenna et al. 2006). Another study conducted by Underwood et al. (2013) proposed similar classification to those of Crundall et al. (2012) defining hazards according to their appearance. An abrupt hazard would appear suddenly in the traffic scene requiring an immediate attention and action, while a gradual hazard unfolds in a way that a precursor would announce its appearance. For example, a motorbike appearing suddenly from an obscured side street would be classified as an abrupt hazard, while a pedestrian approaching slowly towards a zebra crossing would be classified as a gradual hazard. Underwood et al. divided the sample into three groups: motorcycle riders, experienced drivers and inexperienced drivers. They found that gradual hazards were detected significantly more often although these findings were significant only for the motorcycle riders. No differences were found for the number of hazards between experienced and novice drivers regardless of the type of hazard.

Once the driver has identified a precursor they must try to predict what would happen next in the driving scene in order to be better prepared to respond (rather than simply *react*). *Hazard prediction* refers to the ability of drivers to anticipate possible hazardous situations. This process has been studied in a similar way to hazard perception, showing short video clips filmed from the driver’s perspective, however in this case clips are occluded immediately prior to a hazardous event where participants are asked to try and predict how the driving situation is going to develop (making use of precursors). A hazard prediction test also called the “What happens next” task, was developed in 2009 by Jackson, Chapman and Crundall, and has since been validated in different countries by several research groups such as the UK (Crundall, 2016), Malaysia (Lim et al., 2014) and Spain (Castro et al., 2014, 2016). These studies consistently find experienced drivers to outperform novices in predicting hazards.

Once the driver suspects that a hazard will appear, how do they know exactly when the hazard becomes hazardous? Hazard perception is typically measured by recording reaction time of participants which requires the setting of a time window or scoring window. Only the earliest response within each scoring window is considered as correct. For example, when driving on an urban road with parked cars on both sides of the street, suddenly one of the parked cars indicates, and then makes a move to join the road- the car is already a hazard once trying to join the street, while the action of indicating is a precursor to the hazard. Each hazard should be considered individually, as some hazards onsets are easy to define while others could be more difficult (Pradhan and Crundall, 2017).

Many factors, both objective and subjective are likely to influence the hazard response. Individual differences in judging the hazardousness of an event (response criterion) or the time required to process the actual hazard (e.g. Deery, 1999) also play an important role during hazard avoidance. Differences in fixations related to hazard processing have also been found, although with mixed results. On some occasions novices devoted more time fixating the critical hazard (Chapman and Underwood, 1998) and in others less time fixating the critical stimuli in comparison to more experienced groups (Crundall et al., 2012). This last finding may actually reflect the failure to process the critical information as either they are not conscious of the importance of the stimuli they are fixating or are misappraising how hazardous it is. Hazard processing has been assessed by Borowsky and Oron-Gilad (2013) who divided it into different components (hazard perception, hazard categorisation and hazardousness ratings) and they found that risk perception affects both real-time responses of hazard perception and subjective responses such as ratings and categorisations. Risk perception influences participant’s reaction time when is related to the possibility of a collision, while when related to the severity of a possible collision it affects the subjective responses such as hazard categorisation. It is well known that young drivers tend to overestimate their driving skills (De Craen, Twisk, Hagenzieker, Elffers & Brookhuis 2011) and underestimate accident risk perceptions (White, Cunningham & Titchener, 2011).

There is also confusion between the concepts of hazard perception and risk perception, referring to the appraisal of a particular situation as "risk perception", when the correct term would be *risk estimation* (Egea-Capparos, 2012) or *hazard appraisal*. When hazard perception is measured using behavioural measures we refer to the objective perception of a danger (e.g. eye tracking), while hazard appraisal is the subjective estimation of the potential danger and is often measured using self-report data. Deery (1999) in his "Model of processes underlying driving behaviour in response to potential hazards" referred to the importance of the relation between hazard estimation and self-assessment of one's own driving skills. Typically, novice drivers tend to overestimate their driving skills due to lack of experience and overconfidence that they can deal with the hazardous situation. They even may successfully process the hazardous event, yet unsuccessfully calibrate and appraise the levels of risk (De Craen et al., 2011; Horrey, Lesch, Mitsopoulos-Rubens & Lee 2015). One's own criteria of hazardousness play a powerful role in the hazard perception process and participants would remain consistent in their criteria even when they have been provided with a definition of a traffic conflict (Shinar, 1984; Kruysse, 1991).

Finally, when a materialised hazard does occur, the driver is expected to make a response. The type of action will depend on whether the driver has predicted a hazard and planned a potential response, or whether the driver simply reacts to a hazard as it occurs. When a conscious search for a hazard takes place, all the described sub-processes are present and the driver makes a *response*, however when the hazard is just noticed without any previous deliberating, then the driver *reacts* almost automatically to the hazard (Pradhan & Crundall, 2017). All the aforementioned processes involved in hazard avoidance, and some additional ones, have been defined in Table 1.1.

Table 1.1: Definitions suggested by Pradhan and Crundall (2017) for the processes involved in hazard avoidance (taken from Pradhan and Crundall, 2017).

Term	Definition
Hazard Perception	A collection of hazard avoidance sub-processes which variably include hazard searching, hazard prediction, precursor prioritisation, hazard fixation, hazard processing, hazard appraisal, and hazard reaction.
Hazard Avoidance	The over-arching term to describe the process of avoiding a collision with a hazard from initial searching for hazards, through to the successful selection of appropriate response
Hazard Salience	The ability of a hazard to draw attention to itself through bottom-up features including sudden movement, looming, colour, luminance, etc.
Hazard Searching	The direction of overt attention (i.e. eye movements) to areas of the visual scene that are most likely to produce hazards
Hazard Precursor	The clues to an upcoming hazard (e.g. a pedestrian on the pavement is a precursor prior to stepping into the road and becoming a hazard)
Hazard Evidence	The visual evidence contained within an object or location that the driver uses to judge immediate danger (a hazard) or a strong potential for danger (a precursor).
Abrupt Hazard	A hazard that appears without (or ostensibly without) any precursor.
Precursor Prioritisation	Prioritising and labelling precursors most likely to produce hazards for continued monitoring through overt and covert attention.
Hazard Prediction	Predicting a hazard before it appears on the basis of the hazard evidence present in precursors
Hazard Onset	When a precursor turns into a hazard, e.g., a pedestrian steps off the sidewalk into the road, or from behind a parked vehicle; the oncoming car begins to turn across your path, etc.
Hazard Fixation Latency	The time taken to first fixate the hazard. This can be zero if the driver is fixating the precursor at the time it becomes a hazard.
Hazard Processing	The time taken to identify the object as a hazard, primarily reflected in fixation durations and dwell time upon the hazard
Hazard Appraisal	Assessing the level of risk posed by the hazard both in absolute terms and in relation to one's own self-perceived skills
Hazard Mitigation	The act of reducing the probability of a collision with a future hazard by changing one's driving behaviour (e.g. changing lane position, headway) on the basis of prioritised precursors.
Hazard Reaction	Any behavioural outcome from identifying a hazard. This could be positive (braking) or negative (freezing)

Hazard Response	A subsection of hazard reactions composed of deliberate actions (e.g. intentional braking). While more likely to be positive, a poorly chosen intentional response can still be negative (e.g. trying to overtake a braking car when oncoming traffic is too close)

1.6 Conclusion

There is considerable amount of literature that has provided solid evidence that HP tests can reduce fatalities in both UK and Australia, and keep unsafe drivers off the road (Boufous, 2011; Drummond, 2000; Horswill & McKenna, 1999; Wells et al., 2008; Wetton et al., 2011). However, there is also an important amount of research that has failed to replicate these results (Borowsky, Shinar & Oron-Gilad, 2010; Groeger, 2000; Crundall & Underwood, 1998; Underwood et al., 2013; Sagberg & Bjørnskau, 2006). This mixed evidence is especially noticeable when examining research from different countries where each research group has developed their own HP test using different methodologies (e.g. Norwegian HP test, Sagberg & Bjørnskau, 2006-they used a continuous hazardous footage and a secondary task; Israeli HP test, Parmet, Meir & Borowsky, 2014-who discussed that different analysis approach, survival analysis, would be more suitable to deal with missing responses; Singaporean HP test, Yeung & Wong, 2015-they used staged and abrupt hazards; German HP test, Malone & Brünken, 2015-they used CGI footage and included non-hazardous situations). These methodological inconsistencies could have contributed to the lack of replication, which highlights the necessity of a consistent protocol that would contain clear methodological guidelines on how to build a hazard perception test that would be able to provide effects that extend beyond the boundaries of a single country.

The first ever cross-cultural comparison that applied identical HP methodology in two different countries was conducted by Lim et al. (2013). Lim et al. developed a Malaysian-UK HP test that contained video clips filmed in both Malaysia and UK and classified to matched (representing similar hazardous scenarios that could be encountered in both countries) and unmatched (hazards representative only in one of the countries). Malaysian and UK drivers were asked to press the mouse every time they thought they have seen a hazard (a hazard was defined as any object or situation that would make the driver perform an evasive manoeuvre). Lim et al. found that UK participants identified significantly more hazards than Malaysians. Furthermore, UK drivers made many more responses to the Malaysian clips. Interestingly, no differences between the driver groups were found based on experience, with even UK experienced drivers failing to outperform UK novices when viewing UK clips. Lim et al argued that Malaysian drivers might have a higher threshold for deciding that a situation is hazardous, perhaps as a result of desensitisation to hazards through exposure to the more dangerous Malaysian roads, hence their reluctance to respond.

These results raised some questions related to the transferability of the HP, considering that the classical reaction time paradigm may not be suitable for international export due to the confound that the threshold bias creates. In a study previous to this of Lim et al., Wetton et al. (2010) argued that the hazard perception skill is generalised even in unfamiliar locations and that the HP does not need validation, however they used driving footage from the UK (the official UK hazard perception test) and Australia (Brisbane). These two settings are very similar in terms of road structure, architecture and driving behaviour and are not representative for the rest of low- and middle- income countries and their driving environments. Participants from different cultural contexts do not appraise and react to hazards identically. This fact might not be essential for hazard perception purposes in Australia or the UK, however in terms of international export HP methodology should be adapted to the driving context where it is intended to be applied. Is the HP test suitable to be exported as a diagnostic tool in other countries, especially those with more hazardous driving environment? Wallis and Horswill (2007) argued that the criterion bias could also be present in developed countries. As the evidence is inconclusive, further research should focus on examining whether the results of the HP are transferable and generalizable to different driving contexts. In order to be able to address the recommendation of the European Transport Safety Council (ETSC, 2016) and implement the Hazard Perception test in other European countries, an empirical evidence is needed to demonstrate effects that are generalizable beyond borders. Otherwise, it will be difficult to promote this test to international Governments without clear evidence.

1.7 Outline of this thesis

In this thesis, an attempt will be made to test the cross-cultural transferability of the HP test. In Chapter 2, the method of filming driving footage and general methods applied to all studies will be described. In Chapters 3 and 4, different stimuli designs will be compared in order to select the most appropriate for future studies. This new design includes mirror information and different type of car layout. Chapter 5 reports attempts to apply the HP methodology across different countries. Three different driving environments will be compared: China (high accident rate), the UK (low accident rate), and in the middle position is Spain (with an intermediate accident rate, albeit closer to that of the UK). These three countries differ considerably in driving environment, cultural and social driving rules and provide a greater challenge than comparing Australia to the UK (Wetton et al, 2010). Chapter 6 reports the development of a hazard prediction test in an attempt to improve validity ("What happens next") and in Chapter 7, the hazard prediction test is applied across China, Spain and the UK. In Chapter 8, the hazard prediction test is further refined, before being deployed in a completely different cultural context in Chapter 9 (Israel). The final result is a blue print for a culturally-agnostic hazard perception (actually, *prediction*) test that can be promoted to the governments of countries who are in need of successful road safety initiatives. Finally, Chapter 10 will summarise all the major findings of this thesis, offering a discussion of the results and practical implications.

The overall aim of this thesis is to create a methodology which could be followed by other researchers in their attempts to create a hazard perception test. The intention is to develop a test *format* that will be able to differentiate between experienced groups regardless of their cultural background. As the prediction methodology proved to be more successful in discriminating between driving groups overseas, it is expected that it will be more suitable for international export than the traditional hazard perception methodology. Nonetheless, both methodologies will be assessed and clip design for both has been refined featuring mirror information thus allowing the inclusion of wider range of hazardous scenarios.

Lastly, it is expected that the new test format will be suitable for promotion to international governments who can follow it in their attempts to create an HP test as a part of their official driving tests.

Chapter 2

METHODOLOGY

This chapter will begin by briefly reviewing some of the primary methods used in traffic and transport psychology, before detailing the choices made for this thesis. While each subsequent study will contain its own method section, there are some overall considerations and commonalities across the methods that will be detailed in this chapter for reference.

2.1 Driving research

The methodology used in traffic psychology has evolved considerably during the last two decades. Modern equipment has made it both technologically possible and economically feasible to study driving behaviour in ever-more realistic environments. Driving behaviour is often observed within two main settings: ecological and laboratory settings. Particularly, research in traffic psychology has been conducted mostly within laboratory settings under strict control, yet the aim was the understanding of the basic mechanisms behind the processes involved in driving. The most common methods employed within laboratory settings feature driving simulators or video clips filmed from the driver's perspective. Instrumented vehicles have also been employed to test normal everyday driving. It should be noted however that the earliest method used in traffic research was naturalistic observation (Brown, 1967; Dodge 1923, Mourant & Rockwell, 1970) and within the past few years, the laboratory trend has given way to a resurgence in field research utilising the latest sensors to capture real world driving. However, in this thesis only video-based methodology has been applied as the aim was to design a time and cost-effective test affordable for the developing countries (and alternatives such as CGI clips could be too costly).

2.1.1 Methods used in this thesis

For the purposes of this thesis, video-based stimuli will be recorded, edited and used. Some studies will also employ eye movement recording where necessary in order to inform the design of the hazard perception test, and examine the impact of any test modifications on visual search.

Hazard perception (and prediction) skill will be examined using short hazardous video clips created and edited to build hazard perception and hazard prediction tests. Since the aim of this thesis is to create a time and cost-effective tool, which could be applied world-wide, the use of equipment such as driving simulators falls outside the scope of this work.

2.2 Video-based studies

Due to the unceasing debate between ecological validity and experimental control, techniques have been developed to address this problem. Video-based tests were developed in attempt to meet both, providing an ecologically valid context which will require real decisions in real time, and experimental control that ensures all participants see the same scenarios without exposure to real risk. Indeed, drivers cannot interact with video-based stimuli in the same way they can when participating in a driving simulator study, but it is also true that some driving scenarios are difficult to be recreated in a simulator (another reason is the comparison between a real life image vs. images generated by a software). Video-based tests offer the closest to real life hazardous traffic situations, typically using non-staged hazards, although studies have also used deliberately arranged hazards. This type of test appears to be a good alternative in terms of cost and time.

Although some authors argued that video-based tests do not reflect real driving performance (Groeger, 2000), numerous other studies suggest that performance in video-based tests was similar to the actual driving behaviours (Mills et al., 1998; Watts & Quimby, 1979). One might ask why such simple responses are used to measure driving performance? There are several reasons. First, the validity of such tests have been demonstrated (e. g. Crundall et al, 2012; Horwsill & McKenna, 1999; Wetton et al., 2011) and results showed that there is link between good hazard perception skills and accidents rates (Drummond, 2000; Wells et al., 2008). Second, video-based test are not only able to discriminate between safe and unsafe drivers, but are also a cost and time-effective tool. Third, it provides both ecological and experimental validity and permits the measurement of specific elements of the driving task within control settings while exposed to real world hazardous situations.

2.1.1 Hazard perception clips

Hazard perception stimuli have improved in quality over the last 50 years from 50 mm cine film (e.g. Spicer 1964, cited in Pelz & Krupat, 1974) to photo-realistic computer-generated imagery (CGI) used in the current DVSA test (see Figure 2.1). They tend to be clips of variable duration (often no more than 60 seconds) filmed from the driver's perspective as if through the windscreen of a moving vehicle. Research groups tend to record their own footage and create their own tests. Some research-based hazard tests have been filmed with cameras attached to the roof or bonnet of a car (e.g. Shahar, Alberti, Clarke & Crundall, 2010) while others have had cameras attached to the inside of the windscreen. There is no accepted position for placing cameras, though test developers tend to approximate the eye-line of the driver as much as possible.

Hazard tests vary in their presentation. The traditional test is a single-screen video (e.g. McKenna & Crick, 1991). Other variations have extended the field of view across multiple-screens (Shahar et al., 2010), changed the vehicle from which they have been filmed (Crundall & Kroll, 2018; Horswill & Helman, 2003), or added mirror information (Engström, Johansson, & Östlund, 2005; Mackenzie & Harris, 2016; Recarte & Nunes, 2003; Savage, Potter, & Tatler, 2013). For this thesis, the examples of Crundall & Kroll (2018) and Shahar et al., (2010; 2012) will be followed in providing mirror information to supplement the forward view. This new design permits the early prediction of overtaking hazards which was not possible with the traditional hazard perception test.

This image has been removed for copyright reasons

Benda & Hoyos, 1982: Estimating hazards in traffic situations (a)

This image has been removed for copyright reasons

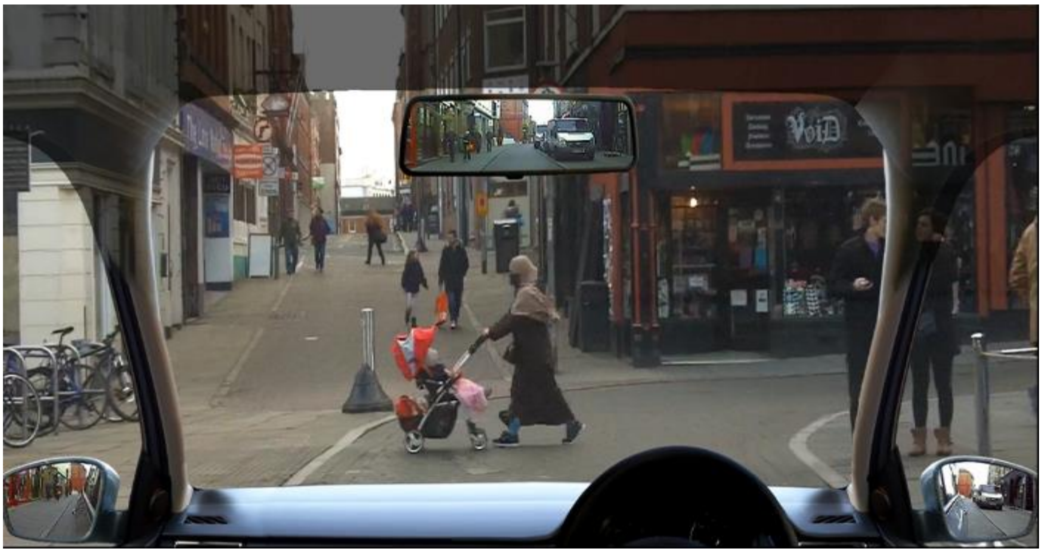
Wetton et al., 2010: The development and validation of two complementary measures of drivers' hazard perception ability (b)

This image has been removed for copyright reasons

Shahar, Alberti, Clark and Crundall, 2010: Hazard perception as a Function of target location and the field of view (c)

This image has been removed for copyright reasons

Underwood, Ngai, Underwood, 2013: Driving experience and situation awareness in hazard detection (d)



Ventsislavova et al., 2018: A cross-cultural comparison of Hazard Perception and Hazard Prediction tests (e)

This image has been removed for copyright reasons

Computer Generated Imagery part of the Official UK Hazard Perception test, 2018 (f)

Figure 2.1: Example images of different hazardous video clips developed by several research groups during the last two decades ((a) to (e)) and the DVSA driving test (f).

2.3 Filming of hazard perception clips

Hazard perception clips were recorded from a moving car in four different countries for various studies in this thesis. This section will provide details on how these clips were filmed, and what their content is.

2.3.1 Camera placement

The footage used for creating the hazardous videos was filmed following the Health and Safety guidelines approved by the Ethics Committee of Nottingham Trent University (see Appendix A). Four cameras were mounted on the car, capturing information coming from the forward view and footage that would normally be seen in the side mirrors, and rear-view mirror (see Figure 2.2). The forward-view camera was mounted via suction mounts to the windshield on the inside of the car in a way that it would not obstruct the view of the driver. The camera used for recording the rear-view mirror footage was placed inside the car on the rear window also via suction mount. Finally, the right and left mirror information was captured by cameras placed on both right and left side windows, on the outside, immediately below the side mirrors. These cameras were positioned to film what would be visible on the side mirrors. These externally mounted cameras were also tethered to the car to prevent loss in the event of a suction mount failing.

For the forward view, a GoPro HERO4 Silver Edition camcorder was used, recording in Full High Definition format (1080p, 16:9 ratio, medium-angle setting). The cameras used to capture the side-mirror and rear-view mirror information were JVC Action Cameras (Model Number: GC-XA1BU; 1080p, 16:9 ratio), (see Figure 2.3).

2.3.2 Filming UK clips

Twenty UK video clips were designed specifically for the purposes of this thesis. The driving footage was filmed in Nottingham and hazards were selected by two experts in the field. Footage was filmed during normal driving, in clear weather conditions and only during day time. Driving and filming was carried out by a native UK experienced driver with a full driving licence.

In order to capture a wide range of hazards, filming took place during several hours per day. Within the UK driving context an average of three valid hazards were captured in 60 to 90 minutes of driving. The available footage was carefully examined by two experts in the field and clips were selected, each containing a hazard defined as an object or event that would require an evasive manoeuvre, including sudden braking or swerving to avoid a potential collision (this definition has been followed to select the rest of the videos in this thesis). All selected hazards represented typical events that a driver could encounter on UK roads in every day driving. The most common scenarios were characterised by pedestrians crossing the road (see Table 2.1 for description of each of the 20 UK hazards, please refer to page 74). None of the scenarios selected for the UK clips were staged.

Figure 2.2: The graphic represents the positions of the four cameras. The green arrows show the placements of the front and rear-view camera and the blue arrows the positions of the left and right cameras. The side mirror graphs the amount of sky, road and car side captured by the cameras.

2.3.3 Filming Chinese clips

The Chinese clips were filmed specifically for a cross-cultural study that forms part of this thesis. The recording of the scenarios took place in and around Beijing. The amount of hours of filming necessary for capturing valid hazardous scenarios was considerably lower in comparison with the UK. An average of nine hazards were captured during one hour of driving. All hazards represented typical events that one could encounter within the Chinese driving environment. The most common event was caused by overtaking and undertaking cars, which seemed to not surprise Chinese drivers (as this behaviour was part of the social driving rules in Beijing). The continuous flow of pedestrians crossing the streets, independently of whether they have been given priority or not, was another common situation in Beijing (for a complete description of the Chinese hazardous scenarios, see Table 2.2, page 76). The driver who participated during the recording was a native Chinese experienced driver from Tsinghua University, and hazards were selected by two expert drivers in the field according to the definition used for the UK clips. Clips were selected by the same experts that have selected the UK videos. None of the hazards were staged.

2.3.4 Filming Spanish clips

The footage that constitutes the hazards for the Spanish hazard perception test was filmed in Granada, also specifically for a cross-cultural study that forms part of this thesis. Filming was carried out during normal driving in daytime with clear weather conditions. This footage was recorded in collaboration with the University of Granada and driving was carried out by a native Spanish expert driver.

The Spanish hazards were selected by the same two experts in the field, according to the definition of a hazardous situation used for the UK and Chinese clips. The selected hazardous scenarios were typical traffic events that any driver would find within the Spanish driving setting (for a full description of the Spanish hazards, see Table 2.3, page 78). In the case of the Spanish videos one hour of footage provided an average of 5-6 hazards, double the amount of hazards filmed in the UK within this amount of time. A common Spanish hazardous event was the weaving behaviour of mopeds, which were observed more often than in the UK clips. None of the Spanish hazards were staged.

a) b)

Figure 2.3: HD cameras employed for filming of the driving footage. On figure 2.3a is represented the GO PRO Hero (front footage) and figure 2.3b represents JVC action camera (left, right and rear view footage)

2.3.5 *Filming Israeli clips*

The Israeli footage was specifically recorded for the purposes of this thesis, in collaboration with Bar Ilan University. The recording took place during day time, with clear weather conditions. All videos were filmed within the radius of Tel-Aviv and Ramat-Gan (which is part of Tel Aviv and is only within 4.3 miles distance of the centre of Tel Aviv). Hazards selected for the Israeli HP test were typical hazards occurring during normal every day driving. The most common hazardous scenario were drivers cutting the path of other drivers while changing lanes.

The average amount of hazards captured within an hour was of 6-7 hazardous scenarios. The footage was filmed by a native Israeli expert driver. Clips were selected by the same experts in the field that have selected the videos for all of the above countries. None of the hazards for the Israeli test were staged (for a full description of the Israeli Hazards, see Table 2.4, page 79).

2.3.6 Which clips were used in which studies

The UK clips were employed for all the experiments in this thesis, except Experiment 8. The Chinese, Spanish and UK hazardous clips were compared together in Experiments 4 and 6 (with ten clips selected from each country). The Israeli hazardous clips were used for Experiment 8. In total, there were 26 Israeli clips: 13 hazardous and 13 quasi-hazardous scenarios. For Experiments 1 and 3, 20 UK clips were employed (ten of the clips were those used for studies 4 and 6). For Experiment 2, ten UK clips were used. Finally, for Experiments 5 and 7, 15 clips (part of the 20 UK clips) were edited into a hazard prediction test.

2.3.7 Editing, synchronisation and clip selection

Driving footage was edited and synchronised using Adobe Premiere Pro. The footage available from the cameras that captured the mirror information was imported into Adobe and synchronised with the forward view footage. A graphic overlay of the interior of a Ford Focus containing mirror placeholders was created in Adobe Photoshop and the synchronised footage was placed within this overlay (see Figure 2.4). The overlay was generated from internal photographs of the Ford Focus which were stitched together in Photoshop. The A-pillars and roof were designed to be transparent. It should be noted that the decision for transparent A-pillars and roof was submitted to examination (Experiments 2 and 3), as the design can be potentially criticised for not being realistic (i.e. real A-pillars are not transparent). However, this design was chosen to mimic the effects of stereopsis and small head movements that allow the driver to access otherwise occluded information (e.g. if one is first in a queue of vehicles at a red traffic signal, the nearest set of signals can easily be hidden by the roof, necessitating the driver to lean forward slightly to be able to look up at the red light). In order to address this design issue, two studies were conducted examining the influence that the transparent car interior has on driver's hazard perception compared to an opaque overlay (see Chapter 4) (see Figure 2.4 c and d for a reference).

During editing, image quality of some clips was slightly improved, with brightness being adjusted for the UK clips due to the cloudy weather which creates darker images when captured by the camera than the real-world. The forward view was captured on a medium width lens which required some anti-fish eye correction to be undertaken. The resolution of all clips was 1920x1080 with 30 frames per second, except for the Israeli videos, which were recorded at 50 frames per second.

Hazards were selected by two experts in Traffic Psychology. The average length of the videos was the following:

20 UK videos: 43750 ms

10 Spanish videos: 42000 ms

10 Chinese videos: 49400 ms

13 Israeli videos: Hazardous clips 20425 ms, Non-hazardous clips: 21648

(a) Chinese clip

(b) Spanish clip

(c) UK clip with transparent car overlay

(d) UK clip with opaque car overlay

(e) Israeli clip

Figure 2.4: Screenshots of the Chinese (a), Spanish (b), UK (c and d) and Israeli (e) hazardous videos. Image (d) contains an example of opaque graphic overlay applied for experiments 2 and 3

2.4 Apparatus

The equipment used for the experiments of this thesis was primarily either E-Prime 2.0 Software (Psychology Software Tools, 2012) or SMI 500Hz Remote Eye tracking Device. For Experiment 8, responses were collected via Qualtrics Software.

2.4.1 E-Prime 2.0

E-Prime 2.0 Software was employed for Experiments 4, 5, 6 and 7. E-Prime is a software package created for the design of experiments within the field of Psychology and Cognitive Sciences. The software offers different functions which allow the design of an experiment (E-Studio), extraction of data for analysis (E-DataAid), file merge (E-Merge) and experimental running (E-Run). All experiments were programmed specifically for this thesis. Stimuli were loaded into the software and the tool was programmed to record reaction time, ratings and written responses of participants, depending on the nature of the particular experiment. Participants responded either using the keyboard or the mouse. The E-Prime software was installed on a Lenovo computer with resolution of 1920x1080 and size of 34.37cm x 56.71 cm. The typical experiment was programmed in the following way: participants were asked to input demographic data (age, sex, years of driving, miles driven in the previous two years, accidents in the previous two years) and then instructions were displayed. Participants were presented with a practice trial where they could practice the task and ask any additional questions. Then the experiment took place where drivers were asked to watch the video clips and either respond to, or predict the hazardous situations. The software was programmed to record all responses automatically.

2.4.2 SMI 500Hz, Eye-tracking equipment

In order to record eye-tracking data, an SMI RED 500 was used. This included a suite of programs to control the eye tracked (I-View X), program and run the studies (Experiment Centre), and analyse the data (BeGaze). Eye-tracking data was recorded for Experiments 1 and 3. The sample frequency was of 500 Hz. The accuracy of SMI RED 500 is 0.4 degrees and the size of its headbox is 39.9cm x 20.1cm. The program additionally collected behavioural responses (e.g. response times to hazards, or selection of a multiple-choice answer).

Participants were typically positioned at a distance of 60 cm of the screen. A four-point calibration was used, where participants were asked to follow a dot until the infrared light captured their pupils and was able to calculate gaze location. On occasion, calibration needed to be undertaken more than once, as the system could not capture the pupil, or participants were not able to follow the point with their gaze in the correct way. If that was the case, they were given an example of how to easily follow the point and calibration was repeated until success. The size of the monitor on which the experiment was displayed was 49 cm x 29.5 cm, with resolution of 1600x900.

Analysis was conducted using the BeGaze software. For each hazard, an onset and offset were defined. In order to capture fixations, areas of interest (AOIs) were created for both precursors and hazards, mirrors and frames (see Figure 2.5). On occasion, the equipment was not able to capture some of the fixations, as they appeared as smooth pursuits. A smooth pursuit would occur when participants move gradually their gaze across the screen, while fixated on a moving object. If this movement occurs during an area of interest (AOI), the equipment does not recognise the fixation as such. In order to avoid losing valid participants, these fixations were identified and calculated manually, by subtracting the hazard onset from the entry of the fixation (the milliseconds in which this fixation started) and dividing the result by the duration of the hazard.

Figure 2.5: The two screen-shots depict the areas of interest for a typical hazard (top image), and left, right and rear-view mirrors and car frame (bottom image).

2.4.3 Qualtrics

Qualtrics is an online survey tool, created for research and data collection purposes. The tool is user friendly and offers more than 85 question types, including the uploading of multimedia footage. As Qualtrics typically collects data from participants who engage with the test outside laboratory control, there are a number of tools that allow one to check the validity of the data (e.g. how much time participants dedicate to a particular question; whether they clicked on a question page; whether questions were displayed to them etc).

Even with the superior design tools of Qualtrics, playing video and collecting meaningful behavioural responses is a challenge. For instance, simply embedding video in Qualtrics results in a progress bar appearing at the bottom of the video. On this basis, participants could then estimate when the clip was going to end (and thus temporally predict the hazard). Bespoke programming was required to overcome this limitation.

Also, videos were set to autoplay at the beginning of each question to avoid participants clicking on the multimedia and stopping it. The duration of videos was timed, which permitted control over the amount of time participants spent watching each video. Access to the settings of the survey and the data was only permitted to the researcher.

Although the data was collected online, prior to the actual experiment participants were shown instructions and were asked for their consent to participate.

2.5 Forward-Backward translation

As this thesis contains a number of studies conducted in other countries, it is important to ensure that the instructions provided to all participants are identical in meaning, despite being written in different languages. To ensure this, the instructions for all cross-cultural studies were subjected to forward-backward translation. The process of forward-backward translation contained several levels of building a conceptual understanding of the tool within each of the target cultures (China, Spain and Israel). It consisted of (1) forward translation, (2) expert panel (3) backward translation (4) final version (International Test Commission, 2010). The first step of the process was conducted by a native speaker of the target language, who was also a transport researcher (familiar with the terminology used) and highly proficient in the English language. The aim of the translation process was to emphasise the conceptual rather than literal translation. Particular English words and phrases do not have a direct equivalent in another language. For instance, to "undertake" does not have a direct translation in Spanish or Hebrew. This fact required a careful examination of the construction of expression in order to define the word in the best possible way.

Furthermore, some definitions also required consideration, as it was necessary to focus on the role a particular concept has within the target culture. The translated text had to be simple and clear, with the language adapted to a non-academic audience, avoiding jargon and terminology which are not familiar to the population outside academia. For instance, one of the key aspects of the hazard perception test is the definition of a hazard or a hazardous situation. It has been challenging to define the concept even within the original cultural context (Pradhan and Crundall, 2017).

Once the first step was completed, an expert panel was convened with bilingual members (English and target language). The aim was to discuss and resolve gaps in the translated version by identifying discrepancies between the new and original version of the tool. When consensus was reached, the test was translated back into English by an independent member not related to the field of work. During that part of the process, the key aspects and terms of the test were highlighted, and attention was given to concepts sensitive to translation across cultures. Discrepancies and confusions were discussed until the final version was satisfactory for all the members.

The next step was to conduct a pilot test of non-drivers. The aim of this step was to find out whether the final version was accessible to all socioeconomic levels. After the pilot, participants were asked about their impression and whether they were able to understand the language.

This Forward-Backward translation procedure was applied specifically for studies, 4, 6 and 8.

2.6 Instructions

Instructions were translated and adapted to each target culture. All participants received instructions in their native language. Prior to the experiment, participants were provided with an information sheet and were asked to sign a consent form. They were informed that participation in each experiment was voluntary and that they had the right to withdraw even if the study has already begun. They were assured that their data was completely anonymous and that only the researcher had access to it. They were given time to ask questions related to their participation before the actual experiment.

All participants received both verbal and written instructions (except for Experiment 8) prior to the beginning of the test. They all saw a practice trial prior to each experiment in order to familiarise themselves with the task and assure they have understood what they are required to do. Each practice trial contained a feedback section, where correct answers were provided. Feedback of performance was given only during the practice trial. If they did not understand the task or answered incorrectly to the practice trial, the researcher provided an additional explanation and feedback. The researcher would not start the actual experiment if participants reported doubts about any of the tasks.

2.7 Ethics

All the studies in this thesis were carried out in accordance with the Declaration of Helsinki and recommendations of the Health Research Ethics Committees of Nottingham Trent University, and were approved by the College Research Ethics Committee of Nottingham Trent University.

In order to develop a video-based test using realistic footage, it was necessary to carry out video-recording during normal driving using cameras attached to the windows of the car. To avoid any possibility of an accident, due to the driver not being familiar to the presence of cameras, driving and recording were performed by a researcher who is an expert driver and has already been trained in driving with cameras.

A risk assessment for on-road filming can be found in Appendix B. A protocol for filming was developed following guidelines from the Health and Safety Executive (see Appendix A) for filming from moving vehicles. Footage was only recorded from journeys that were required for reasons other than filming (i.e. no driving for the sole purpose of filming was conducted).

Informed consent to participate in the studies was obtained from all participants. Only the researcher had access to the data collected within this thesis. In order to anonymise the collected data every participant was given a Unique Identifying Code. Data collected from participants was kept separately from participants' personal details in order to ensure anonymity and confidentiality.

Participants were told that they are able to withdraw their data upon until the point of data analysis. After which any personal data (as defined by the Data Protection Act, 1998) would be destroyed, and only anonymised raw data and summary data will be retained.

This procedure was explained to the participants in the information sheet, prior to the experiment. The consent form confirmed that participants understood the procedure. Participants were informed that following analysis, summaries of non-personal data (demographics, responses in hazard perception test results) may be retained long-term as part of a larger data set and used for publication (unless they had asked for their data to be removed). After each experiment, participants were debriefed.

Table 2.1: Description of the UK clips

Nº	Content	Hazard	Duration of the clips (ms)
1	A car ahead overshoots a red traffic signal. As the cross traffic begins to enter the junction, the reversing light of the car turns on and the car ahead reverses towards you. The clip occludes following initial illumination of the reversing light.	The car in front reverses	64000
2	As you are driving, the car indicates and pulls over blocking your path. As you try to overtake the car and continue your path, a car from behind overtakes you (passing from the right mirror, into the blind spot, before emerging into your forward view). The clip occludes at the point where the overtaking car enters the blindspot.	The car on the right overtakes	34000
3	You are driving along a narrow urban street and the head of a pedestrian is visible above the parked cars on the left. The clip occludes as the pedestrian moves between the parked cars to step into the road.	A pedestrian from the left crosses the street	34000
4	You are driving in a street with shops and parked vehicles on the left side. There is a pedestrian coming out of one of the shops, approaching the street. The pedestrian steps out from between two parked cars on the left just as you accelerate after waiting in standing traffic. The clip occludes when the pedestrian turn his head to look at your car.	A pedestrian from the left crosses the street	49000
5	A distracted pedestrian is walking towards the street. The pedestrian crosses the road from the right without looking. The clip occludes when the pedestrian approaches the road in order to cross.	A distracted pedestrian crosses the street from the right	54000
6	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind undertakes you on the left on a multilane road Once past you, it cuts into your lane and is forced to brake suddenly due to a red traffic light. The clip occludes when the car is no longer visible at the left mirror, but a flash of it is visible on the left window.	A car from the left undertakes	43000
7	There are pedestrians on both side of a narrow street. A pedestrian with a child's push chair enters the road from the right without looking. Her entrance is partially obscured by pedestrians standing on the right. The clip occludes when the push chair is partially visible among the pedestrians.	A pedestrian with a pram crosses the street suddenly from the right	32000
8	You are driving along a narrow urban road. A pedestrian approaches from the right with shopping bags. They are temporally obscured by a large plant pot, before they step out into the road. The clips occludes just as they begin to step out from behind the plant pot.	A pedestrian from the right crosses the street without looking for upcoming cars	46000
9	A bus in a bus lane signals to pull away from a bus stop. Due to parked vehicles ahead in the bus lane, it pulls out into your lane forcing you to stop. The clip occludes following one flash of the indicator from the bus.	The bus in fornt cutts into the lane	31000
10	A car ahead emerges from a side road on the left and crosses your lane (too far away to be considered a hazard). It then indicates again and		

	immediately cuts across your lane once more to park in a layby. This second lane crossing is close enough to your vehicle to constitute a hazard. The clip occludes just before the car starts to cross your lane and following one flash of the indicator of the car.	The car in front turns to the left in the last moment	42000
11	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind overtakes your vehicle on a blind rural bend. The appearance of an oncoming vehicle in the opposite lane forces the overtaking vehicle to pull back into your lane immediately in front of you. The clip occludes when the car is no longer visible at the rear-view mirror and the right mirror.	The car behind overtakes at a blind bend	36000
12	While travelling at speed along a country road, a blind bend ahead reveals a queue of standing traffic, forcing you to slow and stop. The clip occludes immediately after passing the blind bend, when the brake lights of the cars ahead are partially visible	Queue of cars	61000
13	While your car is slowing due to congestion ahead, a pedestrian looks over his shoulder before deciding to run in front of your vehicle forcing to slow more abruptly than otherwise required. The clip occludes at the moment when the pedestrian is visible at the left pillar and is looking at your car.	A pedestrian crosses the street on the left	34000
14	Driving through a school area, two pedestrians are visible on the right. You can see that the woman is looking at the car. The clip occludes at the moment she gesticulates with her arm, thinking that you have given way for her and her child to cross.	Pedestrian from the right crosses the street	70000
15	You are driving along a narrow road with parked cars on both sides, which are obscuring the view ahead. Suddenly an oncoming van appears ahead. The clip occludes at the moment the white van is partially visible.	A van appears suddently behind parked cars	34000
16	While driving along a narrow urban road, it becomes visible a white van parked on the left side of the road. One of the doors of the cargo is open and a pedestrian (the owner of the van) is approaching. This clip was not part of the hazard prediction test.	Pedestrian approaching from the left	35000
17	You are driving along a narrow urban road and stop due to red traffic lights. The car in front aims to turn left, however the upcoming traffic is obstructing its way. The car in front stops even though the traffic lights turned green. This clip was not part of the hazard prediction test.	The car preceding you stops at green traffic lights	42000
18	You are driving along a residential area with pedestrians on the right side of the road. While driving along the road, pedestrians step into the road from the right side. This clip was not part of the hazard prediction test.	Pedestrians approaching from the left	46000
19	While driving the road narrows as you are approaching a road works ahead. Suddenly a pipe becomes visible from the left, hold by one of the workers. The pipe nearly touches your vehicle. This clip was not part of the hazard prediction test.	A worker carrying a large pipe	46000
20	While driving on a single carriageway, one of the upcoming cars signals with the intermittent and then makes a turn, crossing your path. This clip was not part of the hazard prediction test.	A car form the right crossing yur path	42000

Table 2.2: Description of the Chinese clips

N^0	Content	Hazard	Duration of the clips (ms)
1	A pedestrian is visible at the right edge of the road, looking to cross. The pedestrian is obscured by a turning car at the point of stepping into the road. By the time the pedestrian is visible again, he has already stepped into the road becoming a hazard. For experiment 2, the clip occludes just as the pedestrian starts to become visible as the obscuring car moves past.	Pedestrian steps into the road form the right	55000
2	There are parked cars on both sides of a narrow street that might occlude pedestrians. A pedestrian steps into the road in front of you, from between two parked vehicles. The clip occludes as the pedestrian first becomes visible stepping out from between the parked cars.	Pedestrian crosses the road from the left	57000
3	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A cyclist emerges from the side road, obscured by the parked vehicles, and makes a wide turn in front of your vehicle, before cycling towards you. The clip occludes as the front wheel of the bicycle enters the view.	A cyclist emerges from the left side street	26000
4	Your car slows on approach to a junction. A cyclist approaches from the left and is partially obscured by the A-frame of the semi-transparent graphic overlay. The cyclist cuts directly across your path. The clip occludes as the cyclist makes a change in direction to cut across your path.	A cyclist crosses the road from the left side	63000
5	Your car is driving slowly and there are parked cars on the right. A parked car on the right side of your lane indicates late before attempting to pull out in front of you. The clip occludes following one flash of the indicator from the manoeuvring car.	A parked car on the left joins the road	57000
6	A car immediately behind you, visible in the rear-view mirror and left side mirror, decides to overtake by entering a slip road to your left. It is forced to immediately pull back into your lane, in-front of you, as the slip road ends. The clip occludes when the car is no longer visible in the left mirror, but flash of it is visible in the left window.	A car from the left side attempts to undertake	31000
7	A car behind you is visible in the rear-view mirror and right side mirror. The car undertakes you on the left by entering a bus lane. Once past you, it cuts into your lane and is forced to brake suddenly due to slowing traffic ahead. The clip occludes when the car is no longer visible in the right mirror, but it quickly appears next to the right window of your car.	A car undertakes from the right	50000
8	A car behind you is visible in the rear-view mirror and right side mirror. While attempting to exit a multilane road, the car from behind accelerates to undertake your vehicle, forcing you to hold off moving into the desired lane. The clip occludes when the car is no longer visible in the right mirror.	A car undertakes from the left while exiting the main motorway	62000
9	A lorry approaches fast from the right-hand side. The lorry enters the main road from a side road on the right, cutting into your lane. The clip occludes at the moment in which the lorry is about to enter into the main road.	A truck joins the road from the right	38000

0	A car behind you is visible in the rear-view mirror and left side mirror. The car indicates with the front lights. Then undertakes via a bus lane at speed, immediately cutting in front of your vehicle and braking. The clip occludes when the car is no longer visible in the right mirror.	A car is approaching on the emergency lane from the right	55000

Table 2.3: Description of the Spanish clips

N^0	Content	Hazard	Duration of the clips (ms)
1	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A motorcyclist emerges from the side road, partially obscured by the parked vehicles, and enters the main carriageway immediately in front of your vehicle. The clip occludes when the front part of the motor is visible.	A motorcyclist emerges from the right	41000
2	A pedestrian is stood in the road next to a parked car waiting for her friend to exit the car. As your car approaches the driver's door of the parked vehicle opens. The clip occludes when the pedestrians stops next to the car.	The door of a car parked on the right suddenly opens	42000
3	On entering a side road, a rider on a scooter is checking over her shoulder in order to pull out around a vehicle blocking her lane. She then pulls out in front of your vehicle, as you finish turning into the side road. The clip occludes following one flash of the indicator of the scooter.	A scooter pulls out from the left	34000
4	While travelling on a dual carriageway, in the distance a pedestrian enters from the left side of the road. The pedestrian continues to cross the street, forcing you to slow and stop. The clip occludes at the moment when the pedestrian enters the road.	A pedestrian crosses the road from the left	37000
5	A car emerges from a side road and stops in front of you before indicating that it is going to reverse into a parking space at the road edge. The clip occludes following one flash of the indicator from the car.	The car in front stops suddenly	53000
6	A van is approaching fast from a side road on the right. The van tries to pull out and it halts abruptly when already partially out of the road, but nevertheless forces you to brake suddenly. The clip occludes when the van is approaching from the right and almost enters your lane.	A van appears abruptly form the right	48000
7	A car ahead stops on a zebra crossing due to congestion ahead. A pedestrian, unable to cross on the actually crossing, steps into the road slightly in advance of the zebra crossing. As she steps out she is partially obscured by parked vehicles on the right. The clip occludes just as the pedestrian first become visible.	A pedestrian approaches from the left	42000
8	A double-length ('bendy') bus in the right lane indicates and pulls off from a bus stop immediately in front of you, after you have just exited from a roundabout. The clip occludes when the bus turns to enter the road and following a flash of the indicator of the bus.	A bus from the right lane cuts into the right lane	50000
9	While driving on a dual carriageway, a motorcycle undertakes in the right lane and is forced to pull in-front of your car as traffic in the right lane slows due to congestion. The clip occludes when the motorcycle is no longer visible in the right mirror, but part of it is visible at the right window.	A car from the left undertakes	26000
10	A pedestrian is approaching from the left, partially obscured by a pillar. The pedestrian crosses the road ahead from the left. The clip occludes when the pedestrian is approaching the zebra crossing.	A pedestrian approaches from the left	47000

Table 2.4: Description of the Israeli clips

N^0	Content	Hazard	Duration of the clips (ms)
1	You are driving along a narrow residential road, with cars parked on the right side of the street. One of the parked cars turns on its brake lights and another car emerges from the left side of the street. The clip occludes at the moment when the front part of the emerging car is partially visible.	A car emerging from the left	16800
2	While driving you are approaching an underpass. You enter and there is a congestion of cars on both sides. The car on the right signals and cuts your path. The clip occludes with the first flash of the intermittent.	A car on the left side cuts your path	27866
3	You are driving on an urban dual carriageway road and there is an additional lane where there are parked cars. On the right mirror a taxi is visible and it undertakes your car just immediately before colliding with one of the parked cars. The clip occludes at the moment when the taxi is no longer visible at the right mirror.	A taxi undertakes form the right	40666
4	You are waiting due to red traffic lights. The road ahead diverse and there is a congestion of cars on your left side. On your right mirror is visible a van that is approaching towards the diversion. It quickly undertakes your car and squeezes between into your lane. The clip occludes at the moment when the van is no longer visible on the right mirror.	A van undertakes and cuts your path	22300
5	You are driving on a narrow urban road and there is a truck in front of your car. There are parked cars on the right side of the road and a pedestrian is approaching towards the street. The clip occludes at the moment when the pedestrian looks at your car.	A pedestrian steps into the road from the right	16166
6	While you are driving along a congested urban road the car in front diverse into the left lane and then suddenly brakes and cuts your path. The clip occludes after the car's intermittent emits a couple of flashes.	The car on the right cuts your path	21166
7	You are driving along the middle lane of a motorway while the right lane is heavily congested. One of the urban buses from the right lane cuts yours and other vehicles' path. The clip occludes when the bus has slightly started to turn left and have emitted a couple of flashes.	A bus from the right cuts my path	18333
8	You are driving along a residential area with pedestrians crossing the street. You are approaching a roundabout where there is a zebra crossing. Suddenly a cyclist appears from the left and crosses the street. The clip occludes when the cyclist is visible on the right window of the car.	A cyclist from the left crosses the street	19500
9	While driving along a narrow road, it becomes visible that the car in front has its emergency lights on, while still moving. The car then turns left into a side street and starts reversing back into the main road. The clip occludes at the moment when the car starts to reverse.	A car reversing back into to the main street	13000
10	You are driving along a congested urban road, all traffic lights are green. On the right-hand side, there are cars waiting to join the main road. A	A motorbike joins the main	19700

	motorbike appears from the right and approaches the main road without stopping at the red lights. The clip occludes at the moment when the motorbike is about to join your lane.	street from the right	
11	You are driving on a three-way carriageway. An urban bus from the right lane intends to join your lane. The clip occludes when the bus has started slightly to turn left and have emitted a couple of flashes.	A bus cuts your path from the right	14600
12	You are waiting while the traffic lights turn green. You start driving slowly as the traffic is congested. On the left mirror is visible a motorbike which approaches very fast and cuts your path. The clip occludes when the motorbike is no longer visible at the mirror and is visible on your left window.	A motorbike cuts your path from the left	19233
13	You are driving on a narrow road with parked cars on your right side. One of the parked cars is trying to join your path. The clip occludes at the moment when the brake lights of the car are visible.	A parked car from the right tries to join your way	16200
1	You are driving along a narrow street and the traffic lights turn red	No Hazard	31800
2	You are driving along a motorway and the two-lane road narrows into one lane	No Hazard	19733
3	You are driving on a busy urban road and a bus is waiting on your right	No Hazard	22200
4	You are driving along a narrow dual road and there is a cyclist in front of you	No Hazard	21300
5	You are driving along an urban road and the traffic lights ahead turn red	No Hazard	17800
6	You are driving along a busy urban road with pedestrians and cars on both sides	No Hazard	23100
7	You are driving along a busy road with cars constantly undertaking you	No Hazard	21100
8	You are driving along a motorway and there is a motorcyclist in front of you	No Hazard	17300
9	While driving the car in front of you brakes but then continues its way and joins the roundabout	No Hazard	21100
10	You are driving on a three-lane motorway with a busy left lane	No Hazard	25000
11	You are driving in a busy motorway with cars constantly braking in front of you	No Hazard	16900
12	You are driving along a three lane road and traffic circulates around you	No Hazard	23066
13	You are driving along an urban road and there is a lorry in front of you	No Hazard	21033

Chapter 3

DESIGN OF THE HAZARD PERCEPTION VIDEO CLIPS

Experiment 1

Does adding mirror information to the traditional hazard perception clips improve driving performance?

3.1 Introduction

The hazard perception test (HP test) has a long history of (mostly) successful discrimination between safe and unsafe drivers. It has been designed to measure hazard perception skill (HP skill) which has been argued to be the most reliable higher cognitive predictor of collision involvement (Horswill & McKenna, 2004).

The HP skill is a complex one, encompassing different basic processes such as perception, attention and decision making, however it has not been investigated to what extent each of these processes are related to the hazard perception skill. Instead HP tests have ostensibly been designed on the basis of pragmatics (as long as it discriminates between safe and less-safe drivers why worry about the intricacies of the underlying sub-processes?) with no consensus on a clear definition of what a hazardous situation is, which would help for a robust HP test protocol (Pradhan and Crundall, 2017). This is fine as long as the test works, but when a new HP test is found not to work, it is hard to understand why without recourse to a theoretical framework underpinning the skill.

Nonetheless, research in hazard perception has grown. Different research groups have developed their own tests, but without any scientific standardisation in terms of terminology, methods, metrics etc. While some researchers manage to replicate the basic experiential effect (Horwsill & McKenna, 2004; Wallis & Horswill, 2007), others have failed to find such differences (Crundall, Underwood & Chapman, 1999; Groeger et al., 1998; Sagberg & Bjørnskau, 2006).

More recently, new studies which have been conducted outside the UK and Australia (where the HP test has been most investigated) also failed to find differences between the driver groups on the basis of safety (e.g. collision history, driver experience; Borrowsky et al., 2010; Lim et al., 2013; Yeng & Wong, 2015). Due to these inconsistencies the validity of the HP test has been queried (Horswill & McKenna, 2004; Lim et el., 2013; Shahar et al., 2010) and the methodology used by the test has been questioned (particularly, the use of response times as a measure of performance; Jackson et al., 2009; Crundall, 2016). Unfortunately, it is difficult to identify why some tests work and others do not, as tests differ in many ways including presentation mode (e.g. still images, video, simulation), (Crundall et al., 2012; Horswill, et al., 2015; Pradhan et al., 2005; Scialfa, Pereverseff & Borkenhagen 2014; Wallice & Horswill, 2007), required response (e.g. timed response, localised response, predictive accuracy) (e.g. McGowan & Banbury, 2004; Crundall, 2016; Jackson et al., 2009; Wetton et al., 2010, 2011) and what constitutes a hazard (materialised vs. unmaterialised; hidden vs. unhidden; staged vs. naturalistic) (Catchpole & Leadbeatter, 2000; McKenna & Crick, 1991; Pradhan et al., 2009; Ventsislavova et al., 2016; Vlakveld, 2014)

At the start of this thesis, new hazard stimuli were collected and edited, taking advantage of recent advancements in digital video recording technology and video-editing techniques. Given the inconsistent findings, across research groups and countries, this gives rise to the first question of this thesis, namely whether the basic hazard perception effect (i.e. safer drivers responding faster than less-safe drivers) can be replicated with the new stimuli. The definition of safe and less-safe drivers is, however, a difficult one. Past-behaviour is often considered the best predictor of future behaviour, thus one might want to categorise participants based on crash history (Bem, 1980). Unfortunately, crashes are relatively few and far between, and those that drivers can remember may have been caused by factors other than their own skill level or risk-taking propensity. Near-collisions are more frequent, but memory for these can deteriorate quickly (Chapman & Underwood, 1998). One alternative that has frequently been used in the literature is to compare the performance of driver groups who differ on experience. Drivers within the first six months of passing their test have the highest probability of being involved in a collision, especially one that results in a fatality (McCartt et al., 2003; Mayhew et al., 2003). While some argue that poor attitudes to driving risk are to blame, there is considerable evidence to suggest that they simply have not yet developed the relevant skills for safe driving, both in general and specifically in relation to hazard perception (Foss et al., 2011; Williams & Tefft, 2014). Driving style is also closely related to traffic accidents and could be a steady predictor of driving behaviour (reckless and careless drivers are more likely to involve themselves in dangerous driving behaviour) (Taubman-Ben-Ari, & Skvirsky, 2016). However, this chapter is going to focus only on driving skill as it more closely associated to driving performance and information processing (while driving style is associated to driving behaviour). The primary aim of this study was therefore to assess whether the basic experiential effect can be evoked with new stimuli filmed and edited for this thesis.

3.1.1 The impact of mirror information

Typically, the video clips that are used in a traditional HP test only contain visual information related to what happens in front of the vehicle (filmed from a single forward-facing camera mounted internally or externally to a car). This makes it difficult to detect undertaking and overtaking hazards. This fact restricts the type of hazards that can be selected for the test. Although this approach has been criticised (Alberti et al., 2014; Shahar et al., 2010), the UK Government has not indicated any desire to improve the field of view available to HP candidates in the HP test (only the quality of the clips has improved to computer generated imagery CGI). This may be a particular problem for hazard tests in developing countries where there may be a greater amount of overtaking and undertaking hazards (Lim et al., 2013).

There are two easy ways to expand the field of view available in a simple hazard perception test: widen the available forward view (e.g. adding flanking screens to a central display or possibly displaying video in a VR headset), or by adding mirror information to the forward view, allowing the driver to see behind the film-car. Shahar et al., (2010) adopted (and confounded) both methods. In one condition they presented participants with either a single forward view (though including rear-view mirror information), or a three-screen forward view (which also included side mirror information on the left and right screens). They found that responses to hazards were significantly faster when participants saw the three-screen information, though interestingly this included hazards that solely appeared on the central screen. The authors suggested that the greater visual field may increase overall arousal leading to faster responses to hazards that should not ostensibly benefit from a wider field of view. While it appears that a greater field of view had a positive effect on hazard perception reactions, it is impossible to know whether this was due to the wider forward view, or the provision of side-mirror information.

Allen et al., (2005) tested three different types of screen in a simulator study (a single screen, three-screens and a large three screen display, with the last two including mirror information) in an instrumented car cabin. They found that novice drivers behaved differently across the three conditions with more aggressive and "unrealistic" behaviour during the single screen condition (which possibly encouraged an underestimation of risk and over-estimation of the HP skills). Unrealistic and less immersive contexts could evoke less realistic patterns of behaviour, tempting drivers to engage in less realistic behaviour (and they may be less likely to be looking in the appropriate location). Again, however, we do not know whether the inclusion of mirrors or an expanded field of forward view had differential effects on performance.

This field-of-view limitation in the design of the HP test has raised the second question of this study: will the provision of mirror information improve the efficacy of hazard tests? It could equally be asked whether expansion of the forward view improves performance, however this thesis will constrain itself to the technology that is used in current testing centres. Adding mirror information requires effort in the generalisation of the stimuli (see Chapter 2) but does not require any additions to the hardware required to run the test. Conversely, increasing the forward field of view would require the use of VR headsets or multiple-screen set-ups, which makes testing more difficult. This thesis has adopted the approach of designing tests for the simplest possible testing set-up, ideally requiring no more equipment than is currently available in UK driving test centres for the official hazard perception test. It is predicted that provision of mirror information may influence the performance gap between safe and less-safe drivers. Given the paucity of information in the literature, no prediction is made about the direction of this effect in terms of widening or narrowing the gap between experienced/safe and novice/less-safe drivers.

3.1.2 The impact of adding a car interior

While the mirror information could be placed on a forward video view as picture-in-picture, the decision was made to create a graphic overlay of a car interior, with place holders to house the mirror information. This provides a context that allows viewers to understand what the additional video streams reflect (i.e. mirror information), and where to find it. The addition of a car interior may, however, have an independent effect on hazard performance. Not only might the inclusion of the overlay increase immersion (and perhaps make visual behaviour more realistic), the car interior might bring the hazards perceptually closer: without a graphic overlay, the edge of the screen forms the 'windscreen', however, the smaller windscreen area of the graphic overlay means that hazards take up a greater percentage of the windscreen, and may therefore feel closer, and thus more hazardous (see Figure 3.1). In order to assess whether the addition of the car interior itself has any effect on hazard perception, a separate condition included clips with the overlay but without mirror information.

3.1.3 The role of eye movements in hazard perception

The few studies that have investigated the impact of expanding the visual field of view on hazard perception performance have not measured eye movement information or, if they did, the role of experience was not clearly discussed (differences between experienced drivers have not been reported e.g. Shahar et al., 2010). This is a surprising omission given the importance of eye movements during driving, and the development of visual skills as driving experience increases.

The differences between visual strategies of experienced and novice drivers are crucial in terms of collision liability. It has been widely demonstrated that when the roadway becomes more demanding, experienced drivers tend to fixate hazards faster than novices and that novices show longer fixations to hazards (Chapman & Underwood, 1998; Chapman, Van Loon, Trawley & Crundall, 2007; Crundall & Underwood, 1998; Underwood, Crundall & Chapman, 2002). This is due to the less-experienced drivers requiring more time to find, and then process, the hazard. Pollatsek, Fisher & Pradhan, (2006) have reported several studies where novice drivers clearly fixated crucial risky regions of the driving context less frequently than experienced drivers. Similar results were found by Crundall et al. (2012) where novices showed less attention to the critical stimuli. This suggests that novice drivers are not fully aware of the hazardous situation (Crundall et al., 2012), as their fixation durations should increase when confronted with a hazard. Less experienced drivers have not yet developed their hazard perception skill (due to the lack of experience with hazardous situations) enough to be able to meet the demands of the road environment. Thus, novices may need more time to process potential hazards, and therefore maintain their gaze close to the vehicle and central to the view (Alberti, Shahar, & Crundall, 2014; Crundall & Underwood, 1998; Konstantopoulos et al., 2010; Mourant & Rockwell, 1970). In turn, experienced drivers know where to look and what features to prioritise, showing flexible strategies and adapting their visual search according to the requirements of the visual scene (Crundall & Underwood, 1998).

In regards to eye movements specifically to mirrors, research has demonstrated that glances at mirrors do vary with driving experience. For example, Konstantopoulos et al., (2010) found that driving instructors fixated side mirrors significantly more than learners, which is consistent with another previous study by Underwood, Chapman, Bowden & Crundall (2002) where experienced drivers inspected left and right mirrors more than novice drivers. It has been argued that this is due to the different (erroneous) prioritisation strategies that novices use (Konstantopoulos & Crundall, 2008) and even though Olsen et al., (2007) showed that novice drivers are able to learn to inspect more near-side and rear-view mirrors in the first six months, they still had fewer glances in comparison to experienced drivers in some conditions, such as performing secondary tasks (e.g. in-vehicle tasks).

Given research that suggests (a) inexperienced/less-safe drivers should be worse at hazard perception, and (b) they are less likely to use mirrors in some instances, it suggests that any such effects found in this study might be explained, at least partially, by the eye movements of drivers. Accordingly, participant eye movements were recorded using a desktop eye tracker to assess whether the inclusion of mirrors evoked changes in eye movements (perhaps more so in experienced/safer drivers than novices).

3.1.4 Overtaking and undertaking hazards

The present study is also concerned about overtaking and undertaking hazards, and, similar to Shahar et al. (2010), examines whether the mirror information will improve response times to these hazards. There is very little literature that has looked particularly at these types of hazards, although they are considered one of the most dangerous scenarios (Rameau et al., 2016).

Shahar et al. (2010) mentioned the importance of including overtaking hazards in the official HP test given the propensity of such scenarios to contribute to crash statistics, however in their study they did not measure eye movements and differences between experience groups were not reported.

The current study included a variety of hazard types across a selection of edited video clips, though three of these clips included overtaking or undertaking hazards. Though these hazards reflected only 15% of the total hazards in the study, it was considered that these hazards might benefit most from the inclusion of mirror information. Accordingly, this subset of hazards was subjected to additional scrutiny during the analyses.

3.1.5 Overview of the study

To summarise, the current study was interested in answering a number of questions:

1. Could the basic HP experiential effect be replicated using newly recorded and edited footage, designed specifically for this thesis? Will experienced drivers react faster to the pre-defined hazards and will they respond faster to these hazards once they have fixated them?
2. Does the provision of mirror information impact on the efficacy of the hazard perception test? Will mirror information enhance performance on the hazard perception test?
3. Does the mere inclusion of the car interior influence hazard perception performance?
4. Can eye movements help explain any of the above effects? Will experienced drivers fixate hazards faster than the novices?

5. Is perception of undertaking and overtaking hazards influenced by the provision of mirror information? Are these hazards detected faster in the mirror condition than in the forward view condition?
6. Do experienced driver show wider spread of search in comparison to the novices?

The present study compared three types of video clip design: traditional HP forward view video clips, clips including car overlay (but without mirror information) and clips including left, right and rear-view mirror information embedded in the car overlay (see Figure 3.1).

While it was possible to create a fully orthogonal design (including a fourth condition where mirror information was presented without the car interior) this was rejected. The reason for this was, the car interior acts as a placeholder for the mirror information, demarcating it from the background and placing it in context. Providing mirror information without the car interior acting as a placeholder, would result in odd patches of embedded video which may confuse participants.

Experienced and novice drivers viewed one of three sets of clips, and pressed a button as soon as they thought they saw a hazard in the scene (e.g. a car emerging from a side road, a pedestrian stepping out from behind a parked car). It was predicted that experienced drivers will fixate and respond faster to the hazards, and will be more accurate than the novices. It was also predicted that the different conditions (normal view, car-interior view, mirrors view) will impact performance with the mirror condition reporting the highest scores. Specifically, it was expected that during the mirror condition overtaking hazards will be spotted faster than in the other two conditions. Additionally, while novices were expected to explore less the driving environment than the experienced drivers, we did not make any predictions about the amount of extra hazard responses each group will report.

3.2 Method

3.2.1 Participants

Seventy-three participants were originally tested for this experiment, however 16 were excluded due to poor eye calibration. The final sample was composed of 57 participants, with 28 classed as experienced drivers and 29 as novices (19 of these were learner drivers). Participants were divided into experienced and novice drivers using the following criteria: experienced drivers were those who passed their driving test at least two years before the experiment and have driven more than 500 miles the previous year of testing. Novice drivers were those who were either learners, or participants that had passed their test in the previous 12 months, and had driven at minimum 30 miles. Mean mileage for the experienced drivers was calculated considering the miles driven in the previous two years prior to the experiment. For novices, miles since passing their test were used (as they had less than one year of driving experience). In the case of learner drivers, hours of practice were converted into miles and averaged (one hour of practice was equivalent to 20 miles). This criterion has been followed for all the studies in this thesis. The mean mileage for the experienced drivers (mean age=27.2) was 6311.8 miles and for the novices (mean age=21.4) 489.7 miles. The average years of driving for the experienced groups was of eight years and for the novices, four months (post-licence). All participants had normal or corrected to normal vision. The sample was composed of university students and staff, and visitors to the university.

3.2.2 Design

A 2 x 3 between groups design was used for this study, where the factors were the level of experience (experienced vs. novice drivers), type of video design (forward view vs. car layout vs. mirror view) and type of hazard (overtaking vs non-overtaking). Driver's fixations to the hazards, reaction times and extra hazard responses were measured as dependent variables.

The study had three experimental conditions: the first contained 20 hazardous video clips filmed from the driver's perspective where participants could see only the forward view of the driving situation (see Figure 3.1a). These clips were designed similarly to the traditional hazard perception test. For this first condition only one camera was used, attached to the front windscreen of the car. The second condition contained identical video clips as the first one, however in this case a car layout was added to the video image (see Figure 3.1b). The third condition included both the car interior and mirror information: left, right and rear-view mirror footage was added to the forward view image (see Figure 3.1c).

In order to calculate response times (RTs) for the hazards, hazard onsets and offsets were defined for each clip. Onset was defined as the point in which the hazard starts to develop and could eventually pose a threat. This is the moment when a precursor turns into a hazard (e.g. a car ahead begins to edge out of a line of standing traffic in front of the film car; a pedestrian steps off the sidewalk, etc.), and offset was defined as the moment where any later responses would not avoid a collision (e.g a pedestrian already crossing the street which would have required a corrective action by one of the road users to avoid a collision). Each hazard has been considered individually, as the type of hazard varied across different clips. Both onsets and offsets were defined by two experts with several years of experience in the field. Hazard onset times for each clip were subtracted from button-press times to give the RTs. Where participants failed to make a response during a particular clip, they were assigned a maximum response time, equivalent to the hazard offset (following McKenna et al., 2006). To minimize skew in the data a square-root transform was undertaken. The transformed RTs were then standardised into Z-scores using the overall sample mean and standard deviation (SD) for each hazard. This process was necessary because the hazard windows varied in duration, and without standardisation some hazards might exert a greater influence on the final mean than others (following Wetton et al., 2010). While the analysis was conducted on these z-scored, square-root transformed RTs, for clarity of presentation in graphs these figures were converted back into millisecond response times using the mean and standard deviation across all hazards and participants.

Eye movement information was recorded and analysed via SMI Experiment centre and BeGaze software. Fixations recorded by the software that were less than 80ms were not included in the analyses. A threshold of 80ms was chosen in order to exclude any brief fixations that are unlikely to have processed anything. Areas of interest (AOI) were allocated for each of the hazards (starting with the pre-defined hazard onset and finishing with the pre-defined hazard offset; the difference between the two is termed the *hazard window*) and mirrors (for the left, right and rear-view mirror). A rectangle was drawn on each hazard which allowed the AOI to follow the hazard through each frame without missing any of the hazard movements (see Figure 2.5). Following, AOIs for the mirrors were delimitated by a small rectangle.

Participants were randomly allocated to one of the three conditions. Following data loss, this left 21 (ten novices) participants in the forward-view condition, 19 (nine novices) in the car-interior condition and 17 (ten novices) in the mirrors condition.

3.2.3 Materials

The hazard perception test created for this study was comprised of twenty video clips. The video clips were composed of footage filmed in the United Kingdom and were edited in a hazard perception test using Adobe Premiere (see Chapter 2).

The forward view was recorded with a mini HD video camera attached to the inside of the front windscreen and the rear-view mirror footage was recorded via camera attached to the inside of the rear window. The additional side-mirror footage was recorded using cameras attached to the side windows. The mirror footage was synchronised with the forward view and edited into mirror placeholders that were contained in a graphic overlay of the inside of a car. The overlay was designed to be transparent from half way up the A-pillars, allowing the forward view to be seen through these sections of the overlay. Full occlusion of the A-pillars was considered to be too severe, and A-pillar occlusion in the real world can be offset by stereopsis and small head movements (for more information, see Chapter 2). All video clips included one pre-defined hazard (see Table 2.1 for description).

Clips were displayed on a monitor with resolution of 1920x1080 and screen size of 49 cm x 29.5cm. Eye movements were sampled binocularly at 500Hz with an SMI eye tracker computer, and only fixations higher than 80 ms were considered for the analyses. Participants were free to move their head and responded using a mouse and keyboard connected to the laptop.

3.2.4 Procedure

Prior to the experiment, participants received both verbal and written instructions. They were asked to fill in demographic information about their age and driving history (years of driving, miles driven and accidents during the previous two years). Then they were allocated randomly to one of the three conditions (all conditions contained the same hazardous video clips, which appeared in a randomised order). Once allocated, they were asked to sit on the chair in front of the screen as comfortably as they could in order to begin calibration and record their eye movements. They were asked to follow a dot on the screen with their gaze until calibration was completed.

During the study, prior to each video, a small fixation cross appeared on the screen for a second. Participants saw 20 clips and during each clip they were asked to click a button every time they thought they saw a hazard (a hazard was defined as an object or situation where they had to perform an evasive manoeuvre, such as braking or swerving, in order to avoid a collision). Participants were instructed that there was only one correct hazard. Correct identification of a hazard was defined as a response that fell within a temporal scoring window for each hazard. Hazard onset times for each clip were subtracted from button-press times to give the RTs. The experiment took approximately 40 minutes. After the experiment, participants were provided with feedback of their performance (they were given the opportunity to see their own eye movements).

Figure 3.1: Three screenshots of the video clips representing the three different conditions of the experiment. The top panel (3.1a) depicts the video showing only the forward view of the driving scene, the middle panel (3.1b) represents the forward view of the same video, however with a graphic overlay and the bottom panel (3.1c) shows the clip including mirror information

3.3. Results

3.3.1 Eye movement measures

Data were subjected to a 2 x 3 ANOVA where the independent variables were the type of video clips and driver experience, and the dependent variable was the fixations of participants (time to first fixate, and dwell time on hazards).

3.3.1.1 Fixations on hazards

The time taken to first fixate a hazard was calculated by subtracting the time at which a first fixation fell within the hazard AOI and the time of the hazard AOI onset. To account for variation in the hazard window length, this measure was recorded as a percentage of the hazard window at which the hazard was first fixated. For example, a fixation that landed on the hazard after one second of a three second hazard window would therefore score 33%. Zero % would represent a fixation at the very start of the hazard window, while 100% reflects no fixation on the hazard. Smaller percentages reflect smaller times to first fixate the hazard. The total time spent looking at the hazard was calculated considering the total proportion of time that drivers spent looking at the hazard. Then these fixations were converted into percentages by multiplying them by 100. Fixations were measured separately for each of the three conditions and for experience. Missing values (when participants did not fixate the area) were assigned the maximum 100% (i.e. they did not look at the hazard during the scoring window at all).

A 2 x 3 ANOVA was conducted to compare experience vs. the three conditions (forward view, car interior and mirrors) as independent variables. No main effect was found for time-to-first-fixate to hazards for both the experienced groups (M=14.7% for experienced drivers and M=17.3% for novices), ($F(1, 51)=2.66$, $MSe=52.1$ $p=.10$, $\eta_p^2=.05$) and between the conditions (forward view M=15.4%, car interior M=18.8%, mirrors M=13.7%) ($F(2, 51)=2.73$, $MSe=52.1$ $p = .07$, $\eta_p^2=.09$). Experienced drivers did not fixate the pre-defined hazards faster than the novices. However, there was a significant interaction between condition vs. experience ($F(2, 51)=3.95$, $MSe=52.1.9$ $p<.05$, $\eta_p^2=.13$). During the car interior condition novices were considerably slower at fixating the pre-defined hazards, compared to the experienced drivers ($t(17)=-2.50$, $p<.05$), (see Figure 3.2).

Finally, there were no differences between the experienced and novice drivers for the total time spent on fixating the hazards (experienced M = 47.8, novices M=45.5), ($F(1, 51)=.73$, MSe=183.9 $p=.39$, $\eta_p^2=.01$). No differences were found between the conditions either (forward view M=48.7, car interior M=42.5, mirrors M=49.3) ($F(2, 51)=1.44$, MSe=183.4 $p=.25$, $\eta_p^2=.05$) with no significant interactions ($F(2, 51)=1.75$, MSe=183.4 $p=.18$, $\eta_p^2=.06$)

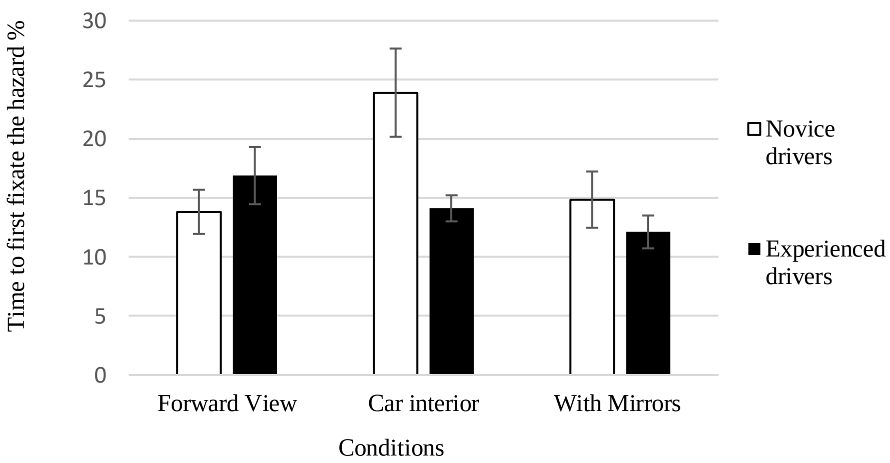


Figure 3.2: Percentages of time to first fixate the hazard across experience and experimental conditions (with standard error bars added)

3.3.1.2 Fixations on mirrors

Fixations on mirrors (outside the hazard window) were summed and then divided by the length of the clip (less the length of the hazard window), before being multiplied by 100 to create percentage dwell on the mirrors. The reason for subtracting the length of the hazard window was that in some occasions the AOI was defined within one of the mirrors, thus confounding the reasons that one may look at the mirrors. Only the third condition of this experiment contained mirror information and the analysis was conducted solely on participants who took part in this condition (seven experienced and ten novice drivers). A 2 x 3 ANOVA was conducted comparing groups across the three mirrors in terms of dwell. A main effect was found for mirrors ($F(2, 30)=18.05$, $MSe=175.3$ $p<.001$, $\eta_p^2=.55$) where pairwise comparisons showed that participants were checking the rear-view mirror significantly more in comparison to the right mirror ($p<.005$) and the left mirror ($p<.05$). Differences were also observed between the right and left mirror, with participants gazing at the right mirror significantly more ($p<.005$). This is due to the fact that in the UK traffic circulates on the left-hand side and hazards are expected to come from the right. No main effect was found for experience ($F(1, 15)=.01$, $MSe=.06$ $p=.90$, $\eta_p^2=.00$), and the interaction was not significant.

3.3.1.3 Spread of search along the horizontal and vertical meridians

This analysis was conducted in order to find out whether there were differences in participants' fixations spread along the horizontal and vertical meridians. For this end, pixel information was extracted and in order to calculate the differences of fixations co-ordinates, standard deviation was used. No differences were found for experience group ($F(1, 51)=1.24$, $MSe=821.89$ $p=.27$, $\eta_p^2=.02$) and between any of the conditions for the horizontal search ($F(2, 51)=2.66$, $MSe=821.89$ $p=.08$, $\eta_p^2=.09$). However, there was a main effect for the vertical spread of search between the conditions ($F(2, 51)=22.7$, $MSe=4969.3$ $p<.001$, $\eta_p^2=.47$), with the forward view showing more vertical spread of search in comparison to the car interior and mirror condition. This result is probably due to the fact that the forward view condition did not contain a car overlay, and therefore more of the visual scene was available to look at.

3.3.1.4 Overtaking hazards and mirrors

Three separate analyses were conducted to look at fixations during the mirror condition for the three videos that contained overtaking hazards, videos, 2, 6 and 11 (see Table 2.1 for description). Again, only participants assigned to the mirror condition were considered for this analysis. The aim was to assess whether the mirror information would benefit the experienced group in terms of faster detection of hazards (as hazards could be anticipated looking at the mirrors). Experienced drivers were not faster to fixate the hazards and no differences were found for any of the videos: video 2 ($F(1, 15)=.22$, $MSe=1.19$ $p=.64$, $\eta_p^2=.01$), video 6 ($F(1, 15)=.59$, $MSe=4.92$ $p=.45$, $\eta_p^2=.03$) and video 11 ($F(1, 15)=.52$, $MSe=7.27$ $p=.48$, $\eta_p^2=.03$). Finally, no differences were found between the experienced groups for glances at mirrors in these three clips; experienced drivers did not appear to rely on mirrors more frequently than the novices.

3.3.2 Behavioural responses to hazards

3.3.2.1 Response time to predefined hazards

Response times to hazards were analysed in order to find out whether experienced drivers would spot hazards quicker than novices. The data were subjected to 2 x 3 ANOVA where condition and experience were the independent variables and RTs or accuracy the dependent variables. No main effect was found for response times: experienced drivers did not react faster to the pre-defined hazards than the novices (M=2349.4ms for experienced, M=2219.8ms for novices), ($F(1, 51)=.06$, $MSe=.11$ $p=.42$, $\eta_p^2=.01$). Faster responses to hazards did not appear to benefit from the mirror information either, with no differences between any of the conditions (M= 2255.4ms for forward view, M= 2249ms for car interior M=2356.6ms for mirrors) ($F(2, 51)=.29$, $MSe=.05$ $p=.75$, $\eta_p^2=.01$).

3.3.2.2 Accuracy

Accuracy (i.e. the percentage of hazards that participants responded to within the hazard window) was also calculated. A button press response to indicate that there was a hazard within the frames of the time window was considered a hit. No main effect was found for either the experience groups (M=64.8% for experienced, M=70.2% for novices), ($F(1, 51)=2.09$, $MSe=531.5$ $p=.15$, $\eta_p^2=.03$) or conditions (M=68.1% for forward view, M=70.3% for car interior, M=63.8% for mirrors), ($F(2, 51)=1.02$, $MSe=259.7$ $p=.36$, $\eta_p^2=.03$).

3.3.2.3 Extra responses to hazards

Extra responses to hazards were also recorded (i.e. responses to potential hazards above and beyond the predefined hazard window, or secondary clicks within a hazard window following an initial response). As clips varied in duration time, the sum of these extra hazard responses for each participant was divided by the duration of the experiment creating a ratio measure of extra-hazard responses. No differences were observed for the experienced groups (M=1.79 for experienced, M=1.84 for novices), (F(1, 51)=.00, MSe=.00 p=.99, η_p^2=.00) or across conditions (M=2.09 for forward view, M=1.51 for car interior, M=1.81 for mirrors), (F(2, 51)=1.19, MSe=1.77 p=.31, η_p^2=.04). The interaction was not significant.

3.3.2.4 Response times after first fixation

The aim of this analysis was to measure whether there were differences between the experienced groups regarding their response times after first fixating the hazard. This analysis is different to those looking at reaction times to hazards, as here the aim is to find out how quick participants are at reacting to hazards once they have first fixated them. However, no differences were found between the experienced groups (M=17.1 for experienced drivers and M=17.5 for novices) (*F*(1, 51)=.05, *MSe*=2.03 *p*=.82, η_p^2=.00) or conditions (forward view M=15.2, car interior M=19.1, mirrors M=17.5) (*F*(2, 51)=1.85, *MSe*=37.8 *p*=.17, η_p^2=.07). No significant interactions were observed either (*F*(2, 51)=.17, *MSe*=37.8 *p*=.84, η_p^2=.00)

3.3.2.5 Response times to overtaking hazards

Videos that contained overtaking hazards were analysed separately to find out whether the additional mirror information influenced response times. The mirror information did not appear to influence hazard detection across the different conditions (F(2, 51)=.25, MSe=1.27 p=.77, η_p^2=.01) or between the groups (F(1, 51)=.81, MSe=1.27 p=.37, η_p^2=.02). No significant interaction was found between videos and conditions either (F(4, 102)=.16, MSe=.90 p=.95, η_p^2=.00)

Figures 3.3a, b, c. illustrate the frequency distribution of the clicks across the different conditions for each of the three overtaking clips during the pre-defined hazard window. As can be seen from these graphs, there is no clear advantage for the mirror condition.

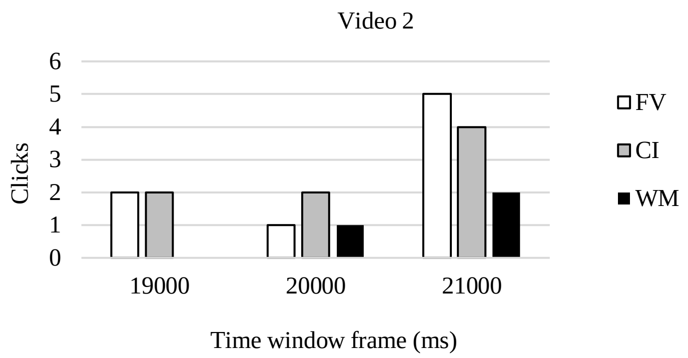


Figure 3.3a: Frequency of clicks across conditions for video 2 (FV = forward view, CI = car interior, WM = with mirrors)

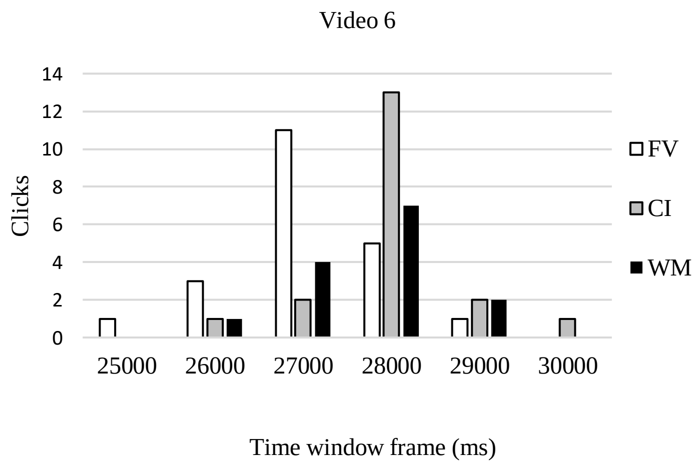


Figure 3.3b: Frequency of clicks across conditions for video 6 (FV = forward view, CI = car interior, WM = with mirrors)

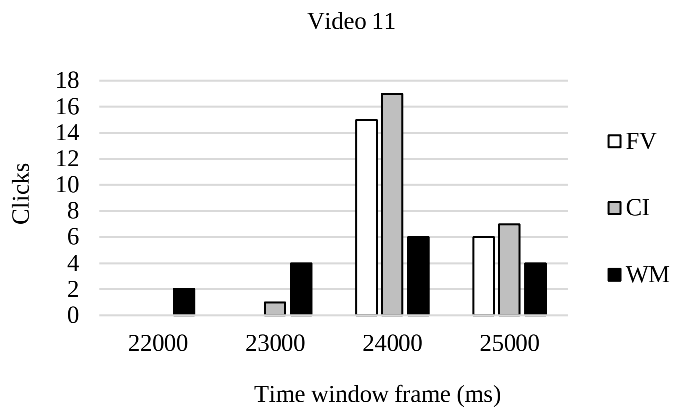


Figure 3.3c: Frequency of clicks across conditions for video 11 (FV = forward view, CI = car interior, WM = with mirrors)

3.4 Discussion

The aim of this study was to examine hazard perception performance using three different types of video design. It was hypothesised that participants will spot and react to hazards faster during the mirror condition and that the basic hazard perception experience effect will be replicated across all conditions with experienced drivers outperforming novices.

None of these hypotheses were accepted and the present study failed to find significant differences for both fixations to hazards and RTs across the three conditions and for the experienced groups. The experienced drivers did not react faster to the pre-defined hazards, nor fixated them quicker than the novices. Neither, it seems, did mirror information enhance (or degrade) hazard perception performance in comparison to the other two conditions. The present results contradict most of the previous literature which has featured eye movements, as the majority of studies have reported clear differences in relation to the visual search pattern between experienced and novice drivers (e.g. Chapman et al., 2007; Crundall & Underwood, 1998; Crundall et al., 2012; 2003). Less experienced drivers are known to detect fewer hazards than their more experienced counterparts (Fisher et al., 2006; Pradhan et al., 2005) and show longer fixation durations while looking at dangerous objects (Chapman & Underwood, 1998). The fact that there were no differences between the experienced groups could suggest that both groups are equally good at spotting hazards. In fact, novices had higher accuracy rates than experienced drivers (although these results were not statistically significant). Does this mean that novices are as good as experienced drivers in detecting hazards? Interestingly, there was a significant interaction between experience and condition, where novices fixated hazards considerably slower during the car interior condition while experienced drivers showed faster fixations. As novices are more prone to maintain a narrow central search of pattern (Crundall & Underwood, 1998), it is not surprising that they were slower at this condition which possibly confused them, while experienced drivers might have been expecting hazards from other possible locations (being more vigilant due to the missing mirror information). This suggests that the car interior addition to the design of the video may make novice's performance somewhat worse. The mere inclusion of a car interior appeared to negatively influence novices' performance by slowing their responses, but it does not seem to affect experienced drivers. It is possible that the inclusion of an interior overlay *without* mirrors created confusion for the novices. This finding supports the idea that video design should be as realistic as possible, as it otherwise could impact negatively participant's responses.

The fact that no differences in performance (for RTs) were found between the mirror vs. no mirror conditions did not suggest though that mirror scenarios restricted or hindered hazard detection. RT means for both mirror and forward view condition were very similar, although means for accuracy were slightly higher for the forward view conditions (but none of these results were significant). Furthermore, the results showed that participants did make use of the mirrors during the experiment. Both rear-view and right mirrors were consulted the most with the rear-view mirror receiving the highest amount of fixations, while the left mirror received the least attention. This pattern reflects previous studies where participants have favoured the offside and rear-view mirrors to the near-side mirror (e.g. Olsen et al., 2007). This result suggests that the mirrors were not used randomly but the information they provided was taken into account, as most of the hazards were coming from the right-hand side (UK driving circulates on the left).

The basic hazard perception effect was not replicated in this study, however this is not the first study that failed to find this effect (Crundall et al., 1999; Groeger et al., 1998; Sagberg & Bjørnskau, 2006; Yeng & Wong, 2015). The traditional hazard perception test has been criticised in terms of validity and whether the scores of the test actually reflect the hazard perception ability. It has been argued that the simple behavioural measure (response within a limited temporal scoring window) could confound the results as there are no clear guidelines on how to define hazard onsets and offsets (Pradhan & Crundall, 2017).

However, it should be noted that the present results could also be due to the sample size which may explain the failure to find an effect by producing false negatives. Doubling the sample size might have yielded this basic experiential effect.

The sample of the present study was not based on self-reported crash data as the studies that have found differences between drivers that are more or less prone to be involved in traffic accidents (Drummond, 2000; Horswill et al., 2010; Wells et al., 2008). The sample in the present study was divided using the number of years since passing the driving test and controlling for miles driven in the previous two years of the experiment. As it is not always possible to use objective collision data (as most of the data is based on self-report crash rates which are subject to memory distortion and self-presentation bias), numerous studies have relied on the years of driving experience. When this strategy has been used, some authors failed to find differences (Crundall et al., 1999; Yeng & Wong, 2015). Other studies have divided their drivers into high risk (young novice drivers) and low risk (mid-age experienced drivers) reporting differences between the groups (Wetton et al., 2010; 2011), while, yet again, others have failed to find a difference between such groups. For instance, a recent study failed to find a response time difference between high-risk and low-risk fire appliance drivers in a study of hazard perception skills in the emergency services (Crundall & Kroll, 2018).

Perhaps, the present results argue for an alternative way of measuring/analysing the results of the HP test. The fact that novices spotted the overtaking hazards as fast as the experienced drivers does not prove however that they have developed their situation awareness and are able to identify such hazards prior to its materialisation (prior to the onset). A recent study by Ābelea, Haustein, Møllera and Martinussen (2018) found that young drivers are poor at detecting hidden hazards and they have suggested specific hazard perception training to overcome this. Their results were previously supported by Crundall et al. (2012) who found that fixating precursors to subsequent hazards was particularly difficult for novices. Specifically, driving instructors and experienced drivers were faster at fixating both behavioural precursors and abrupt hazards when compared to learners. Possibly, this was the case during the present study. Novices noticed the hazards once they had materialised (e.g. an overtaking car), but may have not noticed the precursor to this hazard. Therefore, it might be worth looking at a more sensitive measure, as suggested by Crundall and Kroll (2018), focusing on precursors to hazards. Crundall and Kroll (2018) found that their low-risk, experienced fire-appliance drivers did indeed fixate the precursors more than the high-risk drivers and the control group. Therefore, further research would be necessary in order to examine hazard perception performance considering fixations on precursors while making use of mirror information. This measure might work as a better discriminator between experienced and novice drivers.

In summary, no significant results were found neither between the experienced groups or between the traditional design of the HP test and the mirror condition. Several possibilities have been discussed to address this lack of significance. Increasing the sample size might contribute for differences between the experienced groups or perhaps focusing on more sensitive measure such as examining responses to precursors might yield more reliable results in terms of differentiating between safe and unsafe drivers. Finally, there is also the possibility that learner and novice drivers were better trained on the hazard perception test, due to learners actively practicing the HP test before their driving test and novices still keeping fresh memory of their past practice (which was more recent in comparison to the experienced drivers). If this is the case however, this training only improved their performance on the test and this has not eradicated the danger that novices face in the first six months of driving as indicated by the collision statistics. Future research could look at this possibility.

Chapter 4

DESIGN OF THE HAZARD PERCEPTION CLIPS

Experiments 2 and 3

Does the design of the car overlay impact hazard perception performance? Comparing transparent vs. opaque overlay

4.1 Summary of previous findings

In Chapter 3, the aim was to study three different types of video clip design (forward view, car interior and mirror condition) and find out whether providing a more realistic design, including mirror information, will enhance discrimination between novice and experienced drivers in terms of reaction times and eye movements. It was predicted that experienced drivers would respond faster than novices across all conditions, though especially during the mirror condition where the gap between the experienced groups should be more apparent. There was, however, insufficient evidence to reject the null hypotheses, with no differences for driving experience regarding the performance on the hazard perception test. Experienced drivers did not react or fixate the hazards faster when compared to novices.

This result was surprising given the wealth of evidence that suggests hazard perception skill to be related to driver safety (Boufous et al., 2011; Drummond, 2000; Horswill & McKenna, 2004; Wetton, Hill & Horswill, 2013). Assuming that previous studies have demonstrated a genuine link between hazard perception response times and crash liability, why might the current study have failed to do so?

One possible explanation for these results could be the sample size. A larger sample might have yielded the basic experience effect between the driving groups, though previous studies in the field have demonstrated experiential differences with similar sample sizes (e. g. Borowsky, Shinar & Oron-Gilad, 2010; Smith, Horswill, Chambers & Wetton, 2009; Wetton et al., 2010).

Alternatively, we may question the assumption that our novice drivers had a higher crash liability than the experienced drivers (e.g. Underwood, 2007). However, current crash statistics still show a preponderance of fatalities and collisions in the youngest age bracket (DfT, 2015, 2016; RAC, 2017), and there is no reason to suspect the current sample of novices differs from the crash profile of novice drivers in general.

A third possibility is that learners who took part in the experiment were actively practicing the HP test (by the time the experiment took place), which improved their test performance. The UK hazard perception test was designed to encourage training and raise skill levels in our youngest drivers. However, this possibility needs further investigation as while evidence suggests that certain types of novice driver collision (non-low speed crashes on public roads where the driver has some blame) have reduced as a direct result of the introduction of the official HP test (Wells et al., 2007), there are several studies that have identified skill differences between novice and experienced drivers in the UK since this period (Helman, Grayson & Parkes 2010; McKenna et al., 2006).

Finally, there is the possibility that the current clips lack certain qualities that clips in other successful tests may have. As there is no agreed standard method of developing HP stimuli, and little definition regarding what makes a successful HP clip, it is quite possible that the current stimuli lack undefined elements that are crucial to test success.

The failure to find basic discrimination between novice and experienced drivers hindered the ability of the experiment to identify any potential benefits from adding the car interior and the mirror information to the clips, despite previous studies identifying benefits from adding additional visual and mirror information (Alberti et al., 2014; Shahar et al., 2010). Since no differences were found between the forward view clips and those that contained mirror information it cannot be concluded that this newly developed design is encouraging better performance on the HP test. However, one of the aims of this thesis is to create a more realistic and immersive clip design therefore, in the next two experiments the mirror design will be tested again. This type of design permits the inclusion of a wider range of hazardous situations, including both overtaking and undertaking hazards present in everyday driving, which is potentially important when considering hazards in other countries. Experiment 1 demonstrated that both driver groups made use of the information in the mirrors, suggesting that the information was not superfluous and did influence visual search.

The current chapter will focus on two things. First, the basic hazard perception effect – i.e. a performance difference between driver groups that are categorised on the basis of experience – will be sought once more. To increase the possibility of finding this effect, the sample size will be increased, and the study will focus solely on using clips with mirror information. As the mirrors proved to attract attention in the previous study, they are retained in the hope that they will contribute to an experiential difference.

Eye data and reaction times will be measured again and another design point will be assessed. One of the questions regarding the design of the HP clips used in Experiment 1 that remains unanswered is whether the car overlay should be opaque or transparent. The initial design choice was to make the A-pillars and roof or the car interior transparent in the HP clips. This was chosen as real A-pillars do not completely obscure part of the world: stereopsis and small head movements allow areas of the world otherwise masked by A-pillars to still be visually inspected. Likewise, when stopped at traffic lights that would be otherwise obscured by the roof of the car, the driver may lean forward to bring the traffic lights into view. A completely opaque overlay does not allow behavioural modifications to overcome the obscuration of relevant stimuli, and, while a transparent car interior may not look realistic, it may evoke more realistic scanning behaviour. This assumption will be tested in the current experiment, comparing the hazard performance and scanning behaviour of drivers viewing clips with both opaque and transparent overlays.

4.2 Introduction

4.2.1 The importance of "see-through" design

It has been argued from transparent structures within a car would be very helpful in permitting drivers to spot otherwise obscured hazards (Lindemann & Rigoll, 2017).

There is some existing research on see-through functionality, most of it focused on the so called “blind spots” where visibility is occluded, showing that this occlusion could lead to collision (Cheng et al., 2016; Yashuda & Ohama, 2012). It has also been argued that limited field of view could lead to a high chance of fatality (especially with heavy goods vehicles, see, Niewoehner & Berg, 2005; Summerskill et al., 2015).

During real driving the obscuration by the car A-pillar is offset somewhat by stereopsis and small head movements. However, during an experimental condition, if a completely opaque car overlay is used for the design of the video clips the simulation of stereopsis would not be possible and participants will not be able to see through the opaque frame. Thus, information which would normally be present in the real world could be absolutely occluded and will not allow drivers to perceive the entire objects which otherwise would be visible.

Therefore, the present study focuses particularly on the graphic overlay of the videos. The official hazard perception test contains only a forward view information and it has mainly concentrated on the types of hazards without prioritising other design features such as the mirror information or graphic overlay. These features might be secondary to the main purpose of the test, however its contribution is equally important in order to create a more ecologically valid HP test. The few studies that tested driving performance including mirror information did not focus on these features either (Shahar et al., 2010; Allen et al., 2005). However, in order to create valid and realistic hazard perception clips, every feature should be tested and measured empirically. As a result, two types of car overlay were designed. The first type contained transparent A-pillars and roof which allows to see through them and the second type consisted of fully opaque A-pillars and roof. Although, it can be argued that the transparent car overlay does not reflect a realistic version of a vehicle, it should be noted that this design simulates the fact mentioned above that real-world obscuration by A-pillars is not absolute. For that reason, a transparent overlay is considered to be more behaviourally realistic, if not structurally realistic. In addition, a decision was made to keep the mirror design (they did not enhance performance but neither hindered it). Although there was not a significant difference between conditions, results clearly showed that participants made use of mirrors and showing clips that contain only a graphic overlay hindered novices’ performance (they were considerably slower reacting to the hazards).

In favour of the transparent car overlay is the fact that some videos might contain situations where pedestrians are turning their head to look at the car and decide whether to cross or not the street (see Figure 4.1). This type of information might be occluded by the frames of an opaque overlay which could confuse participants (during an experimental condition it will not be possible to bring objects back into the visual field once they have been occluded by the opaque car overlay).

To the author's knowledge, there is no available research that directly compares the impact the type of a car overlay has on hazard perception performance. There are studies that compare different types of 3D transparent and semi-transparent car interiors (Lindemann & Rigoll, 2017), and the impact of see-through vehicles on driving performance (Yanagi et al., 2015; Rameau et al., 2016). However, none of these studies have looked at the impact an opaque overlay that occludes some of the visual information has on hazard perception skill. For example, Lindemann and Rigoll (2017) conducted a simulation using a see-through interface in a mixed reality driving simulator focusing on user experience (as most of the studies are mainly focused on the technical requirements than users' opinion). They created a transparent spot on the car dashboard which enabled drivers to see the road through it (it was either semi-transparent or fully-transparent). The transparent spot could either follow the movements of the driver's head or appear at arbitrary places within the car. Participants were asked to rate whether they preferred a semi-transparent or fully transparent visibility, relating the ratings to their feelings about irritation, comfort and helpfulness. Lindemann and Rigoll did not find any differences between the conditions. However, when participants were simply asked about their opinion, the semi-transparent version was criticised for being less-helpful in performing the driving task and the full transparent was deemed as unreal and uneasy. However, there was no opaque condition to provide a control condition, and even though participants did not like the design of the overlay, results regarding their driving performance were not significant. This suggests that there is no clear evidence of which type of design is more beneficial in terms of driving performance.

4.2.2 Overview of the study

First, this study (comprising experiments 2 and 3) was conducted as a second opportunity to obtain an experiential difference between novice and experienced drivers. Eye movements and response times were recorded (experiment 3), along with self-reported ratings where participants were asked to rate the hazardousness and realism of each clip (experiment 2). Following the null results of experiment 1, especial focus was placed on drivers' attention to precursors, as a potentially more sensitive measure of hazard skill.

Second, this study was designed to assess drivers' responses and thoughts regarding the opacity of the car overlay. It was predicted that the fully opaque overlay would artificially hinder hazard perception, potentially nullifying any experiential difference (should one be found in the first place).

4.3 Opaque vs. Transparent car layout: hazardous and immersion ratings (Experiment 2)

Following Lindemann and Rigoll (2017), Experiment 2 asked participants to rate how realistic this new design appeared to them. Prior to collecting behavioural measures (in Experiment 3), it was considered that it would be useful to first measure participants' opinions on which type of car interior seemed more realistic and closer to real driving. Thus, a comparison can be made between subjective measures and behavioural data.

4.3.1 Method

4.3.1.1 Participants

Twenty-three participants took part in this experiment (seven learners, seven experienced and nine non-drivers). All experienced drivers had more than five years of driving experience. The sample was composed of University students and visitors to the University during open days.

4.3.1.2 Design and materials

A 2 x 3 design was used where the independent variables were the type of car overlay and experience and the dependent variable was ratings given to the questions.
Ten hazardous stimuli were selected from the 20 video clips used in Experiment 1, specifically videos 1, 2, 3, 5, 6, 7, 10, 12, 13, 15 (see Table 2.1, Chapter 2, for reference). Five of the videos contained transparent car overlay and the other five opaque car overlay (see Figure 4.1).

4.3.1.3 Procedure

Participants received both verbal and written instructions. They were asked to watch each video carefully, and at the end of each clip provide a rating on a Likert scale from 1 to 7 regarding the hazardousness of the scenario (1 was "not at all hazardous" and 7 "extremely hazardous), how close to real driving each video seemed to them (1 was "not at all realistic" and 7 "extremely realistic"), and how much they felt immersed in the clip (1 was "not at immersed" and 7 "extremely immersed"). They provided their ratings on a rating sheet that was placed next to the computer. The stimuli were counterbalanced in the following way: group 1 saw clips 1 to 5 with a transparent car overlay and clips 6-10 with an opaque overlay. Group 2 was presented with clips 1 to 5 with an opaque overlay and 6-10 with the transparent overlay. The clips were presented randomly. They were also assigned randomly to one of the two conditions (opaque or transparent), however none of the participants were aware of the differences between the two sets of clips. All ten videos contained mirror information. Clips were displayed on a Lenovo ThinkPad laptop with dimensions 34.5cm x 19.5cm and resolution 1920x1080. After the session participants were debriefed and asked which type of overlay they preferred.

4.3.2 Results and Discussion

Ratings were subjected to 2 x 3 ANOVA where the type of questions and experience were the independent variables and the rating score the dependent variable (see Norman, 2010, for a justification for using ANOVA to analyse these data).

4.3.2.1 Hazard ratings

No differences were found between the ratings for either of the independent variables. There were no differences between the experienced groups (M=3.93 for experienced drivers, M=3.97 for learners and M= 3.66 for non-drivers), ($F(2, 20)=.33$, $MSe=.48$ $p=.72$, $\eta_p^2=.03$) or conditions (M=3.8 for transparent, M=3.9 for opaque), ($F(1, 20)=.02$, $MSe=.99$ $p=.89$, $\eta_p^2=.00$). Even though all ratings were above the half point on the scale, when a one-sample t-test was conducted to compare the mean ratings of each group of drivers to the mid-point of the scale (3.5), no significant differences were found (no drivers transparent (t(6)=1.89, p=.10), opaque (t(6)=.48, p=.64), learners transparent (t(8)=-.33, p=.74), opaque (t(8)=.91, p=.38), experienced transparent (t(6)=1.54, p=.17), opaque (t(6)=1.26, p=.25)). This suggests that there were no strong overall preferences.

4.3.2.2 How realistic?

No differences were found between the ratings for how realistic was the driving environment, for any of the variables. There were no differences between the experienced groups (M=5.94 experienced drivers, M=5.71 learners and M= 5.16 for non-drivers), ($F(2, 20)=.97$, $MSe=2.76$ $p=.39$, $\eta_p^2=.08$) or conditions (M=5.43 for transparent, M=5.7 for opaque), ($F(1, 20)=2.59$, $MSe=.99$ $p=.12$, $\eta_p^2=.16$). Although there were no differences, means were higher than the midpoint on the scale, which suggests that clips were rated as realistic. To confirm this, a t-test was conducted to compare the means of each group of drivers to the midpoint of the scale. Significant differences were found for each group (no drivers transparent (t(6)=5.78, p<.001), opaque (t(6)=3.62, p<.05), learners transparent (t(8)=3.65, p<.01), opaque (t(8)=3.37, p<.05), experienced transparent (t(6)=5.10, p=.005), opaque (t(6)=14.8, p=.001)). Both layouts were rated as realistic.

4.3.2.3 How close to real driving?

No differences were found for immersion for both experience (M=5.39 for experienced drivers, M=5.03 for learners and M= 4.66 for non-drivers) ($F(2, 20)=.55$, $MSe=2.11$ $p=.59$, $\eta_p^2=.05$) and conditions (M=5 for transparent, M=5 for opaque), ($F(1, 20)=.00$, $MSe=.00$ $p=.94$, $\eta_p^2=.00$). However, according to the mean values, the stimuli were rated as relatively immersive. The t-tests showed that there were differences between the mean ratings and midpoint scale only for the transparent condition for non-drivers and learners (non-drivers, $t(6)=2.86$, $p<.05$; learners, $t(6)=2.82$, $p<.05$) but not for the opaque condition (non-drivers, $t(6)=1.67$, $p=.14$; learners, $t(6)=1.91$, $p=.09$)). In turn, the experienced drivers rated both conditions higher than the midpoint (transparent ($t(6)=7.31$, $p<.001$), opaque ($t(6)=8.93$, $p<.001$)).

4.3.2.4 Preferences

After each session every participant was asked whether they have noticed a change during the duration of the video clips. Results showed that 5 participants did notice the difference from clip to clip (two experienced, one learner and two non-drivers) and five did have the feeling that something changed but did not know what (two experienced, two learners and one non-driver). In turn, 13 (56.5%) of the participants reported that they did not notice anything. Following this, the experimenter showed the difference to the participants and asked them which car overlay they preferred. In total 60.9% of the sample preferred the opaque overlay. Particularly, 71% of the experienced and 71% of the learners opted for the opaque overlay, while the percentage of the non-drivers that preferred the opaque stimuli was a bit lower (54%). Finally, it was explained to participants why the transparent car overlay would be more beneficial in term of driving performance and they were asked again which one they preferred. Only three participants changed their preferences from opaque to transparent.

These findings based on the users' ratings are consistent to those of Lindemann and Rigoll (2017) who failed to find significant differences between the two types of design. However, when participants were simply asked which design type they preferred, they opted for the opaque one as it "looked like a real car". Even though, the experimenter explained the disadvantages of the opaque overlay over the transparent one, only three participants changed their preferences.

While participant preferences are useful to guide the design of stimuli, they should ideally be supported by other data. Despite their preference for the opaque stimuli, there were no differences between the hazard ratings of the clips, and the opaque car overlay was not rated as more immersive or realistic (although both overlays were rated higher than the mid-point of the scale in terms of realism). However, an interesting fact is that when participants were asked how close they felt to really driving, only the transparent condition was rated significantly higher than the average by both non-drivers and learners (experienced rated both condition as significantly higher than the average). This suggests that all drivers and especially learners and non-drivers consider the visibility provided by the transparent condition beneficial and provides an experience that is closer to real driving.

Due to the inconsistencies of the results, it was decided to look at more objective measures such as eye movement data and response times to hazards, which are expected to provide more clear arguments in favour of one of the two designs.

4.4 Opaque vs. Transparent car layout: Behavioural measures (Experiment 3)

4.4.1 Method

4.4.1.1 Participants

Forty-two participants took part in this experiment (21 experienced and 21 novice drivers, N=37 females). The experienced drivers reported a mean age of 21.3 years, average mileage of 5928.6 miles and average driving experience of three years and three months. The novice drivers (15 learners) reported a mean age of 18.5 years, average mileage of 480 miles and two months of average driving experience. All participants had normal or corrected to normal vision. The sample was composed of university students and staff, and visitors to the university.

4.4.1.2 Design

A 2 x 2 between groups design was used for this study, where the factors were the level of experience (experienced vs. novice drivers) and type of car overlay design (transparent vs. opaque). Driver's fixations to the hazards and reaction times were measured as dependent variables.

The 20 hazardous stimuli applied for this experiment were identical as in Experiment 1 (see Chapter 2, Table 2.1), however this time all videos contained mirror information with ten of the videos containing transparent car overlays and ten containing opaque car overlay (see Figure 4.1).

Eye movement information was recorded and analysed identically as in experiment 1. This time AOIs were defined also around the dashboard and car frames delimitating the exact shape of these (please, refer back to Figure 2.5).

4.4.1.3. Materials and apparatus

Materials and apparatus were identical to those used in the mirror condition in Experiment 1 (see Chapter 3), with the exception that the overlay of the car interior could either be completely opaque, or have transparent A-pillars and roof, as shown in Figure 4.1.

4.4.1.4 Procedure

Participants received both verbal and written instructions. The procedure was very similar to that of Experiment 1 regarding the calibration, eye tracking measures and response times.

Participants were asked to watch 20 clips containing a pre-defined hazard, and press a button every time they thought they saw a hazard. Clips were displayed on the same eye tracking equipment as in experiment 1. All participants saw ten clips with the transparent car overlay and ten with an opaque one (see Figure 4.1). The experiment was counterbalanced in the following way: one quarter of the participants saw videos 1-10 with the transparent car overlay and videos 11-20 with the opaque overlay. For the next quarter of the sample, videos 11-20 were transparent and 1-10 opaque, and so on (see Table 4.1). Each participant was randomly assigned to one of the four groups. This counterbalancing was applied to avoid order effects. Participants were naïve to the difference between the types of overlays at the start of the experiment, and all stimuli were randomised within a single block. Participants were also asked to rate how hazardous, how realistic and how close to real driving it made them feel after each clip using a Likert 1 to 7 scale where 1 was “not at all” and 7 “extremely”. After the session, participants were asked whether they had noticed any differences between the clips. They were then debriefed regarding the differences between the two types of car overlay and were shown the different types of clips again in order to elicit qualitative feedback.

The experiment took approximately 40 minutes. After the experiment participants were provided with feedback on their performance (they were given the opportunity to see their own eye movements).

Table 4.1: Counterbalancing order across groups

Groups	Video design	
Group 1	Videos 1-10 transparent	11-20 opaque
Group 2	Videos 11-20 transparent	1-10 opaque
Group 3	Videos 1-10 opaque	11-20 transparent
Group 4	Videos 11-20 opaque	1-10 transparent

Figure 4.1: The top image represents a clip that contains a transparent car overlay and the bottom image the same clip with opaque car overlay

4.4.2 Results

4.4.2.1 Eye movement measures

The method of analyses was identical to the one used in Chapter 3 (please, refer back to Chapter 3, section 3.3) and data were subjected to 2 x 2 ANOVA where the independent variables were type of car overlay and experience and the dependent variable participants' fixations.

4.4.2.1.1 Fixations on hazards

First, time-to-first-fixate the hazards was calculated (as in Chapter 3). The data were then subjected to a 2 x 2 ANOVA with experience and overlay-type as the independent variables. No differences were found for time-to-first-fixate the hazard between the experienced groups (M=13.5% for experienced drivers and M=15.2% for novices) (F(1, 40)=1.63, MSe=35.2 p=.21, η_p^2=.04) or between the conditions (M=14.7% for transparent frame and M=13.9% for opaque frame), (F(1, 40)=.33, MSe=41.3 p=.57, η_p^2=.00). The interaction between condition and experience did not reach significance (F(1, 40)=1.63, MSe=41.3 p=.67, η_p^2=.00). Contrary to what was expected, the transparent frame did not enhance faster hazard detection from any of the groups. The fact that it provided more visibility did not seem to decrease the time taken to first fixate the hazard in comparison to the opaque frame.

The total time spent looking at the hazards (the dwell of the eye within the relevant Area of Interest (AOI)) was also calculated, by adding the duration of each fixation within the AOI and then multiplying it by 100 to convert it into a percentage. No differences were found neither for experience (M=39.1 % for experienced drivers and M=40.1% for novices), F(1, 40)=.08, MSe=293.8 p=.78, η_p^2=.00), nor between the two conditions (M transparent= 38.6, M opaque=40.6), (F(1, 40)=.59, MSe=132.2 p=.45, η_p^2=.01). Novices did not fixate longer the hazards than the experienced drivers as it was expected. The interaction between experience and condition was not significant (F(1, 40)=.00, MSe=132.2 p=.94, η_p^2=.00).

4.4.2.1.2 Fixations on mirrors and frames

All fixations on mirrors were summed and divided by the length of the corresponding video clip and then subtracted by the length of the hazard window (for reference, please refer back to Chapter 3, point 3.3.1.2). To create a percentage dwell on mirrors, fixations were multiplied by 100. A main effect was found for mirrors (F(2, 80)=84.7, MSe=1.16 p<.001, η_p^2=.68). Pairwise comparisons revealed that differences lay specifically between the rear-view mirror and the other two mirrors (p<.001). Participants looked significantly longer in the rear-view mirror in comparison to the right mirror ($p<.001$) and the left mirror ($p<.001$). No main effect was found for experience ($F(1, 40)=1.65$, $MSe=13.6$ $p=.21$, $\eta_p^2=.04$), however there was a significant interaction for mirrors vs. experience ($F(2, 80)=9.95$, $MSe=1.16$ $p<.001$, $\eta_p^2=.19$). Experienced drivers spent significantly more time looking at the right mirror than the novices, however the novices spent significantly more time looking at the rear-view mirror (see Figure 4.2).

AOIs were defined also around both the transparent and opaque overlay, and the dashboard bits in order to find out whether participants were using the visual information provided by the transparent frame (see Figure 4.1). All fixations on the frame and dashboard were summed and divided by the length of the corresponding video clip and then subtracted by the length of the hazard window. To create a percentage dwell on mirrors, fixations were multiplied by 100. Results showed that there was not a significant difference between the total fixations on the transparent or opaque frame (M transparent =56% vs. M opaque =56% ($F(1, 40)=.01$, $MSe=.12$, $p=.91$, $\eta_p^2=.00$), nor there was a significant interaction between type of frames and experience ($F(1, 40)=.36$, $MSe=.12$ $p=.55$, $\eta_p^2=.00$). There were no significant differences between the dwells on the dashboard during the two conditions (M transparent=0.98% vs M opaque=1.11%) ($F(1, 40)=1.74$, $MSe=.20$ $p=.19$, $\eta_p^2=.04$), however there was a significant interaction between condition and experience. Experienced drivers dwelled significantly more on the dashboard during the opaque frame condition ($F(1, 40)=4.97$, $MSe=.21$ $p<.05$, $\eta_p^2=.11$), (see Figure 4.3) .

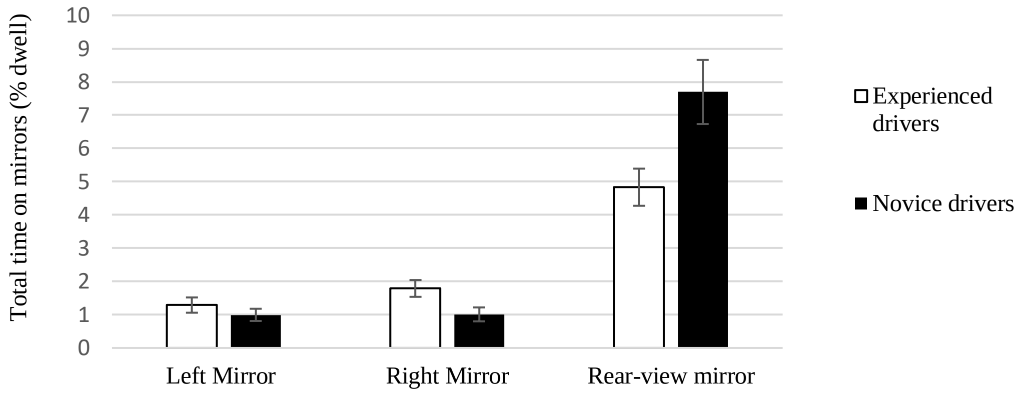


Figure 4.2: Total time spend looking at the left, right and back mirror across experience (with standard error bars added)

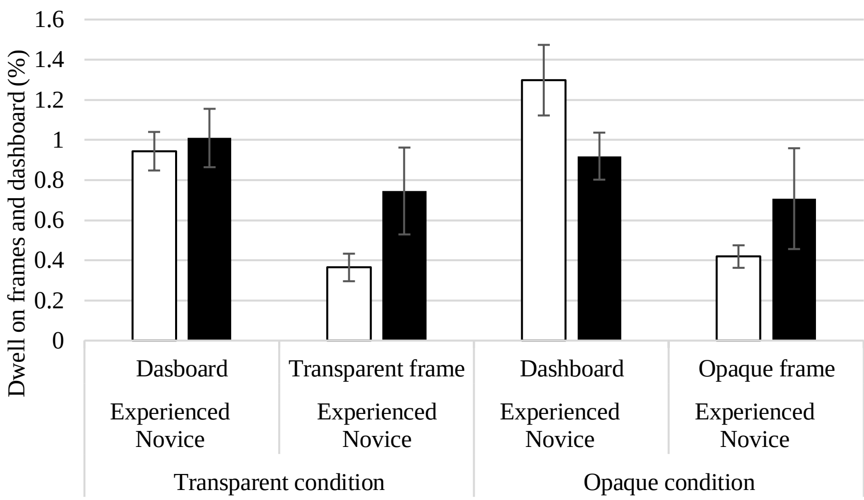


Figure 4.3: Percentages of dwells on frames and dashboard across both conditions and experience (with standard error bars added)

4.4.2.1.3 Spread of search along the horizontal and vertical meridians

Participants' fixations (measured in pixels) along the horizontal and vertical meridians were also analysed using standard deviation. Significant differences were found for the horizontal spread of search between experienced and novice drivers (M experienced =315.4 px, M novices =291.3 px), (F(1, 40)=4.90, MSe=2490.6.3 p<.05, $\eta_p{}^2$=.11). Experienced drivers explored the driving environment significantly more and showed a wider spread of search in comparison to the novices. There was no main effect for condition or a significant interaction for the horizontal scanning.

No main effect was found for the vertical spread of search for both experience and condition, however there was a significant interaction between condition vs. experience (F(1, 40)=7.69, MSe=118.8 p<.005, $\eta_p{}^2$=.16). The experienced drivers showed a wider vertical search when watching the videos with the opaque car overlay, while novices had considerably wider vertical search during the transparent overlay videos (see Figure 4.4).

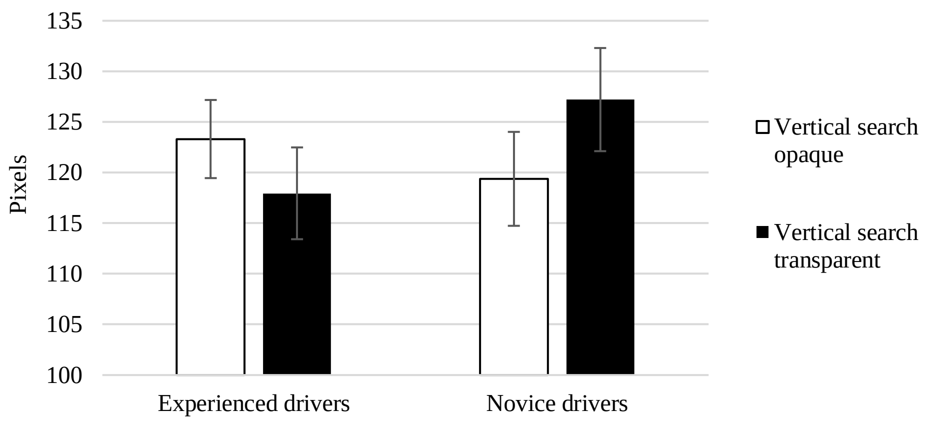


Figure 4.4: Vertical scanning of the driving environment across experience (experienced vs. novices) and conditions (transparent vs. opaque) (with standard error bars added)

4.4.2.1.4 Precursor analysis

Precursors to hazards were also analysed. A precursor is the clue from the visual scene that would help the driver predict how the hazardous situation is going to develop. Experienced drivers are more likely to notice precursors earlier than novices, as they have been driving for a longer time, and have been exposed to a greater number of similar hazardous events. A precursor AOI window (exactly 2000 ms for each video) was defined just prior to the hazard onset. Immediately after the precursor AOI, the hazard AOI was defined. Precursor net dwell times were calculated. The units were converted from ms to percentages. Net dwell time is defined as the sum of sample fixation durations for all gaze data samples that hit the area of interest (AOI). There was a marginal experience effect (M=32.2 for experienced, M=28.1 for novices), ($F(1, 40)=3.43$, $MSe=102.2$, $p=.07$, $\eta_p^2=.08$). Even though this effect was not statistically significant, there was marginal evidence to suggest experienced drivers devoted a greater amount of their attention to precursors.

4.4.2.2. Behavioural responses to hazards

The method of analyses was identical to the one used in Chapter 3 (please, refer back to Chapter 3, point 3.3) and data were subjected to 2 x 2 ANOVA where the independent variables were type of car overlay and experience and the dependent variable participant’s accuracy and RTs.

4.4.2.2.1 Response time to predefined hazards

Response times to predefined hazards were analysed (please, refer back to Chapter 3, point 3.3.2.1 for a detailed description of the method of calculation of the RTs) and no differences were found for the experienced groups. The basic experience effect of the HP test was not replicated in this experiment, which is consistent with Experiment 1. Experienced drivers did not react faster to the pre-defined hazards than the novices (M=2030 ms for experienced, M=2077 ms for novices), ($F(1, 40)=.62$, $MSe=.30$ $p=.44$, $\eta_p^2=.02$). No significant differences were found between the two conditions either. The transparent frame did not produce faster RTs for either of the groups (M= 2038.7ms for transparent, M=2068ms for opaque) (F(1, 40)=.32, MSe=.06 p=.57, η_p^2=.00). This result is contrary to the hypothesis, as it was expected that the greater visibility offered in the transparent condition would allow faster detection of the hazards.

4.4.2.2.2 Accuracy

Accuracy (i.e. the percentage of hazards that participants responded to within the hazard window) was also calculated with no differences for the experienced groups (M=33.8% for experienced, M=36.4% for novices), (F(1, 40)=1.11, MSe=130.1 p=.29, η_p^2=.03) or between conditions (M=34.5% for transparent, M=35.7% for opaque), (F(1, 40)=.24, MSe=122.8 p=.63, η_p^2=.01). Experienced drivers did not identify more hazards than the novices. The interaction was not significant either (F(1, 40)=.47, MSe=122.7 p=.49, η_p^2=.01).

4.4.2.2.3 Extra responses to hazards

Ratios for extra responses to hazards were analysed (please refer back to Chapter 3, point 3.3.2.3) but no differences were observed between experienced and novice drivers (M=2.03 for experienced, M=1.59 for novices), (F(1, 40)=.2.82, MSe=4.10 p=.10, η_p^2=.07) or across conditions (M=1.78 for transparent, M=1.84 for opaque), (F(1, 40)=.70, MSe=.07 p=.41, η_p^2=.01). The opacity of the overlay did not influence participants' additional clicks during the video.

4.4.2.3 Ratings

Participants were again asked to rate the hazardousness, realism and immersion (as in Experiment 2) of each video using a scale from 1 to 7 (1 "not at all" and 7 "extremely so") (see Table 4.2). However no significant differences were found for any of these variables (which is consistent with Experiment 2). There were no differences in the ratings regarding immersion and realism between the transparent and opaque car overlay. None of them was rated more realistic or hazardous than the other. There were no differences for the ratings between the experienced groups either. None of the interactions were significant.

Table 4.2: Means of the ratings (from 1 to 7) for each question: How hazardous do you think that particular clips was? (Q1), How close to real driving did that clip look? (Q2), How much did you feel 'you were there' in that clip?(Q3) across the experience groups

	Transparent			Opaque		
Questions	Q1	Q2	Q3	Q1	Q2	Q3
Experienced	4.05	5.39	5.262	3.96	5.40	5.19
Novice	3.91	4.97	4.84	4.08	5.00	4.97

Similarly to Experiment 2, participants were asked whether they noticed anything that changed during the duration of the clips and about their preferences regarding the transparent and opaque car overlay. Only three participants (out of 42) reported that they have noticed that the car overlay changed and six reported that they had noticed that something changed but did not know what it was (some of them mentioned they could see better, or became more aware of the mirrors or even mentioned “blind spots”). The rest of the sample (78.6%) did not notice the change.

Regarding their preferences nine (21.4%) participants preferred the opaque overlay and 33 (78.6%) the transparent. These results are contrary to those in Experiment 2, as previously participants opted for the opaque overlay. However, in Experiment 2, participants were asked to merely watch the videos without actively searching and detecting hazards, while in this experiment they had to pay attention and spot possible hazards.

4.5 General discussion

In this study two types of graphic car overlay were compared: transparent vs. opaque incorporated into short hazardous video clips. These clips were presented to novice and experienced drivers with reaction times and fixations to hazards recorded. Participants were also asked to rate the videos in terms of hazardousness, realism and immersion, as per Experiment 2.

This second attempt to find a basic experiential effect also failed, despite doubling the relevant sample sizes. There were no differences in terms of reactions to hazards and no differences were found for fixations either. This time the sample size was augmented, however experienced drivers did not respond faster to hazards, or fixate them sooner. In Experiment 1 (Chapter 3) drivers with an average of eight years of driving experience were tested and in Experiment 3 experienced drivers had an average of three years of driving experience and none of these groups independently of the years of driving showed better performance on the HP test. The attempt to replicate the basic experience effect failed in this experiment too, however this time significant differences were found between experienced and novice drivers for the horizontal spread of search which is consistent with the previous literature (Crundall & Underwood, 1998; Crundall, Underwood & Chapman, 1999), meaning that this effect might be susceptible to the sample size. Literature has been consistent so far on replicating that experienced drivers explore significantly more of the driving environment in order to be able to deal with more complex visual scenes when compared to novices (Crundall & Underwood, 1998; Mourant & Rockwell 1970, 1972; Underwood, Chapman, Bowden, & Crundall, 2002). The fact that we failed again to find differences for reaction times, but we did find that experienced drivers showed wider spread of search (after augmenting the sample) might suggest that the latter effect is more robust.

As noted previously, one explanation for the lack of differences between the driver groups could be the possibility that learner drivers have become trained on hazard perception to such an extent that we should no longer expect to see differences between novice and experienced drivers in the UK. Despite the evidence that the official test has reduced UK collisions (Wells et al., 2007) there is evidence that even highly experienced drivers can be trained to improve their hazard perception skills (Horswill et al., 2013; Castro et al., 2016). Coupled with the fact that the heightened crash liability of novice drivers still persists, this suggests that there is still work to be done. Indeed, it is possible that UK novices are extremely well trained in undertaking the hazard perception test, but less well trained in spotting actual hazards on the road. Perhaps the availability of online tools that permit practising the hazard perception test has an important role in training learner drivers to spot and react to hazards, however only within artificial scenarios.

Although no differences were found for the overlay design and experience, there was a significant interaction between experience and mirrors. Experienced drivers spent significantly more time looking at the right mirror than the novices, while novices spent significantly more time looking at the rear-view mirror. In order to see whether there were differences in terms of identifying early precursors, fixations that fell within a two second window prior to the hazard onset were calculated. Thus, it was possible to identify whether experienced drivers spent more time looking at the right place prior to the hazard. A marginal difference was found between the groups with experienced drivers spending more time on fixating the precursor prior to the hazard onset than novices. Although not significant, these last results point towards some evidence that precursors might be a better discriminator between experienced and novice drivers (Crundall et al., 2012, Crundall & Kroll, 2018). Coupled with the analysis of spread of search, which replicated the typical finding that more experienced drivers search more widely in the horizontal plane (Chapman & Underwood, 1998), the data suggest that the clips contain qualities that are sensitive to driving experience, yet these differences do not feed into the traditional measure of hazard response time.

Participants did not spend significantly more time looking at the transparent areas of the overlay, compared to those same areas in the opaque condition. It cannot be concluded that they were attending more to the visual scene through the transparent A-pillars and through the roof of the car interior. Furthermore, neither of the two car overlays appeared more hazardous, realistic or immersive to participants when they were asked to rate them. These results were identical for both Experiment 2 and 3; there were no differences between the ratings when participants were only shown the videos without asking them to spot hazards and when they were actually asked to actively search for hazards. Although participants did not differ in their ratings, when these same ratings were compared to the mid-point of the scale the transparent condition was rated significantly more "close to real driving" than the opaque by the learners and no drivers. This yields some evidence in favour of the transparent design as learners and non-drivers considered the transparent design more immersive.

These results do not provide definite evidence that suggests that adding mirror information or transparent frame to the HP video clips will enhance better performance or discriminate clearly between experienced and novice drivers. Neither, however, do the data provide evidence against the addition of these elements. It should be noted that there are pragmatic reasons why this type of design (the inclusion of mirrors, with a transparent interior) should be maintained for future research. Providing mirrors in the HP video clips reflects the information that would be available to drivers during real life driving while the transparent placeholder allows drivers to perceive essential cues which otherwise would be unfairly hidden by the opaque overlay.

Several studies have already suggested improvements to be considered for the design of the traditional hazard perception test (Alberti et al., 2014; Crundall, 2016; Lim et al., 2013; Pradhan and Crundall, 2017; Shahar et al., 2010), however no changes related to the design of the video clips have been made by the DVSA (with exception of the improvement of the quality of the clips, by upgrading the image into CGI.).

The traditional HP test has been studied mainly within the Western driving environment. It would be interesting to study HP performance within different driving context with participants that have not been exposed to any HP training and find out whether there will be differences between experienced and novices drivers. There have already been attempts featuring Malaysian drivers (Lim et al., 2014) however no differences were found between the experienced groups in terms of reactions to hazard. For that reason, the next study will feature Chinese Spanish and UK sample in order to compare performance on the HP test and see whether this time there will be a basic experiential effect between novice and experienced drivers. If UK novices are now trained to the same level as experienced drivers, then perhaps the basic experiential effect will be easier to find in other countries that do not have a national hazard perception test.

Chapter 5

AN INTERNATIONAL PERSPECTIVE ON HAZARD PERCEPTION

Experiment 4

Testing the Hazard perception test in three different countries: China, Spain and the UK

The material covered in this chapter has been adapted from previously published paper in Ventsislavova, P., Crundall, D. Baguley, T., Castro, C., Gugliotta, A. Garcia-Frnandez, P., Zhang, W., Ba, Y. & Li, Q. (2019). A comparison of hazard perception and hazard prediction tests across China,Spain and the UK. *Accident Analysis and Prevention, 122*, 268-286.
https://doi.org/10.1016/j.aap.2018.10.010

5.1 Summary of previous findings

Many researchers agree that the hazard perception skill is, perhaps, the only higher order cognitive skill to relate to crash-risk and that hazard perception tests have huge potential for reducing collisions around the world. Surprisingly, the previous two studies described in Chapter 3 and 4 failed to find significant differences in performance on the test between the experienced and novice drivers, despite efforts to test different range of experience and increase the sample size in the last experiment. One of the possibilities discussed was that UK novice drivers must practice hazard perception in order to pass the official DVSA test, which might explain why it is difficult to find experiential differences: in essence, the novice cohort could now be as practiced at hazard perception as the more experienced drivers. Eye movement data also failed to yield significant differences, and even though the design of the video clips was improved (including mirror information and transparent frame to provide better visibility) aiming to enhance hazard detection, this did not show any differences in performance neither between the conditions nor between the groups. Glances at the mirrors were consistent between the studies in Chapters 3 and 4, showing that novices have not yet developed the habit to check the side mirrors while experienced drivers examined both left and right mirrors.

Furthermore, though the transparent overlay did not appear to directly affect hazard perception performance in comparison to a fully opaque overlay, participants (especially learners and no drivers) did rate the transparent overlay as more immersive (however, only when their ratings were compared to the mid-point of the scale). This does not reflect increased attention to the transparent overlay per se and it cannot be concluded that participants did take advantage of the transparency of the overlay. When directly asked which they preferred, participants chose the opaque overly in Experiment 2 and the transparent overlay in Experiment 3. Given that the latter experiment involved an active search for hazards, this is perhaps the more relevant choice.

Also, differences in vertical spread of search were observed in Experiment 1 where the car overlay was present in all conditions. There was also a significant interaction in Experiment 3 with the experienced drivers exploring more vertically with the opaque video clips while novices showed more vertical search during the transparent condition. This suggests that experienced drivers might have looked for additional information such as precursors to hazards or other clues occluded by the opaque frame. These findings demonstrate that the design of the clips does influence participants' eye movement behaviour, although it did not enhance faster hazard detection in terms of response times in both Experiments 1 and 3.

When fixations upon precursors were analysed, there was only a marginal significance between the experienced and novice drivers. This is surprising considering previous studies that consistently found differences between the groups. However, the data shows that experienced drivers fixated more precursors than the novices (even though only marginally) which suggests that this type of measure/analysis might be more reliable than the response latencies to hazards.

As the HP test has been developed over the past 50 years on pragmatic, rather than theoretical grounds, it is difficult to determine a reason behind the lack of validity. Assuming that UK novice drivers may now be well versed in hazard perception since the introduction of the formal test in 2002, the following study will examine differences in performance in international samples.

Many studies have been conducted in other countries, seeking experiential differences in hazard perception tests. Unfortunately, it is difficult to compare these studies across countries, as different research groups have designed their own hazard perception tests without following any official guidelines (as there is no official protocol that explains how to build a hazard perception test). The wide variety of tests prevents any sensible comparison between countries.

One study has however used the same test methodology across two countries, comparing response times to hazards of UK and Malaysian drivers on both UK and Malaysian clips (Lim, Sheppard & Crundall; 2013). They also failed to find the basic experiential effect in both groups, though other cross-cultural effects were apparent (e.g. the response frequency of UK drivers was extremely high in Malaysian clips, compared to Malaysian driver responses). The following study provides a second opportunity to investigate cross-cultural differences, measuring HP performance of drivers from China, Spain and the UK, using clips filmed in all three countries.

5.2 Introduction

5.2.1 An international perspective on hazard perception testing

When countries are compared on the number of road fatalities, accounting for population size, a handful of European countries typically dominate the safest spots at the top of the table (e.g. Sweden, Norway, Denmark, Switzerland, the UK, and the Netherlands; OECD/ITF, 2015; WHO, 2015). Thus it behoves researchers in these countries to identify which of their own safety initiatives contribute significantly to their national safety record and to assess whether these interventions are suitable for export to other countries who may benefit in their own attempts to reduce on-road injuries and fatalities. Certainly, the introduction of the national HP test in the UK has ostensibly decreased road collisions. As noted in Chapter 1, Wells et al., (2008) reported a 17.4% reduction in non-low speed collisions (where blame could be attached) linked to the introduction of the test in 2002. This result demonstrates the significant impact that the HP test has had on UK road safety, and raises the possibility that this could be of equal use to other countries who are facing even greater road safety challenges.

The first country to implement the HP test as a part of their licencing procedure was Australia (specifically Victoria in 1996) and UK, Queensland (Australia) and the Netherlands have since followed the example and developed their own HP tests. In addition, there are research groups around the world who have developed HP tests in their own countries, including Spain, Germany, The Netherlands, Israel, Singapore, Malaysia, Canada, Hong Kong, China, Japan, and New Zealand (e.g. Borowsky et al., 2010; Cheng, Ng & Lee, 2011; Cocron, Bachl, Früh, Koch & Krems, 2014; Gau, Yu & Hou, 2015; Malone & Brünken, 2016; Isler, Starkey & Williamson, 2008; Lim et al., 2013; Rosenbloom et al. 2011; Scialfa et al., 2014; Shimazaki, Ito, Fujii, & Ishida, 2017; Vlakveld, 2014; Wang, Peng, Liang, Zhang, & Wu, 2007; Yeung, & Wong, 2015). Unfortunately, the results of many studies from around the world are mixed, with some researchers discriminating between safe and less-safe drivers on the basis of hazard responses (e.g. Horswill, et al., 2015; Wallice & Horswill, 2007), while others fail to find this basic effect (Sagberg, & Bjørnskau, 2006; Lim et al., 2013; Yeung & Wong, 2015).

5.2.2. Are hazards culturally specific?

It is difficult to pinpoint the reason why some studies successfully discriminate between safe and less-safe drivers, while others do not, as the precise design of these various tests can differ on many crucial points. The most interesting difference between these studies is the country in which they are conducted. Both the stimuli (the video clips containing the hazards) and the participants, are culturally specific to the region. There are wide cultural differences in the nature of driving, including both the legal and social rules that govern acceptable behaviour, which in turn influence the nature of the hazards. It is possible that some types of hazard are more prevalent in particular countries, and that some of these hazards may be less successful in discriminating between driver groups (Crundall et al., 2012; Crundall 2016) or are simply unsuitable for a hazard perception test. For instance, when Crundall was filming clips in Malaysia for the Lim et al. (2013; 2014) studies, he reported that many hazards did not make the final cut. The majority of these rejected hazards were interactions between the film car and motorcycles, which would overtake without warning and cut in front of the film car, necessitating urgent braking in some instances (Crundall, 2015, personal communication). The immediate appearance in the camera view of these motorcycles, would not have provided the more experienced drivers in the study with any precursors (i.e. visual clues) to help them predict the occurrence of the hazard (Pradhan & Crundall, 2017), and thus could not have discriminated between safe and less-safe drivers. Provision of mirror footage in the current studies however, allowed such hazards to be included.

This touches on several other reasons why differences in findings might arise between research studies: there is no accepted standard for what constitutes a hazard, or how these clips should be edited and then presented, or even what response should be collected from participants. Many research teams adopt an individual approach to developing hazard perception tests, making it difficult to compare studies across different countries when they have employed different methodologies and used different sets of clips. In fairness, it should be noted that this is not necessarily a problem just across countries, as there are several studies conducted within the UK (including the studies conducted in Chapter 3 and 4 of this thesis) that have failed to replicate the basic behavioural discrimination between experienced and novice drivers (e.g. Crundall et al., 1999, Underwood et al., 2013).

5.2.3 Comparison of HP performance between UK and Malaysian drivers

To the author's knowledge, only Lim et al. (2013; 2014) have measured performance on the exact same test across two different countries[5]. In 2013 they compared Malaysian and UK drivers' hazard perception performance on clips filmed in both countries. They found that the UK drivers responded to many more hazards than Malaysian drivers, especially when they were presented with Malaysian clips. A difference between novice and experienced drivers did not materialise however (in both Malaysian and UK

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

participants). The authors suggested that cultural differences in hazard criterion (the internal threshold at which one considers an event to be a ‘hazard’) impacted more on test performance than experience. As Malaysian drivers typically encounter more hazards on the road than UK drivers, these events become normalised to the extent that a scenario must be extremely dangerous before they consider it to be a ‘hazard’, rather than simply an everyday event. The hazardousness of the Malaysian driving environment was also evident from the extra hazard responses reported by the participants. Malaysian clips received higher rates of extra hazard responses and UK drivers reported more extra hazard responses than the Malaysian drivers. Malaysian drivers appeared to have a higher threshold for hazards than UK drivers. However, without finding a difference between UK novice and experienced drivers, they could not firmly conclude that the hazard perception test could not transfer between countries (as they could not establish the effect in the UK in the first place with their clips).

5.2.4 Overview of the study

The current study will assess whether the hazard perception methodology could successfully discriminate between experienced and inexperienced drivers across three countries (China, Spain and the UK), paving the way for the design and export of a culturally agnostic test. While most studies of HP performance across countries use different stimuli and different test formats, the current study will use the same clips and identical methodology, across a cohort of participants recruited in the three countries. All participants saw three sets of clips, with one set filmed in the UK, one set filmed in Spain, and a third set of clips filmed in China (e.g. all UK participants saw clips from China, Spain and the UK, etc.). These three selected countries have very different cultures and traffic collision statistics. The World Health Organisation (2015) estimates the road fatalities of China, Spain and the UK to be 18.8, 3.7 and 2.9 deaths per 100,000 population, respectively, with an estimated 260,000 annual fatalities in China (though the officially reported number was just over 58,000 for 2013). Differences between officially reported statistics and WHO estimates reflect a number of measurement difficulties, such as trying to equate different definitions of a road collision fatality across different countries. For instance, in the UK an individual must die with 30 days of a collision to be counted, whereas in China the deadline for inclusion is seven days.

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

Key to this study was the requirement that the country-specific tests were as similar as possible in all other ways. Thus, all clips from each country were filmed and edited for this specific test, rather than co-opting previously captured video footage for inclusion. From the experience of filming in Malaysia, thought was given as to how best capture hazards that might not be suitable for the single-camera forward view favoured in the official UK and Australian tests. In order to accommodate the potential increase in overtaking hazards that may occur outside the UK, the use of mirror information (as used in Chapters 3 and 4) was thought to be essential. This was the only possible way to provide information about precursors to hazards that appear from behind the film-car. It was predicted that this test format would discriminate between experienced and inexperienced drivers in each country (using experience as a surrogate for crash likelihood), though there remained the possibility that experience might not show a difference in the UK sample due to the unsuccessful previous attempts.

It was also considered likely that experience might interact with the clip origin and participant nationality, such that any experiential effects may only be apparent in clips from the home country for those participants. Such findings would at least demonstrate that the test *format* is culturally agnostic (i.e. the test may discriminate between experienced and novice drivers using home country clips), if not the actual stimuli filmed across the three countries (i.e. cultural differences in the stimuli prevent the experiential effects from transferring to non-native contexts).

Finally, it is expected that the Chinese driving environment will be considered as the most hazardous when compared to the other two countries. This will be apparent by the number of extra hazard responses the Chinese clips will receive as well as the high hazardous rating scores (as there will be more potential hazards worth reporting). Furthermore, it is predicted that UK drivers will report more extra hazard responses than the Chinese and Spanish drivers, thus showing the differences in hazard identification thresholds.

This research was conducted collaboratively with the University of Granada and Tsinghua University. The current author was the lead researcher, and the collaboration was put together solely for the purposes of conducting research for this PhD thesis (i.e. this study was not already being undertaken as part of any other research project).

5.3 Method

5.3.1 Participants

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

Demographic details for each group can be found in Table 5.1. Over all three countries, the average experienced driver was 29.5 years old, passed the driving test in

2005 (with ten years of experience), and drove an average of 11804 miles per year. The average inexperienced driver was 21.1 years old, passed the driving test in 2014 (less than one year of experience) and had an annual mileage of only 63 miles.

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

Table 5.1: Mean demographic values for participants

Demographics	Chinese Participants		Spanish Participants		UK Participants	
	Novice	Exp'd	Novice	Exp'd	Novice	Exp'd
Total N (female N)	31 (13)	19 (2)	25 (17)	26 (8)	24 (13)	23 (21)
Age	22	28.6	19.2	35.9	22.1	24.1
Post-licence Experience (years)	1	6	1	17	1	7
Annual Mileage	82.8	4274.3	28	22041.9	78.2	9095.7

5.3.2 Materials and apparatus

To create the hazard perception stimuli filming was undertaken in China, Spain and the UK (for more information about the filming of the stimuli, see Chapter 2). The driver of the film car in each country was an experienced, native driver, with previous experience of conducting driving-safety research. Filming took place across a variety of times, but always in daylight and with clear weather conditions. The filmed environment in each country included city driving (Beijing, Granada, Nottingham), suburbs, and rural locations. Ten clips were chosen from each country to create the hazard perception test (with 30 clips selected in total). Clips varied in length from 31s to 64s and each clip included one *a priori* hazard identified by our team of transport researchers from across the countries. All hazards were captured naturalistically. In addition to the actual hazard (see Table 5.2 for a description of the individual hazards), these clips typically included several other potential hazard sources (i.e. precursors that did not develop into hazards, Pradhan & Crundall, 2017). Screen shots from each country can be viewed in Figure 5.1.

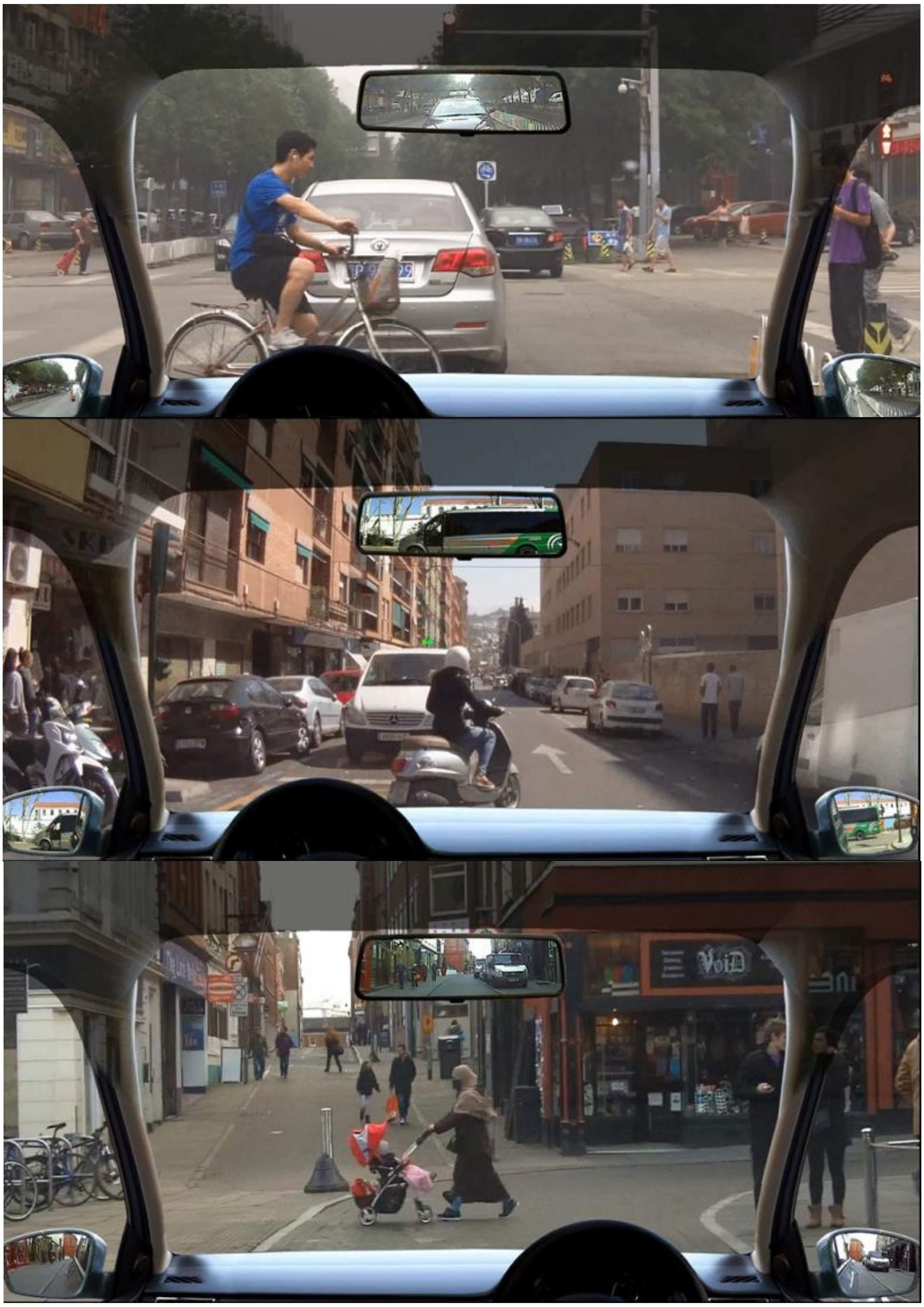


Figure 5.1: Three screen shots taken from hazard perception clips filmed in China (top panel), Spain (middle panel) and the UK (bottom panel).

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

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

Table 5.2: A description of the a priori hazards selected within each clip

Clip Number	Hazards	Duration of the clip (ms)
	CHINESE CLIPS	
1	A pedestrian is visible at the right edge of the road, looking to cross. The pedestrian is obscured by a turning car at the point of stepping into the road. By the time the pedestrian is visible again, he has already stepped into the road becoming a hazard.	55000
2	There are parked cars on both sides of a narrow street that might occlude pedestrians. A pedestrian steps into the road in front of you, from between two parked vehicles.	57000
3	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A cyclist emerges from the side road, obscured by the parked vehicles, and makes a wide turn in front of your vehicle, before cycling towards you.	26000
4	Your car slows on approach to a junction. A cyclist approaches from the left and is partially obscured by the A-frame of the semi-transparent graphic overlay. The cyclist cuts directly across your path.	63000
5	Your car is driving slowly and there are parked cars on the right. A parked car on the right side of your lane indicates late before attempting to pull out in front of you.	57000
6	A car immediately behind you, visible in the rear-view mirror and left side mirror, decides to overtake by entering a slip road to your left. It is forced to immediately pull back into your lane, in-front of you, as the slip road ends.	31000
7	A car behind you is visible in the rear-view mirror and right side mirror. The car undertakes you on the left by entering a bus lane. Once past you, it cuts into your lane and is forced to brake suddenly due to slowing traffic ahead.	50000

8	A car behind you is visible in the rear-view mirror and right side mirror. While attempting to exit a multilane road, the car from behind accelerates to undertake your vehicle, forcing you to hold off moving into the desired lane.	62000
9	A lorry approaches fast from the right-hand side. The lorry enters the main road from a side road on the right, cutting into your lane.	38000
10	A car behind you is visible in the rear-view mirror and left side mirror. The car indicates with the front lights. Then undertakes via a bus lane at speed, immediately cutting in front of your vehicle and braking.	55000
	SPANISH CLIPS	
1	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A motorcyclist emerges from the side road, partially obscured by the parked vehicles, and enters the main carriageway immediately in front of your vehicle.	41000
2	A pedestrian is stood in the road next to a parked car waiting for her friend to exit the car. As your car approaches the driver's door of the parked vehicle opens.	42000
3	On entering a side road, a rider on a scooter is checking over her shoulder in order to pull out around a vehicle blocking her lane. She then pulls out in front of your vehicle, as you finish turning into the side road.	34000
4	While travelling on a dual carriageway, in the distance a pedestrian enters from the left side of the road. The pedestrian continues to cross the street, forcing you to slow and stop.	37000
5	A car emerges from a side road and stops in front of you before indicating that it is going to reverse into a parking space at the road edge.	53000
6	A van is approaching fast from a side road on the right. The van tries to pull out and it halts abruptly when already partially out of the road, but nevertheless forces you to brake suddenly.	48000
7	A car ahead stops on a zebra crossing due to congestion ahead. A pedestrian, unable to cross on the actually crossing, steps into the road slightly in advance of the zebra crossing. As she steps out she is partially obscured by parked vehicles on the right.	42000
8	A double-length ('bendy') bus in the right lane indicates and pulls off from a bus stop immediately in front of you, after you have just exited from a roundabout.	50000
9	While driving on a dual carriageway, a motorcycle undertakes in the right lane and is forced to pull in-front of your car as traffic in the right lane slows due to congestion.	26000
10	A pedestrian is approaching from the left, partially obscured by a pillar. The pedestrian crosses the road ahead from the left.	47000
	UK CLIPS	
1	A car ahead overshoots a red traffic signal. As the cross traffic begins to enter the junction, the reversing light of the car turns on and the car ahead reverses towards you.	64000
2	You are driving in a street with shops and parked vehicles on the left side. There is a pedestrian coming out of one of the shops, approaching the street. The pedestrian steps out from between two parked cars on the left just as you accelerate after waiting in standing traffic.	49000
3	A distracted pedestrian is walking towards the street. The pedestrian crosses the road from the right without looking.	54000
4	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind undertakes you on the right on a multilane road.	43000

	Once past you, it cuts into your lane and is forced to brake suddenly due to a red traffic light.	
5	There are pedestrians on both side of a narrow street. A pedestrian with a child's push chair enters the road from the right without looking. Her entrance is partially obscured by pedestrians standing on the right.	32000
6	A bus in a bus lane signals to pull away from a bus stop. Due to parked vehicles ahead in the bus lane, it pulls out into your lane forcing you to stop.	31000
7	A car ahead emerges from a side road on the left and crosses your lane (too far away to be considered a hazard). It then indicates again and immediately cuts across your lane once more to park in a layby. This second lane crossing is close enough to your vehicle to constitute a hazard.	42000
8	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind overtakes your vehicle on a blind rural bend. The appearance of an oncoming vehicle in the opposite lane forces the overtaking vehicle to pull back into your lane immediately in front of you.	34000
9	While travelling at speed along a country road, a blind bend ahead reveals a queue of standing traffic, forcing you to slow and stop.	70000
10	While your car is slowing due to congestion ahead, a pedestrian looks over his shoulder before deciding to run in front of your vehicle forcing to slow more abruptly than otherwise required.	35000

5.3.3 Design

A 2 x 3 x 3 mixed design was used. The between-group factors were the driving experience of participants (experienced vs. inexperienced) and their nationality (Chinese vs. Spanish vs. UK). The within-group factor was the clip origin (China vs. Spain vs. UK). The dependent variables included the percentage of hazards that participants correctly identified and their response times to these hazards. Extra hazards responses (additional button-presses inside the hazard window or made outside the hazard window) and hazardous ratings were also measured.

Correct identification of a hazard was defined as a button response that fell within a temporal scoring window for each hazard. At the end of each clip participants were also required to rate each clip for the level of hazardousness presented on a Likert scale from 1-7, with higher numbers reflecting increasing levels of danger (with 'not at all hazardous' to 'extremely hazardous' as the anchors). Clips from the three countries were presented in three different blocks (10 clips per country).

5.3.4 Procedure

Participants were seated in front of the screen and viewed the on-screen instructions in their native language. They were asked to fill in a demographic questionnaire which included information such as age, sex, year of obtaining their driving license, driving collisions in the past 12 months, and miles/kilometres driven in the past 12 months. Participants were seated 60 cm from the screen.

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

5.4 Results

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

There were two main measures of interest: the percentage of hazards correctly identified (i.e. where a response was made within the temporal scoring window), and the response times associated with these mouse clicks. Both of these measures were subjected to 2 x 3 x 3 mixed ANOVA comparing performance across driver experience (experienced vs. novice drivers), participant nationality (Chinese, Spanish, or from the UK), and the origin of the clips (China, Spain or the UK). Between-group effects were further explored with Tukey tests while within-group effects were explored with planned repeated contrasts comparing Chinese drivers to Spanish drivers, and Spanish drivers to UK drivers (reflecting the rank order of these countries in road safety according to the World Health Organisation; WHO, 2015).

5.4.1 Response accuracy to hazards

Participants were considered to have correctly responded to a hazard if they pressed the mouse button within the hazard window for each specific clip (the mouse cursor was not visible on the screen and the location of the mouse was not important). Hit rates (accuracy) were calculated for each participant across the ten clips filmed in each of the three countries, and were turned into percentages for analysis via a 2 x 3 x 3 ANOVA. A main effect of drivers' nationality was found ($F(2, 142)=12.03$, $MSe = 1.96$, $p < 0.001$, $\eta_p^2 = 0.145$). Post Hoc Tukey tests revealed that UK drivers responded correctly to more hazards than Chinese drivers (78.9% vs. 64.5%, $p < 0.001$), while Spanish drivers also detected more hazards than Chinese drivers (73.3% vs 64.5%, p = 0.005). There was no difference between UK drivers and Spanish drivers.

A main effect was also found for clip origin ($F(2,284)=7.89$, $MSe = 1.69$, $p < 0.001$, $\eta_p^2 = 0.053$). Planned repeated contrasts found Chinese hazards to be less frequently identified than Spanish hazards (68.7% vs. 72.5%, p = 0.02), though there was no difference between responses to Spanish and UK hazards across all participants (72.5% vs. 75.1%, p = 0.09).

The omnibus interaction for participant nationality and clip origin approached statistical significance ($F(4,284) = 2.30$, $MSe = 1.69$, p = 0.059, $\eta_p^2 = 0.031$; see Figure 5.2), with the planned contrasts isolating the interaction component to the comparison of the responses to the Spanish and UK clips (*F*(2,142) = 4.92, *MSe* = 2.62, p = 0.009, η_p^2 =0.065).

To investigate this further, three Bonferroni-corrected, repeated-measures t-tests compared the performance on the Spanish clips to that of the UK clips for each of the three participant groups. Only the Spanish group were found to perform differently on the two clips sets, responding to more hazards in the UK clips than in the Spanish clips (78.8% 72.0%), t(50) = 3.3, p = 0.002). None of the other interactions or main effects, including any analysis that included driver experience, were significant.

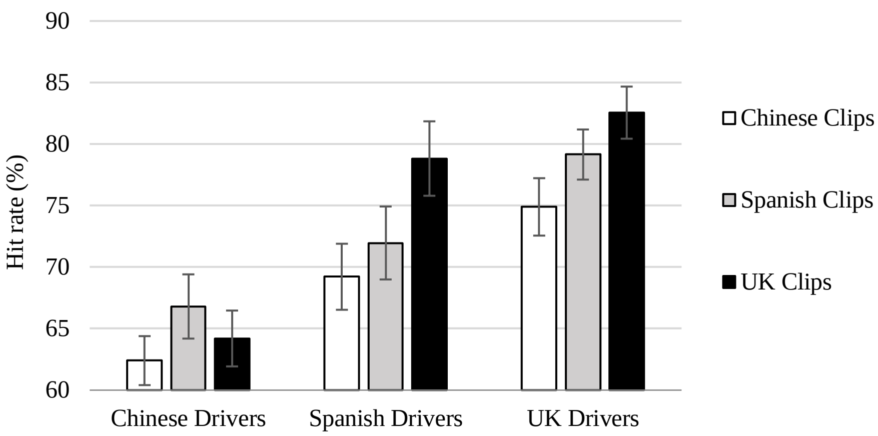


Figure 5.2: The percentage of hazards correctly responded to by all participant groups across the three countries of origin for the clips (with standard error bars added)

5.4.2 Response times to hazards

The transformed data (following Wetton et al., 2010, please refer back to Chapter 3) were subjected to a 2 x 3 x 3 mixed ANOVA. A main effect of participant nationality was found (*F*(2,142) = 18.46, *MSe* = 0.14, $p < 0.001$, η_p^2 = 0.206). Post hoc Tukey tests identified that all groups were significantly different to each other (all *p*s < 0.05) with UK drivers responding the quickest, followed by Spanish drivers, with Chinese drivers responding the slowest (with means of 2387 ms, 2829 ms, and 3120 ms, respectively, where these are reconverted from z-scores using group means and standard deviations).

An interaction was found between clip origin and participant nationality (F(4, 284) = 5.08, MSe = 0.10, p < 0.001, η_p^2 =0.067). Planned contrasts allowed comparison of Chinese clips with Spanish clips, and Spanish clips with UK clips, across the nationality factor. The interaction was found to lie between the Spanish clips and UK clips (F(2,142) = 9.38, MSe = 0.25, p < 0.001, η_p^2 =0.117). As can be seen in Figure 5.3, while the Chinese participants responded most slowly across all clips, they are markedly worse on the UK clips. None of the remaining main effects or interactions was significant. Thus, as with the analysis of the percentage of hazards responded to, there was no effect of driver experience.

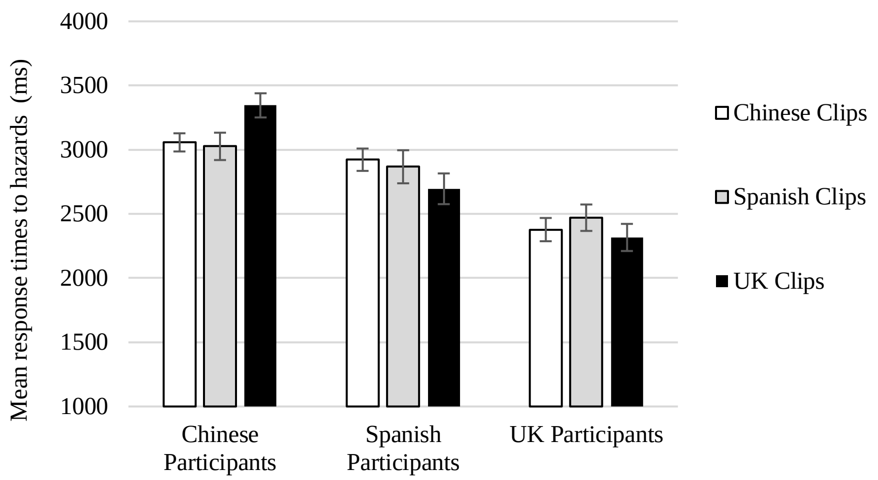


Figure 5.3: The mean response time of all three driver groups to hazards across the three clips sets recorded in each country (reconverted from square-root and z-score transformed means, with standard error bars added)

5.4.3 Extra hazard responses

In addition to the analysis of the two main DVs above, the number of additional responses that participants made while watching the clips was also calculated, above and beyond those responses that correctly identified the hazards. The extra hazard responses (EHR/m) measure was compared across clip origin, participant nationality and driver experience via a 3 x 3 x 2 mixed ANOVA. A main effect was found for participant's nationality ($F(2,142)=8.50$, MSe =1.44, $p < 0.001$, η_p^2 =0.107). Post Hoc Tukey tests revealed that differences were found specifically between Chinese and Spanish participants ($p < 0.005$) and between Chinese and UK participants ($p < 0.005$). Both Spanish and UK participants reported significantly more extra hazard responses per minute in comparison to the Chinese drivers. There were no differences between Spanish and UK participants.

A main effect was found for clip origin, $F(2,284)=86.53$, MSe =0.41, $p < 0.001$, η_p^2 =0.38. Chinese clips (M = 2.12) received the greatest number of extra hazard responses per minute, followed by the Spanish (M = 1.92) clips and finally the UK clips (M = 1.18). Planned comparisons confirmed significant differences between the Chinese and Spanish clips ($F(1,142)=7.49$, MSe = 0.76, $p < 0.005$, η_p^2 =0.050) and between the Spanish and UK clips ($F(1,142)=107.49$, MSe =0.74, $p < 0.001$, η_p^2 =0.431). The UK clips were considered to have significantly fewer hazardous events worth reporting than the Spanish clips, which in turn had fewer hazards than the Chinese clips.

A significant interaction was found across clip origin and nationality ($F(4,284)=5.53$, MSe =0.41, $p < 0.001$, η_p^2 =0.072). Planned comparisons showed that the interaction lay between the Chinese and Spanish clips ($F(2,142)=6.06$, MSe =0.76, $p < 0.005$, η_p^2 =0.079). As can be seen in Figure 5.4, both the Chinese and UK participants viewed the Chinese clips and the Spanish clips to be equally hazardous in terms of EHR/m. Spanish participants, however, made more responses to the Chinese clips compared to the Spanish clips. The UK clips appeared to be considered as the least hazardous by all three countries.

A significant interaction was also found across clip origin and driver experience, ($F(2,284)=3.21$, MSe =0.41, $p < 0.05$, η_p^2 =0.022. Planned comparisons revealed the interaction to lie in the comparison of Chinese and Spanish clips (F(1,142)=4.52, *MSe* =0.76, $p < 0.05$, η_p^2 =0.031), with experienced drivers making a greater number of extra hazard responses per minute to Chinese clips than to Spanish clips. There was no difference in the extra hazard responses of the inexperienced drivers to Chinese and Spanish clips (Figure 5.5).

A Pearson correlation was calculated to assess the relationship between the percentage accuracy of each participant across all clips, and the number of extra hazard responses per minute averaged over all clips. There was a significant correlation between the two variables (r(146)=0.37, $p < 0.001$). This suggests that a successful hit rate appears related to the overall number of extra hazard responses that participants made, despite having already removed the four UK participants who were considered 'excessive responders' (more than 3 SDs above the average participant per block, i.e. more than 60.4 responses during the 10 clips from a particular country).

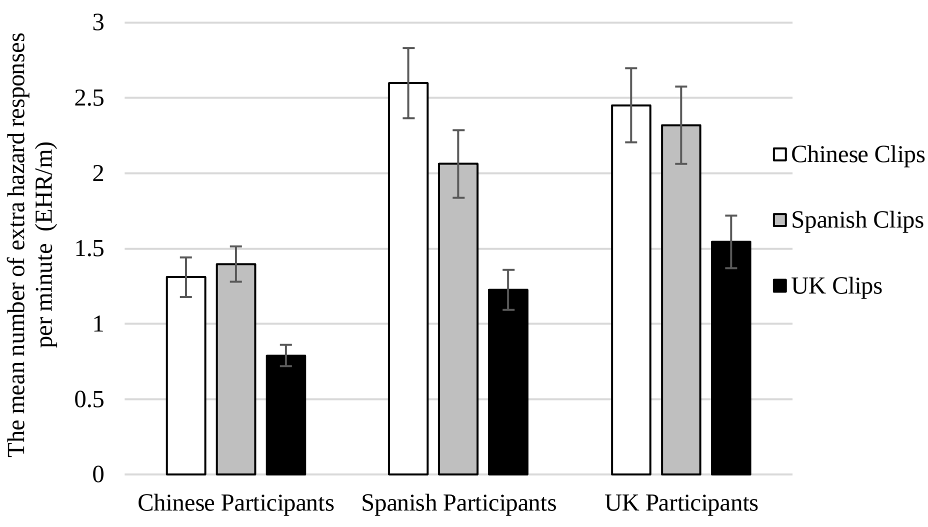


Figure 5.4: The mean number of extra hazard responses per minute for each block of clips across the three groups of participants (with standard error bars added)

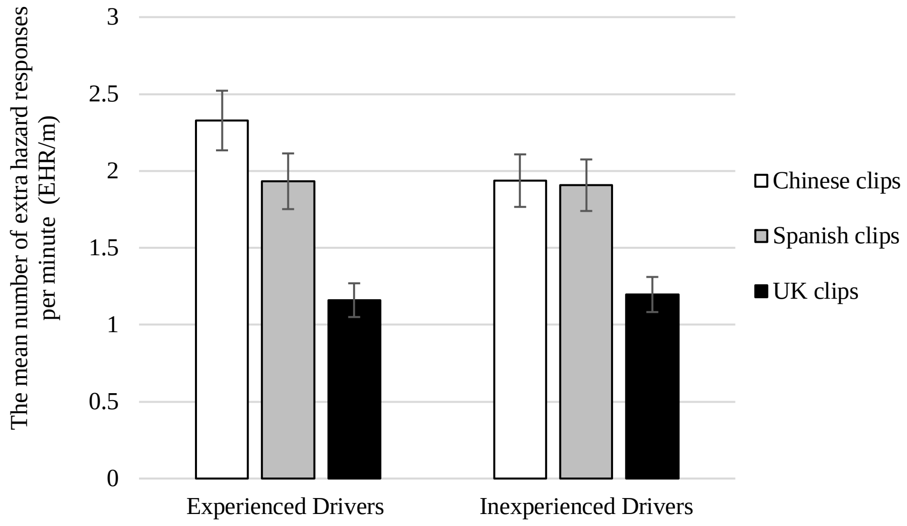


Figure 5.5: The mean number of extra hazard responses per minute for each block of clips across experienced and inexperienced drivers (with standard error bars added)

5.4.4 Hazardousness ratings for all hazard clips

Following each clip, participants were asked to provide a hazardousness rating on a scale of 1 to 7 (where higher numbers reflect greater levels of perceived hazardousness). These ratings were averaged across clips for each participant according to country of origin and were subjected to a 2 x 3 x 3 ANOVA (participant experience vs. nationality vs. clip origin) (see Norman, 2010, for a justification for using ANOVA to analyse these data).

There was a main effect of clip origin ($F(2,284) = 62.1$, $MSe = 0.263$, $p < 0.001$, $\eta_p^2 = 0.30$). Repeated planned contrasts compared clips from the three countries and found that Chinese clips were rated as more hazardous than Spanish Clips (4.5 vs. 4.14; $F(1, 142) = 30.9$, $MSe = 0.62$, $p < 0.001$, $\eta_p^2 = 0.179$), and the Spanish clips were rated as more hazardous than the UK clips (4.14 vs. 3.84; $F(1, 142) = 26.5$, $MSe = 0.52$, $p < 0.001$, $\eta_p^2 = 0.157$).

Nationality also produced a significant main effect ($F(2, 142) = 20.7$, $MSe = 0.78$, $p < 0.001$, $\eta_p^2 = 0.226$), with Spanish drivers giving the highest ratings (4.73) followed by UK drivers (4.14), with Chinese drivers giving the lowest hazardousness ratings (3.60). Tukey tests revealed each possible comparison to be significant (ps < 0.01). The final factor of driver experience did not reveal a difference between the two groups (4.2 vs. 4.1 for experienced and inexperienced drivers, respectively).

An interaction was noted between clip origin and participant nationality ($F(4, 284) = 5.17$, $MSe = 0.263$, $p < 0.001$, $\eta_p^2 = 0.07$). As can be seen in Figure 5.6, only Spanish drivers rated the Chinese clips as more hazardous than the Spanish clips, mirroring the EHR/m results. Also, while both Chinese and UK participants rated Spanish clips as more hazardous than UK clips, the Spanish drivers did not make this distinction.

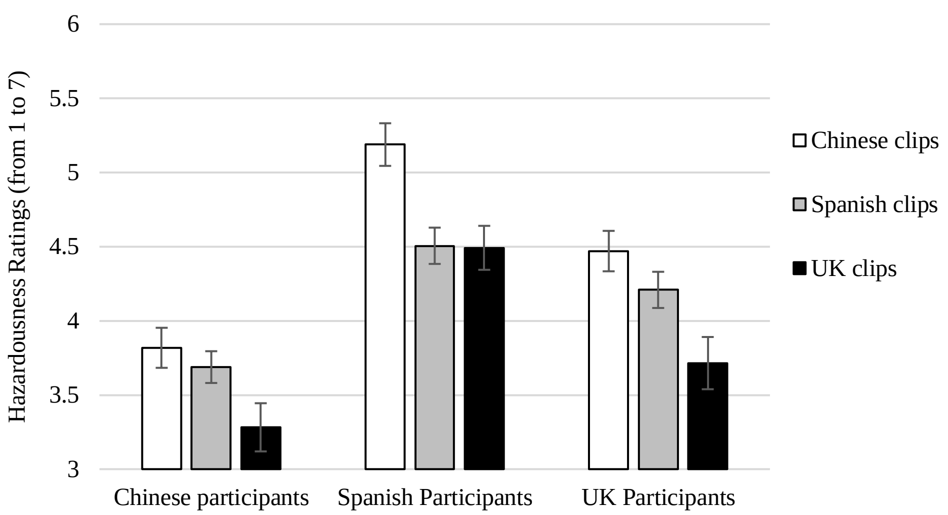


Figure 5.6: The hazardousness ratings of participants across three nationalities and the clip sets from each country. Ratings were given on a 1-7 scale, where higher ratings indicate greater perceived hazardousness (standard error bars added)

5.5 Discussion

The aim of this study was to assess whether the hazard perception test is suitable for export to different driving cultures. This was a novel endeavour as, though many research groups around the world have investigated hazard perception in their own countries, they have done so with vastly differing methodologies making it difficult to compare the validity of the tests across different regions. Only one previous attempt has been made to assess hazard perception skills of drivers from different countries using the same clip set, but the results of that study were inconclusive (Lim et al., 2013). Therefore, three different countries were tested using an identical hazard perception methodology in order to find out whether the test will discriminate between experienced groups from different cultures. The current experiment revealed no differences between experienced and inexperienced drivers in regard to the two main dependent measures: response times to hazards and the percentage of hazards correctly responded to. Similarly to Lim et al. (2013), it is hard to conclude that the hazard perception methodology is suitable for export to other countries when we cannot discriminate between any driver group on the basis of experience, regardless of nationality. If at least the UK clips could produce a difference between UK experienced and inexperienced drivers, then one could feel comfortable that the basic test replicated previous work in the field, but this was not the case. Admittedly, in order for this effect to have risen to our attention it would have had to evoke a three-way interaction between participant nationality, clip origin and driver experience. Failures to hit such high goals may always raise suspicions of a lack of statistical power, but even when UK experienced drivers' accuracy rates and response times are directly compared to those of UK inexperienced drivers (just using UK clips), even these simple t-tests cannot reach significance ($M_{\text{experienced accuracy}} = 8.13$ vs. $M_{\text{inexperienced accuracy}} = 8.38$, $t(45) = -0.57$, $p = 0.57$; $M_{\text{experienced RTs}} = 2373.6$ ms vs. $M_{\text{inexperienced RTs}} = 2402.4$ ms, $t(45) = 0.13$, $p = 0.88$). One possibility was that the scoring windows defined for this study (including Experiments 1 and 3) were too lax. Therefore, more conservative scoring windows were set and additional reaction time and hit rate analyses were performed. This analysis was a minor one and depending on the hazard, either the onset or offset were slightly changed (by removing a certain amount of frames and making the window "smaller"). However, the results were identical as the reported ones, with no significant differences between the experienced groups. It is possible that more significant changes to the window would have produced effects. However, examining different types of scoring windows was not the aim of this chapter and considering that the results were not different from the original, this analysis has not been described in the methods and analyses section.

While this lack of significance contradicts many studies that have previously demonstrated such HP tests to discriminate between experienced and inexperienced drivers (e.g. Wallis & Horswill, 2007; Horswill et al., 2008; Deery, 1999), it has been already noted in the introduction that failure to find this effect is not without precedent (e.g. Sagberg, & Bjørnskau, 2006; Lim et al., 2013; Yeung & Wong, 2015).

Despite the failure to find experiential differences, several other interesting findings were noted that suggest the typical HP approach might be culturally sensitive. First, it was notable that Chinese drivers made fewer hazard responses over all the clips, both in terms of identifying the *a priori* target hazards, and in their additional responses to other potential hazards (EHR/m). In contrast, Chinese clips evoked the greatest number of extra hazard responses per minute across all participants (most strongly within the Spanish participant group). This suggests that the Chinese clips are more complex, and contain more precursors than the Spanish and UK clips. This in itself is unsurprising as China has the highest collision rate of the three countries, and is therefore likely to have more potential hazards. Of greater interest is the possibility that our Chinese drivers, who are continuously exposed to a higher frequency of potential hazards, were therefore less sensitive (or more accepting) of hazards from all three countries. This possibility once again raises the problem of criterion bias influencing the simple push-button response required in the traditional hazard perception methodology.

Chinese drivers were also slower to respond to hazards across all three countries than other drivers, while the UK drivers were the fastest. The slow responses of Chinese drivers may stem from their high threshold for reporting hazards, which seems apparent in their lower frequency of hazard responses, though the faster responses of the UK drivers may reflect their previous exposure to the national UK test.

The high threshold of Chinese drivers when responding to hazards is also reflected in the hazardousness ratings provided by participants following each clip.

These ratings mostly follow the pattern of extra hazard responses per minute (EHR/m), with Chinese clips being rated as the most hazardous across all participants, yet with Chinese drivers providing the lowest ratings of hazardousness across all clips. Thus, the explicit ratings of hazardousness again support the notion that the Chinese drivers face the greatest level of on-road complexity, and are correspondingly desensitised to high levels of hazardousness.

Even though the Chinese clips evoked the greatest number of EHR/m, they were also noted to produce the fewest responses to the *a priori* hazards across all participants. This suggests that the number of potential precursors available in the environment might degrade ability to identify the correct target (Crundall, 2016). This could be due to an increase in distracter stimuli that demand attention away from the target, similar to increasing set size in a visual search array (Palmer, Ames & Lindsey, 1993; Dobkins & Bosworth, 2001; Wolfe, 1994), or perhaps due to an attentional depletion in processing resources, such as an attentional blink (e.g. Chun, 1997, Lagroix et al., 2012). In the face of such demanding visual complexity it is understandable that Chinese drivers might increase their threshold for deciding whether something poses an explicit hazard.

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

This raises the possibility of developing a cohesive and culturally-agnostic typology of hazards. Some attempts have been made in the literature to distinguish between coarse categories of hazards (e.g. latent vs. overt hazards; developing vs.

abrupt hazards; behavioural vs. environmental prediction hazards; see Pradhan & Crundall, 2017, for a review), but there is an opportunity to classify hazards at a finer level. It is likely that some hazards will be more effective discriminators of driver safety than others (e.g. Crundall, et al., 2012; Crundall, 2016). If a hazard typology can have a degree of consistency across cultures, then this increases the value of developing such a system of categorisation.

To summarise, it appears that this current hazard perception test is not appropriate to export to other countries. First, it does not discriminate between experienced and inexperienced drivers which is considered to be a mainstay of test validity. While disappointing, this is not necessarily insurmountable. Some hazards are likely to be more successful in discriminating between experienced and inexperienced drivers (Crundall et al., 2012; Crundall, 2016) and it could be possible to collect new hazards that add to the discriminative ability of the test. Therefore, a clip by clip comparison was graphed to compare raw RTs of driving groups at least within their home countries (see Figure 5.7 a, b, c). At first glance, only Spanish experienced drivers seem to react faster to the hazards (although not significantly). In addition, it is noticeable that experienced drivers often perform similarly to novices. This is especially apparent for the UK sample where on some occasions novices were faster than experienced drivers (e. g. hazards involving pedestrians or abrupt hazards such as queue of standing traffic). As a result, the results of the UK sample are consistent with those in Chapter 3 and 4 failing to find the basic experiential effect.

However, in each country there is at least one hazard that shows considerable difference between experienced and novice drivers. Interestingly, experienced drivers from either country reacted much faster to hazards that contained indicator signals, reversing lights or braking lights (UC 1, 7, CC 5, SC 5). This supports, in part, that some hazards (especially those containing precursors) are more successful in discriminating between experienced and novice drivers, and is consistent with the findings of Crundall et al. (2012). Nonetheless, a valid test should contain a variety of hazards and it is expected that most of these hazards should be able to discriminate between the experienced groups. At least this is the premise behind the HP test.

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

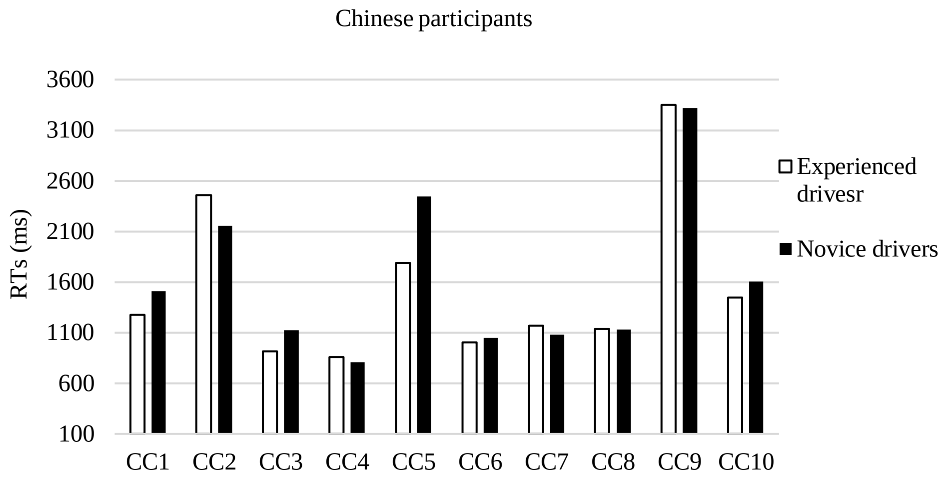


Figure 5.7 a: RTs of the Chinese participants across all 10 Chinese clips

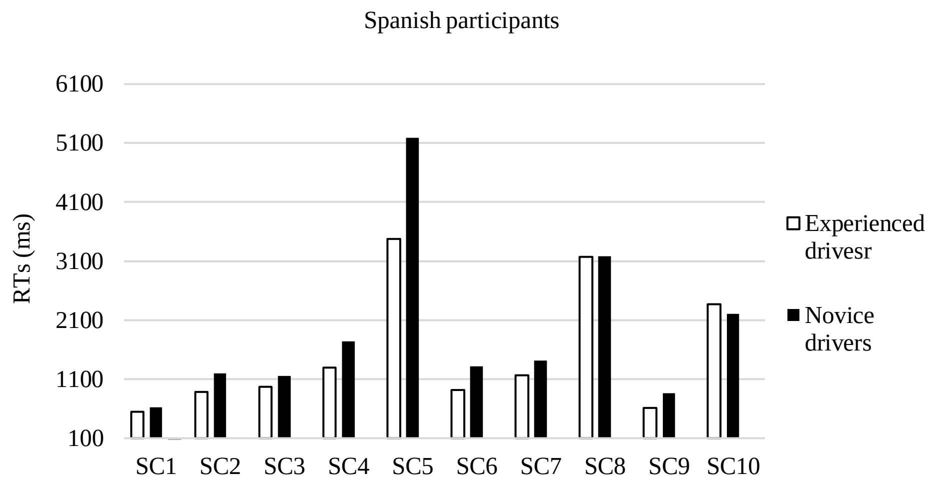


Figure 5.7 b: RTs of the Spanish participants across all 10 Spanish clips

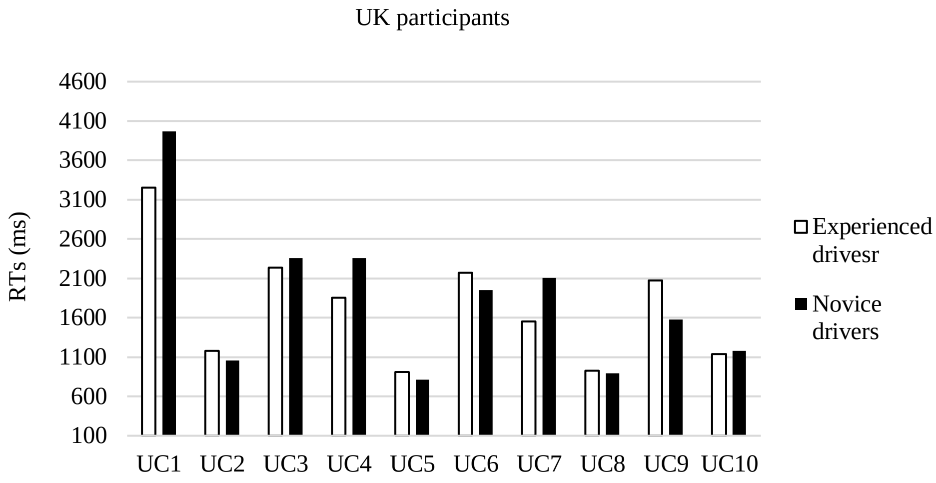


Figure 5.7 c: RTs of the UK participants across all 10 UK clips

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

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

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

In conclusion, the current hazard perception test, based on the traditional UK methodology produces more differences between groups on the basis of nationality than driving experience, and is greatly influenced by the context of clips. No differences were observed between the UK experienced and novice drivers. There was not a significant difference even within the samples from China and Spain which might suggest that it is the measurement and interpretation of simple response times that leads to problems related to criterion bias and overestimation of novices' HP skills. It should be acknowledged however that this is not the only HP methodology that we could have implemented, and that the variations employed by many other researchers may have produced a better test. However, when considered alongside the problems of criterion bias, and issues related to the measurement and interpretation of simple response times, there appears to be little evidence that allows us to commend the export of this particular hazard perception test methodology to other countries. Rather than creating a culturally-agnostic test of drivers' higher-order cognitive skills, we have created a culturally sensitive measure that cannot yet discriminate between safe and less-safe drivers based on experience.

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

Chapter 6

THE HAZARD PREDICTION TEST

Experiment 5

Free-response hazard prediction test: occlusion points

The material covered in this chapter has been adapted from previously published paper in Ventsislavova, P., & Crundall, D. (2018). The hazard prediction test: A comparison of free-response and multiple-choice formats. *Safety Science, 109*, 246-255. https://doi.org/10.1016/j.ssci.2018.06.004 (see Appendix D)

6.1 Summary of previous findings

The hazard perception test is considered one of the most successful initiatives in the UK which has demonstrated both retrospective and prospective sensitivity to crash-likelihood. Suggestions have been made for international export of the hazard perception methodology aiming to repeat its success in different countries. Several research groups from different countries have developed their own hazard perception test, though they have used different methodologies which makes cross-cultural comparisons difficult. Only one previous attempt has been made to assess the hazard perception skill of drivers cross-culturally using the same clip set, however with inconclusive results (Lim et al., 2013; 2014). Therefore, in Experiment 4, drivers from three different countries (China, Spain and the UK) were compared using identical HP methodologies. The results revealed considerable differences in the driver groups. For instance, the Chinese participants were the slowest to respond to the hazards, identifying significantly fewer hazards, and made fewer responses overall, in comparison to the Spanish and UK drivers. Conversely, UK drivers showed faster responses and identified the most hazards. Despite finding these effects of participant nationality, there were no significant differences between experienced and inexperienced drivers in regard to their hazards responses. Both groups performed identically in the test and, therefore, it cannot be concluded that the traditional methodology of the hazard perception test is transferable to other countries, as test-validity could not be established in the UK sample in the first instance. Not only there were no experiential differences, but the cultural differences of the driving groups appeared to significantly influence the way they approached the test. Chinese drivers rated the clips as less hazardous than the other driver groups, which may account for their slower response times. UK clips were considered especially non-hazardous by the Chinese drivers, reflecting their extremely slow HP responses to this clip set. This was ostensibly due to differences in cultural hazard thresholds. On the basis of the higher traffic collision statistics in China, compared to the UK and Spain, it is safe to assume that Chinese drivers are likely to encounter many more hazards on the road in every day driving. This increased exposure to hazards presumably desensitises the Chinese drivers to the relative seriousness of some hazardous events, increasing their thresholds for reporting them.

In addition, a correlation was identified between the number of *a priori* hazards that participants responded to within the scoring window and the overall number of extra hazard responses that participants made. This raises a clear concern for the traditional hazard perception methodology, as it appears that the high performance of individuals may be influenced by clicks falling within the scoring window that do not necessarily reflect responses to the *a priori* hazard. The national UK test tries to dissuade excessive clicking by giving a score of zero if participants click too frequently, but it is not clear where this threshold is, or whether it is effective in reducing the relationship between overall frequency of clicks and hazard hit rate.

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

As the traditional hazard perception test failed to find differences between the experienced groups, an alternative *hazard prediction* test was created in Experiment 5 based on initial studies that have already been conducted in the UK and Spain (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009). The hazard clips were edited to occlude just as the hazard begins to develop, and participants were asked ‘What happens next?’.

6.2 Introduction

6.2.1 The hazard prediction test

The hazard prediction test differs from the traditional hazard perception test in that it forgoes response times in favour of accuracy for predicting what happens next following an occlusion that occurs just as the hazard begins to develop. It is argued that this test format removes several potential problems associated with recording response times to hazards (Crundall, 2016). For instance, traditional response-time measures require a scoring window to be defined. If a response is made between the onset and offset of a hazard, then the response is considered to be correct. However, there are no clear guidelines on how to define onsets and offsets, and there is always the possibility that excellent drivers will spot very subtle cues to upcoming hazards, and respond just before the scoring window (which would be counted as a miss). Even if drivers do press within the scoring window, we do not know if they are responding to the actual hazard, or to some other less hazardous aspect of the scene (see Crundall, 2016, for an argument as to why localised hazard responses are not necessarily a suitable solution for a lack of accuracy in the traditional test). Finally, ‘hazard perception’ is confounded by post-perceptual processes, such as criterion bias: expert drivers may delay or refrain from responding to hazards because they believe the unfolding event falls within the boundaries of their driving skill (Pradhan & Crundall, 2017). The hazard prediction test (or ‘What happens next?’ test) mitigates these confounds by removing reliance on response times, replacing them with the accuracy of drivers to predict what happens next following occlusion of the developing hazard. This removes the need for scoring windows, and the reliance on an internal threshold to judge whether ‘what happens next’ is actually a hazard.

A number of studies have demonstrated the ability of the hazard prediction test to successfully discriminate between safer, experienced drivers, and less-safe, inexperienced drivers (Jackson et al., 2009; Crundall, 2016; Castro et al., 2014, 2016, Ventsislavova et al, 2016, Gugliotta et al, 2017; Lim et al., 2014). Several of these studies have also developed this occlusion-based methodology through a number of targeted experiments focusing on design elements. For instance, Jackson et al. (2009) demonstrated that an occlusion is necessary to discriminate driver groups, rather than just pausing on the final frame. A freeze-frame provides an unrealistic amount of time for novice drivers to identify clues to the impending hazard, whereas an occlusion ensures that the driver must be looking in the right place at the right time. As safer drivers are more likely to prioritise those areas of the scene that may develop into hazards, the occlusion is therefore more likely to identify the safest drivers (Crundall & Kroll, 2018).

Crundall (2016) addressed a number of methodological questions, including the impact of clip length on predictive accuracy. He found that longer clips resulted in lower prediction accuracy, especially for novice drivers, suggesting that novices suffer a greater vigilance decrement over time. In a separate experiment, Crundall manipulated the occlusion point. The results demonstrated a decline in prediction accuracy as the occlusion point became more temporally distant from the hazard. The novice/experienced driver distinction remained however and did not interact with the occlusion point. Thus it seems that hazards can be extrapolated from relatively early information (in that particular case, over a second prior to hazard onset), though at a reduced level of accuracy. Participants' confidence ratings in their predictions fall to baseline levels however at the most distal occlusion points.

These initial studies suggest that the hazard prediction test can provide a robust and simpler alternative to the more traditional hazard perception test. To this end, a new hazard prediction test was developed for the current study, as an alternative to the response-time hazard perception tests used so far. It was hoped that this new test might finally evoke some experiential differences.

This new test employed the UK clips used in the previous studies. Novice and experienced drivers were tested on this new prediction task in order to assess whether it could discriminate between these groups. If this result is found, it paves the way for using this alternate test variant in across different countries. This study required free responses, to the question 'What happens next?' to be typed into a response box, following the methodology of Jackson et al., (2009).

6.2.2 Overview of the study

To validate this new version of the test experienced and inexperienced drivers were recruited to take part. In addition to the basic validation of this test, one of the questions raised by Crundall (2016) was revisited: when should one occlude the clip to maximise discriminability? As noted above, Crundall found that occlusions that were proximal to the hazard resulted in the most accurate predictions. While the most distal occlusions (an average of 1200 ms prior to hazard onset) significantly reduced accuracy, participants were still able to predict the hazard on 57% of trials (with a free response). Surprisingly, the change in occlusion point did not affect the discriminability of the driver groups. One might suppose that experienced drivers would fare relatively better than novices at more distal occlusions, as they may make better use of weaker hazard evidence (Pradhan & Crundall, 2017). However, as occlusions become extremely separated from the hazard all drivers must eventually reach the same nadir, in the absence of even subtle cues to the nature of the upcoming hazard. Neither of these eventualities occurred in the study of Crundall (2016), possibly due to the particular occlusion points that were chosen and the nature of the individual hazards in those clips. Given the need to validate the new free-response version of this prediction test on experienced and inexperienced drivers, this also provided the opportunity to investigate whether a novel set of stimuli produced different findings when the occlusion points were varied across three levels: proximal (temporally closest to the hazard), intermediate, and distal occlusion points.

6.3 Method

6.3.1 Participants

Sixty-one participants took part in this experiment. The sample was divided into experienced and inexperienced drivers. Experienced drivers (N = 30, mean age = 21.5 years) had three years of mean driving experience since passing their driving test and inexperienced drivers (N = 31; mean age = 19.5 years) had less than one year of post-test driving experience, and included 15 learner drivers. Mean mileages for experienced and inexperienced drivers were 4553.3 miles and 290.9 miles in the year prior to the study, respectively. Participants were also asked to report whether they had suffered any traffic accidents during the past 24 months. However, only five participants reported that they have been involved in a traffic collision during that period (two novices and three experienced). All participants had normal or corrected-to-normal vision and were university students and staff.

6.3.2 Design

A 2 x 3 between-group factorial design was employed, where the independent variables were the driving experience of participants (experienced vs. inexperienced) and the occlusion points (proximal vs. intermediate vs. distal). Five cells contained ten participants, with the distal/inexperienced group containing 11 participants. The proximal occlusion point was the closest to the onset of the hazard, and would typically show several frames of the hazard following onset. This meant that participants who were looking in the right place at the right time, would have their predictions confirmed. For instance, if the participant believed the pedestrian ahead may step into the road, a fixation on this pedestrian immediately prior to occlusion (proximal condition only) would confirm this, as the pedestrian would be seen to move a leg in the direction of the road. Intermediate occlusion points were an average of 618 ms earlier than the proximal occlusion points, and would contain evidence of precursors, but without any confirmation of the hazard triggering. In this case, the pedestrian might be seen walking towards the edge of the pavement, perhaps turning her head in the direction of the film car. Finally, the distal occlusion points were an average of 1222 ms earlier than the proximal points. Typically, these would still contain precursors to the hazard, but the evidence of hazardousness would be weak. For instance, the pedestrian may be visible on the pavement, but still be walking parallel to the road with no obvious intention to cross. See Figure 6.1 for an alternative example of these three occlusion points.

The dependent variable was the percentage of the accuracy with which participants correctly identified the hazards across a total of 15 clips. Following occlusion, participants were presented with a free-text entry box to provide their answer. These typed responses were subsequently coded with one point awarded for a correct answer. Participants were instructed to consider the identity of the hazardous object, it's location within the scene, and how the event would unfold (e.g. "the pedestrian on the left is about to step into the road in front of me"). Where participants failed to report the three suggested items in their answer, but it was still unambiguously correct, they were still awarded the point. For instance, the response "The pedestrian is about to step into the road" could still be awarded the point if this was the only visible pedestrian, even though the response failed to locate the hazardous object as being on the left-side pavement. Two independent evaluators scored each answer. Cohen's kappa was 0.88 for the distal occlusion point, 0.88 for the intermediate occlusion point, and 0.80 for the proximal occlusion point. Where raters disagreed in their initial rating, a final score was agreed via discussion.

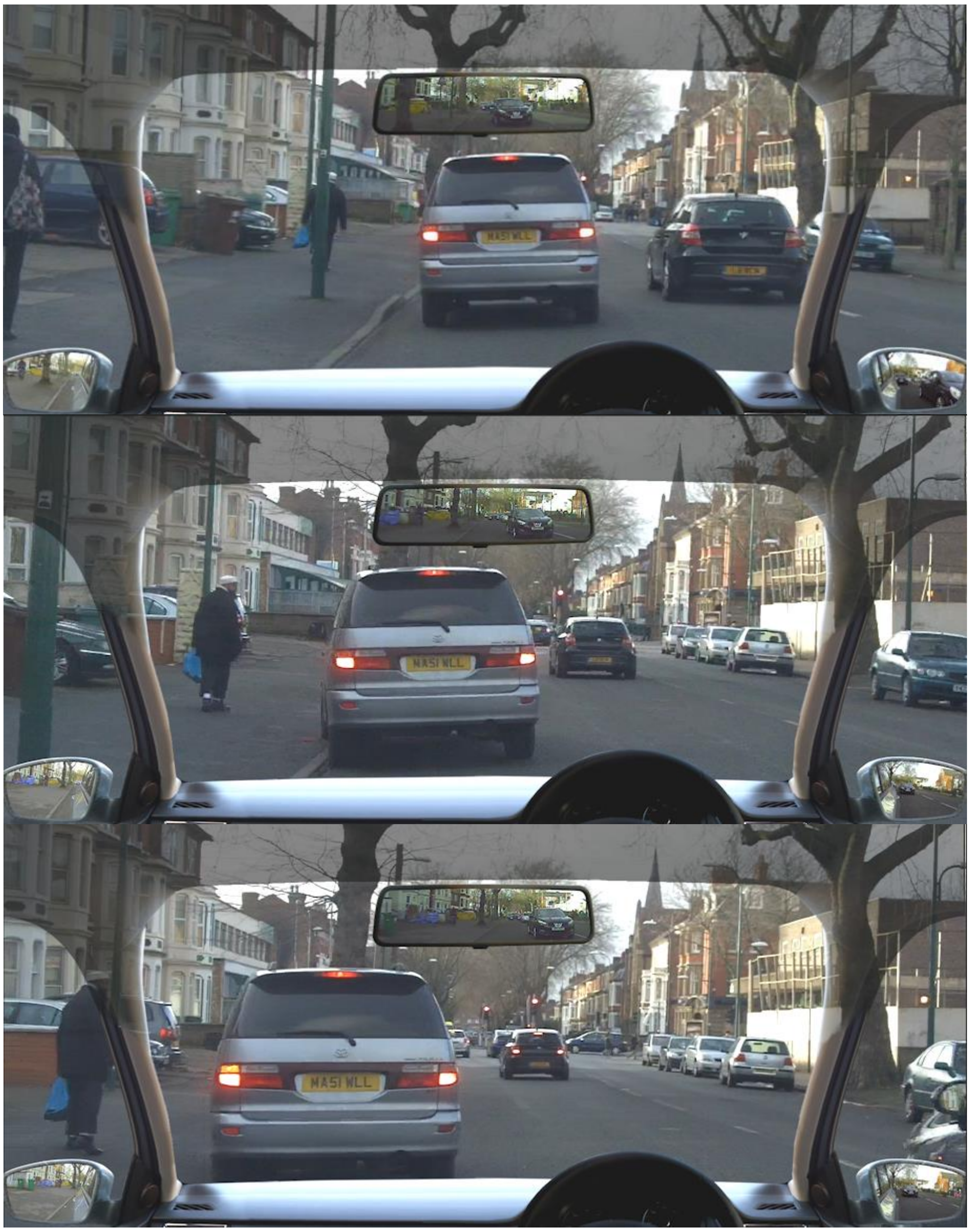


Figure 6.1: Three frames illustrating the three occlusion points for one clip. While driving along a road, the car in front pulls over. This requires the film car to overtake the stopping vehicle. However, a safe driver should be aware that there is a vehicle visible in the right side mirror that might pose a problem (top panel; distal occlusion). As the clip progresses, the car in the side mirror disappears, as it moves into the blind spot ready to overtake. Thus, the evidence that this vehicle is going to pose a problem increases considerably (middle panel; intermediate occlusion). Finally, the overtaking car becomes visible in the forward view confirming that the car from behind is indeed overtaking and poses a hazard for the film car's intended action (bottom panel; proximal occlusion)

6.3.3 Materials

Fifteen UK clips were selected from a corpus filmed specifically for this thesis (all of which have been used in previous studies in this thesis). Five of the clips were removed due to not being suitable for a hazard prediction test. None of these clips contained a precursor which could have helped identify how the hazardous situation was going to develop. Table 6.1 lists the hazards. The duration of each clip varied between 31s and 64s prior to editing the occlusions.

The clips were edited to occlude (i.e. cut to an immediate black screen) at three separate points: proximal, intermediate and distal. A proximal cut would allow several frames of the actual hazard to be seen (providing the participant was already looking at the location of the imminent hazard), confirming the participant's prediction. Intermediate occlusions and distal occlusions occurred an average of 618 and 1222 ms earlier than the proximal occlusion, respectively.

Clips were displayed on a Lenovo (ThinkPad) computer running E-Prime 2.0 Software (Psychology Software Tools, 2012), with a resolution of 1920x1080 and screen size of 34.5cm x 19.5cm. Participants responded using the keyboard.

6.3.4 Procedure

Once participants gave consent, they were seated 60 cm from the screen and asked to answer demographic questions, including age, sex, year of obtaining driving license, driving collisions in the past 24 months, and miles driven in the past 12 months. Participants were then given both verbal and on-screen instructions for the test. They were told that they would watch 15 short video clips from a driver's perspective. They were asked to watch each clip carefully, as at some point the clip would stop and be occluded by a black screen. Following this, they were asked to describe "What happens next?" (i.e. how the driving situation was going to develop), by typing their answer in a free-response box on the screen (with a 150 character limit). They were instructed to include in their answer any source of potential hazard, the location of that source, and how the situation was going to develop, though the emphasis was on predicting what happens next, regardless of how hazardous the event might be. Following each answer, participants were asked to rate how hazardous they thought the predicted event would be for them, using a 7-point Likert scale (where 1 is 'not hazardous at all', and 7 is 'extremely hazardous'). The clips appeared in a randomised order.

Participants then watched a practice trial where they had the opportunity to familiarise themselves with the task. This practice trial also included feedback: following occlusion and their answer, the clip was replayed without an occlusion allowing participants to see what the hazard actually was. This practice trial was the only time when they received feedback about their performance. Once participants felt comfortable with the instructions and the practice, they began the experiment. In total, the experiment took an average of 20 minutes.

Table 6.1: A description of the *a priori* hazards selected within each clip

Nº	Video clips (*with occlusion points italicised*)	Hazard	Duration (ms)
1	A car ahead overshoots a red traffic signal. As the cross-traffic begins to enter the junction, the reversing light of the car ahead illuminates, and this car then reverses towards you to get out of the junction. *The clip occludes following initial illumination of the reversing light.*	Reversing car ahead	64000
2	As you are driving, the car indicates and pulls over blocking your path. As you try to overtake the car and continue your path, a car from behind overtakes you (passing from the right mirror, into the blind spot, before emerging into your forward view). *The clip occludes at the point where the overtaking car enters the blind spot.*	A car from behind overtakes you	34000
3	You are driving along a narrow urban street and the head of a pedestrian is visible above the parked cars on the left. *The clip occludes as the pedestrian moves between the parked cars to step into the road.*	The pedestrian from the left crosses the street	34000
4	You come to a standstill due to congestion in an urban street with shops and parked vehicles. A pedestrian exits a shop and then steps out from between two parked cars on the left, just as the traffic moves and you begin to accelerate. *The clip occludes when the pedestrian turns his head to look at you, while stepping forward.*	A pedestrian from the left crosses the street	49000
5	A distracted pedestrian is walking towards the street. The pedestrian crosses the road from the right without looking. *The clip occludes as the pedestrian begins to cross.*	A distracted pedestrian crosses the street from the right	54000
6	A car behind you is visible in the rear-view mirror and left side mirror. As you approach traffic lights, this car undertakes you and pulls in front of you, forcing you to brake. *The clip occludes when the car is just emerging from the blind spot.*	The car from the left undertakes the camera car	43000
7	There are pedestrians on both side of a narrow street. A pedestrian with a child's pushchair enters the road from the right without looking. Her entrance is partially obscured by pedestrians standing on the right. *The clip occludes when the pushchair is partially visible among the pedestrians.*	A pedestrian with a child's pushchair enters the road from the right without looking	32000

8	You are driving along a narrow urban road. A pedestrian approaches from the right with shopping bags. He is temporally obscured by a large plant pot, before he steps out into the road. *The clips occludes just as he begins to step out from behind the plant pot.*	A pedestrian from the right crosses the street without looking for upcoming cars	46000
9	A bus in a bus lane signals to pull away from a bus stop. Due to parked vehicles ahead in the bus lane, it pulls out into your lane forcing you to stop. *The clip occludes following one flash of the indicator from the bus.*	The bus pulls out into your lane	31000
10	An oncoming car cuts across your lane to park in a layby on your left. *The clip occludes just before the car starts to cross your lane, following one flash of its indicator.*	The car in front cuts across your lane	42000
11	A car behind you is visible in the rear-view mirror and right-side mirror. The car from behind overtakes your vehicle on a blind bend. The appearance of an oncoming car in the opposite lane forces the overtaking vehicle to pull back into your lane immediately in front of you. *The clip occludes when the overtaking car begins to emerge from your blind spot, and the oncoming car becomes obvious.*	The car behind overtakes on a blind bend	34000
12	While travelling at speed along a country road, a blind bend ahead reveals a queue of standing traffic, forcing you to slow and stop. *The clip occludes immediately after passing the blind bend, when the brake lights of the cars ahead are partially visible*	Queue of standing traffic	70000
13	While your car is slowing due to congestion ahead, a pedestrian looks over his shoulder before deciding to run in front of your vehicle forcing you to slow more abruptly than otherwise required. *The clip occludes at the moment when the pedestrian looks directly at you, as he steps forward.*	A pedestrian crosses the street from the left	35000
14	Driving through a school area, two pedestrians are visible on the right. You can see that the woman is looking at the car. *The clip occludes at the moment she gesticulates with her arm, thinking that you have given way for her and her child to cross.*	Pedestrians from the right cross the street after indicating to the camera car	42000
15	You are driving along a narrow road with parked cars on both sides, which are obscuring the view ahead. Suddenly an oncoming van appears ahead. *The clip occludes at the moment the white van is partially visible.*	A van appears suddenly behind parked cars	46000

6.4 Results

All participants were given a score out of 15 clips for the accuracy of their predictions. The scores were based on the interpretation of the typed free responses by two expert raters. Cohen's Kappa for each condition was acceptable (=>.8), with disagreements reconciled through discussion.

Participants' scores were converted to percentages and subjected to a 2 x 3 Analysis of Variance (ANOVA; experienced vs. inexperienced drivers x 3 levels of occlusion point). A main effect was found for driving experience with experienced drivers being more accurate when predicting hazards than inexperienced (54.9% vs.

45.6%; $F(1,55) = 7.13, MSe = 161.2, p < .01, \eta_p^2 = .11$). A main effect was also found across all levels of occlusion point, with the distal, intermediate and proximal conditions producing average accuracy rates of 28.6%, 50.7%, and.72.3%, respectively ($F(2,55) = 60.6, MSe = 161.2, p < .001, \eta_p^2 = .69$). Planned repeated contrasts were conducted comparing distal to intermediate, and intermediate to proximal occlusion points. Both were significant (*ps* < .001).

Despite a visible trend to suggest that the distal condition is the weakest discriminator of the driver groups (see Figure 6.2), no significant interaction was found between experienced group and occlusion point ($F(2,55) = .94, MSe = 151.2, p = .39, \eta_p^2 = .03$).

Ratings of hazardousness were also subjected to a 2 x 3 ANOVA but no differences were found across experience or occlusion points. The average hazard rating of experienced and inexperienced drivers for the distal clips was 3.46 and 3.40, respectively. For the intermediate clips, experienced drivers gave a rating of 3.76 while inexperienced drivers rated them at 3.72. Finally, for the proximal clips, the ratings for experience and inexperienced drivers were 3.25 and 4.04.

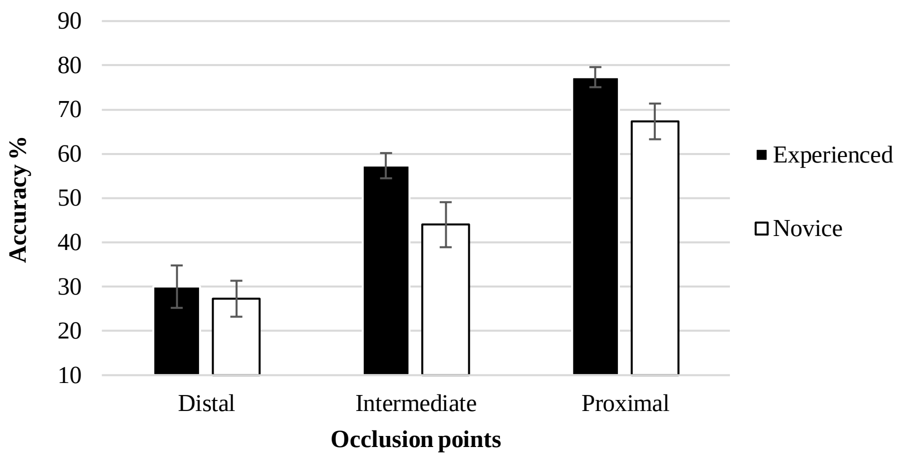


Figure 6.2: Percentage of correct responses across all three occlusion points and experienced groups (with standard error bars added)

6.5 Discussion

The current study has adopted an alternative methodology to test differences in performance between experienced and novice drivers. Since no differences were found between the experienced groups with the traditional hazard perception format (not only with the international samples but also with the UK sample), a new hazard *prediction* test was developed and trialled. Unlike the hazard perception test, the hazard prediction test successfully discriminated between the driver groups, with the experienced drivers outperforming the novices. This newly developed test has provided an insight into the act of hazard prediction, demonstrating that adding an occlusion to clips that were previously non-discriminative, is sufficient to evoke the hypothesised group differences. This result adds to the small, but growing, literature that has successfully employed the occlusion format (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009; Lehtonen et al., 2017; Lim et al., 2014; Ventsislavova et al., 2016). This is especially important given the more equivocal nature of the findings in the field when using more traditional response-time hazard perception tests (Horswill, 2017; Pradhan & Crundall, 2017).

The primary finding of the current experiment was that this new version of the hazard prediction test, using a completely different set of clips from those used by Crundall (2016), has successfully discriminated between experienced and inexperienced drivers. With appropriate occlusion points, the clips successfully elicited more accurate responses from experienced drivers compared to inexperienced drivers.

The occlusion-point factor ostensibly produced identical findings to those of Crundall (2016), with declining accuracy as the temporal separation between occlusion and hazard increases. Crundall (2016) failed to find an interaction, and the current interaction between experience and occlusion point also failed to reach significance. However, a closer look at the data is suggestive that the ability of clips to discriminate between the driver groups at the most distal level of occlusion is noticeably degraded. While the data do not support any strong conclusions, there is a clear suggestion that cutting the clip too early reduces the link between precursors and hazards to such an extent that all drivers approach a performance floor. In support of this, one should note that the overall accuracy to the distal condition in the current study (28.6%) is far below the 57% score in the comparable condition of Crundall (2016). This suggests that the distal occlusions in the current study are cut more severely in relation to the actual hazards. This cannot be due to absolute differences in distal occlusion points across the Crundall study (2016) and the current one (1200 ms vs. 1222 ms), so must therefore be due to differences in unfolding nature of the hazard precursors. It is likely that the lag between first evidence of a hazard, and the actual hazard beginning to develop, was greater in the Crundall (2016) clips, than in the current selection.

The information provided by the precursor is essential in terms of anticipating hazards prior to their onset. It has even been argued that being able to detect precursors is as important as spotting the actual hazard (Borowsky, Shinar & Oron-Gilad, 2010; Crundall et al., 2012), as drivers who make use of the cues from the environment will be better able to predict, mitigate and avoid collisions with hazards. More experienced drivers are better at anticipating precursors due to them facing similar situations previously (possibly via pattern matching against a store individual instances that one has been exposed to, Groeger, 2000). Furthermore, if drivers do not make use of precursors to anticipate a hazardous situation then the hazard becomes an abrupt onset to which a driver would *react* but not *respond* (with the latter action implying that the driver selects an action from a list of possible responses; Pradhan & Crundall, 2017). Younger participants (regardless of driving experience) have faster reactions to simple abrupt onsets (which means that abrupt onset hazards are likely poor discriminators between novice and experienced drivers), though fast reactions do not necessarily translate into safe behaviour (Crundall et al., 2012; Yeung & Wong, 2015).

In regard to hazardousness ratings, no differences were noted across either factor, mirroring the results of Jackson et al. (2009), who also failed to find significant differences regarding the hazardousness of the unfolding events. A possible explanation for this is that drivers are basing their ratings more on the nature of the scene that they *do* see (in the 30 seconds, or so, prior to occlusion), rather than the predicted event itself, which predominantly remains unseen. Drivers who do not know 'what happens next' following occlusion are unlikely to consider that the current clip is about to increase in danger. Even those who do correctly predict the hazard have no guarantee that they are correct, and a lack of confidence is likely to drag down the highest hazardousness ratings. Despite these considerations, it would still be expected that those drivers who correctly predict the event to give a higher hazard rating than those who had no idea what was about to happen. As the experienced drivers produced more significant predictions than the novice group, one would therefore expect (all things being equal) that experienced drivers would also rate the clips as more hazardous on average. The failure to find such a difference once again argues for contamination by criterion bias: experienced drivers are more likely to predict the upcoming event, but discount the danger it poses. Fortunately, as hypothesised, any such criterion bias did not influence the primary score of prediction accuracy. This result supports the use of this test-variant for avoiding the confounding effects of reference to internal hazard thresholds.

In conclusion, the current study has validated the prediction paradigm by demonstrating an experiential difference in predictive accuracy, using the same clips that failed to find group differences using the more traditional response time methodology. Furthermore, it has been demonstrated that the choice of occlusion point has limited influence upon the discriminability of the test, within reasonable boundaries. If set too early however, none of the participants are likely to score highly due to the removal of all precursors to the upcoming hazard. Logically, it also follows that setting the occlusion too late will equally remove any experiential differentiation, due to a ceiling effect.

Having demonstrated that the prediction test is more sensitive to driving experience in this context, the next question is whether this test format can be successfully exported to Spain and China. Given the results of this current study, and the suggestion in Chapter 5 that criterion bias may have confounded the international samples, the hazard prediction test offers a potentially more viable option for a culturally-agnostic test. The following chapter will return to Spain and China to assess this.

Chapter 7

CROSS-CULTURAL HAZARD PREDICTION

Experiment 6

Testing the Hazard Prediction test in three different countries: China, Spain and the UK

The material covered in this chapter has been adapted from previously published paper in Ventsislavova, P., Crundall, D. Baguley, T., Castro, C., Gugliotta, A. Garcia-Frnandez, P., Zhang, W., Ba, Y. & Li, Q. (2019). A comparison of hazard perception and hazard prediction tests across China, Spain and the UK. *Accident Analysis and Prevention, 122,* 268-286. https://doi.org/10.1016/j.aap.2018.10.010

7.1 Summary of previous findings

Experiment 5 was the first study of this thesis that yielded significant differences between experienced and novice drivers. All the studies prior to Experiment 5 failed to find these differences. In the previous studies, the traditional version of the hazard perception test was tested with different driving sample in terms of driving experience and within different cultural contexts. All of the studies failed to find the basic HP experiential effect. Therefore, a new alternative version of this test was created for Experiment 5. The hazard prediction test was first reported as an assessment that could differentiate between novice and experienced drivers by Jackson et al., (2009) though the underlying methodology dates back further (McKenna & Crick., 1997). Since then it has been successfully validated in several studies, consistently finding differences between experienced and novice drivers. It has been even argued that this test removes many of the methodological problems of the hazard perception test, such as the difficulty associated with setting hazard windows and the problem of criterion bias (Lim et al., 2014; Pradhan & Crundall, 2017). The aim of Experiment 5 was to assess whether a prediction test using the current stimuli would work (i.e. find a novice/experienced driver difference), and how the test should be precisely designed. The results demonstrated clear differences between the experienced and novice drivers, with experienced drivers predicting significantly more hazards following occlusion. There were also differences between the three types of occlusion points with the distal occlusion points decreasing prediction accuracy while the proximal occlusion points showed an improvement. Although the distal occlusion point was the hardest to predict, there was not a significant interaction between occlusion points and experience. Therefore, it was not possible to conclude that one general occlusion point was better than another. These results were identical to those of Crundall (2016), pointing towards a robust methodology, even though a completely different set of stimuli was used (including mirror information, which Crundall, 2016, did not use) and a different driving sample.

As this test showed clear differences between the experienced groups, the aim of the next study was to apply the test within the Chinese, Spanish and UK driving contexts and compare it to the results of the hazard perception test in Experiment 4. It was expected that this test would be better able to identify performance differences between the experienced groups within the different driving contexts.

7.2 Introduction

The act of hazard perception contains a number of sub-processes including searching for hazardous precursors, predicting which hazard is most likely to occur, monitoring the prioritised locations, spotting and processing the eventual hazard, and then responding in a timely and appropriate manner. Indeed, the whole process of avoiding a hazard on the road is poorly reflected within the term ‘hazard perception’ and recently Pradhan and Crundall (2017) have argued that ‘hazard avoidance’ is a more appropriate overall term. While ‘hazard perception’ is not a broad enough term to capture the whole hazard avoidance process (such as selection of the most appropriate behavioural response; see Ventsislavova et al., 2016), neither does it confine itself to a perceptual process. Evidence from Experiment 4 was noted, and from other studies, which suggests that post-perceptual processes, such as comparison of the demands of the unfolding hazard to one's own perceived skill, may influence the response. With such a nebulous definition of hazard perception, it is unsurprising to find that it is not completely clear on what the traditional HP test is measuring.

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

A second alternative to overcome the problems inherent in the traditional HP methodology is to pinpoint a more specific sub-process that can be more precisely measured. Pradhan and Crundall (2017) have defined these different sub-processes, one of which is the act of *hazard prediction*. This process is akin to Endsley's (1988a, 1995) third level of situation awareness: projection of future states and locations of objects on the basis of their current configuration and trajectories. The driver collects evidence from all potential hazard precursors and predicts whether any of them will come into conflict with her own vehicle. Should this process identify an imminent hazard, the driver prepares to act accordingly. This sub-process lies at the heart of all hazard avoidance, and is likely to be the key skill that traditional hazard perception tests are imperfectly measuring. Rather than letting the clips play all the way through, clips in the *hazard prediction test* are cut short, occluding as soon as the hazard begins to develop. Instead of asking participants to make a speeded response to the hazard, they are simply asked 'What happens next?', with their responses coded as correct or incorrect. This rests on the assumption that safer drivers know where to look for precursors to potential hazards, and can process, prioritise and monitor these precursors accordingly. This gives them the best possible chance of looking in the right place at the right time (i.e. looking at the precursor just as it begins to develop into a hazard before the screen is immediately occluded). Less safe drivers are less likely to be looking in the most appropriate locations and will therefore have a reduced chance of predicting the hazard (Crundall & Kroll, 2018).

7.2.1 Advantages of the hazard prediction test

This purer measure of hazard prediction skill offers several advantages over the traditional hazard perception methodology. First, it provides a measure of accuracy that is unavailable to traditional hazard perception tests (without some form of hazard localisation in the response, which may bring with it a new set of confounds). Secondly, it removes the need for temporal scoring windows which may penalise very good drivers who press slightly too soon. Thirdly, it removes the controversy of dealing with missing response time data. The traditional approach of recording the maximum possible RT in otherwise empty cells (McKenna et al., 2006) has been argued to distort results (see Parmet, Meir & Borowsky, 2014, who recommend the use of survival analysis). The hazard prediction test avoids this problem by dropping RTs as the main measure.

A fourth benefit is that it reduces the possibility that the test instructions are interpreted differently across the cultures. Terms like "hazard" and "hazardousness" are inherently prone to individual differences in interpretation (Wetton et al., 2011), and thus cultural differences are highly probable. Despite the best efforts in Experiment 4 (forward-backward translation, recruiting Chinese and Spanish experimenters to run the experiments in their respective countries), our participants may have had significantly different understanding of what constitutes a hazard. With the hazard prediction test however, this problem is removed by simply asking "What happens next?".

Finally, the hazard prediction test should remove criterion bias. There is no implicit or explicit motivation for participants to compare an unfolding hazard to their own self-perceived skill when responding. Instead, they simply report what happens next, regardless of how hazardous they believe the imminent event would be for them personally (though self-perceived hazardousness can still be captured after they have made the prediction if required). If the cultural sensitivity of the test used in Experiment 4 is, at least in part, due to the confounding of criterion bias with the traditional speeded hazard response, then a new test based just on this prediction element of the skill may be more robust (Jackson et al., 2009; Castro et al 2014; Lim et al., 2014; Crundall, 2016).

There have been several attempts so far to apply the hazard prediction test in different cultural driving contexts and they have all been successful. The hazard prediction test discriminated between experience and novice drivers within the Spanish driving context with different driving profiles (Castro et al., 2014; Ventsislavova et al., 2016), within the UK driving context where it was first validated (Crundall, 2016; Jackson et al., 2009), within the Finnish driving context with a sample of infant cyclists (Lehtonen et al., 2017) and within the Malaysian driving context (Lim et al., 2014). This latter study was in direct contrast to Lim et al., previous attempts (2013) to find experiential differences with a more traditional hazard perception methodology.

7.2.2 Overview of the study

The hazard prediction test for this study was created using the same clips employed in Experiment 4. Following occlusion, participants typed their responses to what they believed would happen next. A new cohort of experienced and inexperienced drivers was recruited across the three countries for this second experiment. It was predicted that the prediction test would be more successful than the hazard perception test in discriminating between the driver groups, and that the test would demonstrate fewer cultural sensitivities.

7.3 Method

7.3.1 Participants

A hundred and fifty-three participants took part in this study. The sample was composed of 50 Chinese, 52 Spanish and 51 UK drivers. One participant was later excluded (from the UK sample) due to difficulties categorising the individual as experienced or inexperienced. All of the participants held a full or a learner-driver licence from their country. Participants were split into two sub-groups of experienced and inexperienced drivers following the method used in Experiment 5. In China, 26 experienced drivers were recruited (mean age of 25.3, an average of five years of post-licensure experience, and a mean annual mileage of 5474 miles) and 24 inexperienced drivers (mean age of 22.7, an average of one year of post-licensure experience, and a mean mileage of 33.7 miles). In Spain, 27 experienced drivers were recruited (mean age of 40.9, an average of 21 years of post-licensure experience, and a mean mileage of 20183 miles) and 25 inexperienced Spanish drivers (mean age of 20.2, an average of 1 year of post-license mean experience and mean mileage of 28.9 miles). In the UK 23 experienced UK drivers were recruited (mean age of 24.4, an average of seven years of post-license mean experience and mean annual mileage of 5587 miles), along with 27 inexperienced UK drivers (mean age of 19.4, an average of one year of post-license experience, and a mean annual mileage of 266.7). Across all countries, the mean age of experienced drivers was 30.2 years, with an average of 11 years of post-licensure experience, and they had driven an average of 10415 miles in the previous year, while inexperienced drivers had a mean age of 20.8, with an average of one year of post-licensure experience, and had driven an average of 109.6 miles (see Table 7.1).

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

Table 7.1: Mean demographic values for participants

Demographics	Chinese Participants		Spanish Participants		UK Participants	
	Novice	Exp'd	Novice	Exp'd	Novice	Exp'd
Total N (female N)	24 (9)	26 (4)	25 (21)	27 (4)	27 (18)	23 (23)
Age	22.7	25.3	20.2	40.9	19.4	24.4
Post-licence Experience (years)	1	5	1	21	1	7
Annual Mileage	33.7	5474	28.9	20183	266.7	5587

7.3.2 Materials and apparatus

The apparatus and stimuli for this experiment were the same as those used in Experiment 5, though the video clips were edited to stop immediately prior to the appearance of the hazard for the current experiment (immediately following hazard onset), with the clip occluded by a black screen. The edited clip always gave enough information for participants to deduce what would happen next in the driving scene providing they were looking in the appropriate location just before occlusion (Jackson et al., 2009). At the end of each clip a black screen was displayed. The duration of the clips varied between 12 s. and 58 s.

The study was subjected to a forward-backward translation procedure similar to the one conducted in Experiment 5. UK instructions were subjected to a Chinese and Spanish forward-backward translation (following the guidelines of International Test Commission; ITC, 2010). During the translation process an emphasis was made on the importance that participants clearly understand that they are required only to predict how the driving situation is going to develop independently of whether they consider it hazardous or not. As participants were required to provide typed responses, each response was translated individually and independently by the same bilingual experts in Experiment 5. Then, translated responses were compared in order to ensure that all content represents the original responses at all accuracy. The translation into Chinese and Spanish was performed by the same team as in Experiment 5 consisting of three bilingual experts with a high level of expertise in Chinese and Spanish culture, traffic regulations and driving habits.

As an example, consider clip 1 from the Chinese block (see Table 7.2). In this clip (as used in Experiment 5) a pedestrian looks to cross the road from the right but is then obscured by a turning vehicle. When the vehicle has finished the manoeuvre, the pedestrian is already crossing the road in front of you. For the current hazard prediction test, this clip was edited to end in the middle of the obscuring vehicle's manoeuvre, at a point where part of the hazardous pedestrian emerging in the road can be seen. An experienced driver should notice the pedestrian before the vehicle turns, and therefore should monitor the trailing edge of the obscuring vehicle to assess whether the pedestrian has indeed entered the road. The briefest glimpse of the re-emerged pedestrian is only likely to be spotted if the driver is aware of the unfolding hazard and is actively seeking the pedestrian.

Table 7.2: A description of the occlusion points of the a priori hazards selected within each clip

Clip Number	Occlusion points of the a priori hazards (*see, Chapter 5 for description of the full hazard*)
	CHINESE CLIPS
1	The clip occludes just as the pedestrian starts to become visible as the obscuring car moves past.
2	The clip occludes as the pedestrian first becomes visible stepping out from between the parked cars.
3	The clip occludes as the front wheel of the bicycle enters the view.
4	The clip occludes as the cyclist makes a change in direction to cut across your path.
5	The clip occludes following one flash of the indicator from the manoeuvring car.
6	The clip occludes when the car is no longer visible in the left mirror, but flash of it is visible in the left window.
7	The clip occludes when the car is no longer visible in the right mirror, but it quickly appears next to the right window of your car.
8	The clip occludes when the car is no longer visible in the right mirror.
9	The clip occludes at the moment in which the lorry is about to enter into the main road.
10	The clip occludes when the car is no longer visible in the right mirror.
	SPANISH CLIPS
1	. The clip occludes when the front part of the motor is visible.
2	The clip occludes when the pedestrians stops next to the car.
3	The clip occludes following one flash of the indicator of the scooter.
4	The clip occludes at the moment when the pedestrian enters the road.
5	The clip occludes following one flash of the indicator from the car.
6	The clip occludes when the van is approaching from the right and almost enters your lane.
7	The clip occludes just as the pedestrian first become visible.
8	The clip occludes when the bus turns to enter the road and following a flash of the indicator of the bus.
9	The clip occludes when the motorcycle is no longer visible in the right mirror, but part of it is visible at the right window.
10	The clip occludes when the pedestrian is approaching the zebra crossing.
	UK CLIPS
1	The clip occludes following initial illumination of the reversing light.

2	The clip occludes when the pedestrian turn his head to look at your car.
3	The clip occludes when the pedestrian approaches the road in order to cross.
4	The clip occludes when the car is no longer visible at the left mirror, but a flash of it is visible on the left window.
5	The clip occludes when the push chair is partially visible among the pedestrians.
6	The clip occludes following one flash of the indicator from the bus.
7	The clip occludes just before the car starts to cross your lane and following one flash of the indicator of the car.
8	The clip occludes when the car is no longer visible at the rear-view mirror and the right mirror.
9	The clip occludes immediately after passing the blind bend, when the brake lights of the cars ahead are partially visible.
10	. The clip occludes at the moment when the pedestrian is visible at the left pillar and is looking at your car.

7.3.3 Design and procedure

The design of the study was identical to that of Experiment 5, except for the dependent variable. Instead of a response time measure to the hazard, the screen occluded immediately prior to the hazard fully developing, and participants were asked to type what they thought happened next into a text entry box on the screen.

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

Typed responses were later coded with one point given for each correct answer (ideally specifying what and where the object of interest was, and how the event would unfold) and zero points for an incorrect answer (the coding procedure was identical as the one in Experiment 6).

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

7.4 Results

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

7.4.1 Accuracy

To test whether there were differences in the accuracy of hazard prediction performance across the factors a 2 x 2 x 3 ANOVA was conducted. The between-groups factors were the experience level of drivers (experience vs. inexperienced drivers) and their nationality (Chinese vs. Spanish vs. UK). The within group factor was the origin of the clip (China vs. Spain vs. UK).

As predicted, a significant main effect was found for driver experience when predicting the hazardous situation (F(1, 146)=7.62, MSe = 263.70, p = 0.007, η_p^2=0.050), with experienced drivers outperforming novices (52.1% vs. 45.0%, respectively). This effect did not interact with participant nationality or clip origin.

A significant difference was also found for clip origin (F(2, 292)=10.15, MSe = 222.66, p < 0.001, η_p^2 = 0.065). Repeated contrasts found the Spanish clips to evoke significantly poorer performance than both Chinese clips (43.5% Vs. 51.5%, F(1, 146)=19.90, MSe = 428.96, p < 0.001, η_p^2= 0.120) and UK clips (43.5% Vs. 49.7%, F(1, 146)=9.91, MSe = 434.30, p < 0.005, η_p^2=0.064).

Though no main effect of driver nationality was found, there was an interaction between clip origin and nationality (F(4,292)=4.48, MSe = 222.66 p < 0.005, η_p^2 = 0.058). Repeated contrasts identified the interaction to lie in the comparison of Chinese clips to Spanish clips across participant groups (F(2,146) = 5.48, MSe = 428.96, p = 0.005, η_p^2=0.070). From Figure 7.1 it can be seen that the UK and Spanish clips evoke similar patterns of behaviour across all participant groups, however the Chinese clips disrupt this pattern with Chinese drivers performing particularly well.

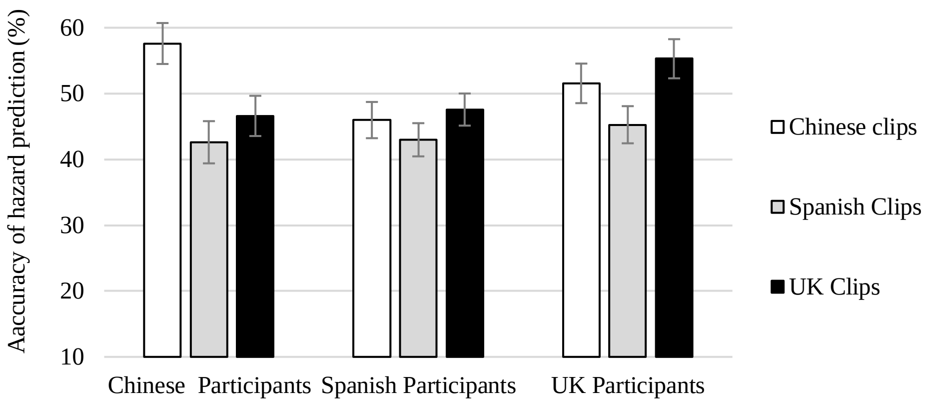


Figure 7.1: The mean percentage of accuracy of predicting the hazards of all three driver groups across the three clips sets recorded in each country (with standard error bars added)

7.4.2 Hazardousness ratings for hazard prediction

As with the hazard perception study, participants were asked to provide a hazardousness rating on a scale of 1 to 7 (where higher numbers reflect greater levels of perceived hazardousness). These ratings were averaged across clips for each participant according to country of origin and were subjected to a 2 x 3 x 3 ANOVA (across experience, nationality and clip origin).

There was a main effect of clip origin ($F(2,294) = 55.56$, $MSe = 0.313$, $p < 0.001$, $\eta_p^2 = 0.274$). Repeated planned contrasts comparing the three country tests demonstrated that Chinese clips were rated as more hazardous than Spanish Clips (4.26 vs. 3.73; $F(1, 147) = 69.3$, $MSe = 0.607$, $p < 0.001$ $\eta_p^2 = 0.321$) with no difference between the Spanish clips and the UK clips (3.73 vs. 3.63; $F(1, 147) = 2.57$, $MSe = 0.655$, $p = 0.11$ $\eta_p^2 = 0.017$).

Nationality also produced a significant main effect ($F(2, 147) = 11.75$, $MSe = 0.969$, $p < 0.001$, $\eta_p^2 = 0.138$), with Spanish drivers giving the highest ratings (4.24) followed by UK drivers (4.02), with Chinese drivers giving the lowest hazard ratings (3.33). Tukey tests revealed Chinese drivers to give lower ratings than both Spanish and UK drivers (ps < 0.01), with no difference between Spanish and UK drivers ($p=0.48$).

In regard to driver experience, though experienced drivers were not found to give significantly different ratings to those provided by novices overall (3.86 vs. 3.88), driver experience did interact with clip origin ($F(2, 294) = 4.38$, $MSe = 0.313$, $p < 0.05$, $\eta_p^2 = 0.29$). Planned comparisons revealed the interaction to lie in the comparison of Chinese and Spanish clips ($F(1, 147) = 8.48$, $MSe = 0.607$, $p < 0.05$, $\eta_p^2 = 0.055$). As can be seen in Figure 7.2, both driver groups rated Chinese clips as more dangerous than Spanish clips, but this effect is more pronounced in the experienced driver group.

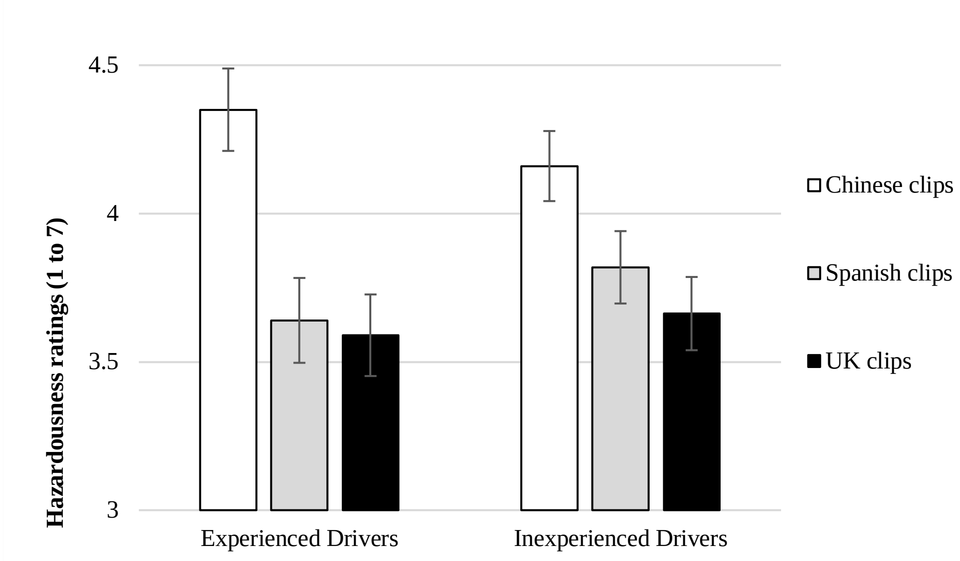


Figure 7.2: Hazardousness ratings for the hazard prediction test for the experience groups across clip origin (with standard error bars).

7.5 Discussion

Unlike the hazard perception test of Experiment 4, the hazard prediction test successfully differentiated between experienced and novice drivers, with the experienced drivers outperforming the inexperienced across all nationalities. These results are consistent with the limited previous research, demonstrating that the prediction test is a more robust discriminator of driver experience than the traditional hazard perception test (Lim et al., 2014, Castro et al., 2014, Crundall, 2016). The superiority of the hazard prediction test is all the more convincing in that the group differences were found using the same clips as the unsuccessful hazard perception test in Experiment 4. In addition, no interaction was found between experience and participant nationality demonstrating that the prediction test is less sensitive to cultural differences than the hazard perception test.

There were, however, still differences between the responses in regard to clip origin as one might expect. In contrast with Experiment 4, Spanish hazards appeared to be the most difficult to predict when compared to both Chinese and UK hazards. This was not reflected in the hazardousness ratings attributed by drivers to the clips, where

Chinese clips were once again rated as most hazardous. This suggests that the Likert scores were more likely reflecting general visual clutter, complexity and congestion, rather than the *a priori* hazard in particular. This is understandable in the current experiment as drivers never saw the full *a priori* hazard. Instead, the participants presumably referenced other potential hazards that they had seen in the clip in order to provide a hazard rating. As Chinese clips evoked the most extra-hazard responses in Experiment 4, it is safe to conclude that these clips contain more potential hazard precursors, and that this fact is probably producing the hazardousness ratings in this experiment.

Certainly, the evidence from this experiment does not suggest that the prediction test remains uninfluenced by the location in which they are filmed, but the fact that different environments produce different accuracy rates is not surprising. Neither it is surprising that some drivers perform better when viewing clips filmed in their own country. This does not however detract from the claim that the prediction test is a more culturally-agnostic form of assessment than the hazard perception test. The crucial point is that this test discriminates between experienced and inexperienced drivers regardless of which country they come from.

7.6 Comparison of the two tests

It is possible to directly compare the performance on the two tests, using the prediction accuracy from experiment 6 and the percentage of hazards that received a speeded response in experiment 4 (though note that we cannot claim that all responses that fell in the scoring window in the hazard perception test were referencing the actual hazard – this is one of the problems with the traditional HP methodology). In the analysis reported below we only focus on the main effect of *test type* (whether accuracy scores differ across the hazard *perception* and hazard *prediction* tests), and any emerging interactions with test type.

Accuracy rates for the two tests were compared with a 2 x 2 x 3 x 3 mixed ANOVA with the factors of test type (Perception vs. Prediction), participant experience (inexperienced vs. experienced), participant nationality (Chinese, Spanish or from the UK), and clip origin (China, Spain, UK). The results showed that there was a main effect for the type of test ($F(1,288)=182.07$, $MSe = 230.22$, $p < 0.001$, $\eta_p^2=0.387$). Participants scored higher on average for the hazard perception test compared to the hazard prediction test (72.1% vs. 48.6%; $p < 0.001$). A significant interaction was also found for test-type and experience ($F(1, 288)=5.36$, $MSe = 230.22$, $p < 0.05$, $\eta_p^2=0.018$). Despite the prediction test appearing more difficult than the perception test, it is clear that the benefit of experience only holds for hazard prediction (52.1% vs. 45.0%) rather than hazard perception (see Figure 7.3).

A significant interaction was found for test type and nationality ($F(2, 288)=4.78$, $MSe = 230.22$, $p < 0.01$, $\eta_p^2=0.032$). As can be seen from Figure 7.4 the variation in performance across the nationalities was significantly greater in the hazard perception test (reflected in the main effect of nationality found in Experiment 4), than in the hazard prediction test (with no significant main effort of nationality in Experiment 6).

Finally, there was a significant interaction between clip origin and test type ($F(2, 576)=12.73$, $MSe = 196.20$, $p < 0.001$, $\eta_p^2=0.042$). The within-subject contrast revealed a difference between the Chinese and Spanish clips ($F(1, 288)=23.39$, $MSe = 399.17$, $p < 0.001$, $\eta_p^2=0.075$). In the hazard perception test the Chinese hazards were the hardest to detect (68.7%) and UK hazards were the easiest (75.1%), however for the prediction test, Spanish hazards appeared to be the hardest to predict (44.3%) and Chinese the easiest (51.8%) (see Figure 7.5).

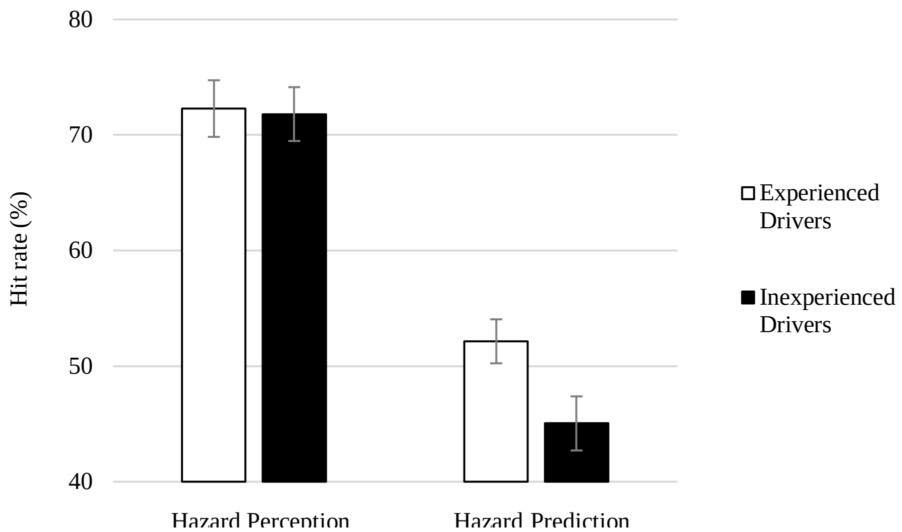


Figure 7.3: The mean percentage of the hazards correctly identified/predicted for both hazard perception and hazard prediction tests across driving experience (with standard error bars)

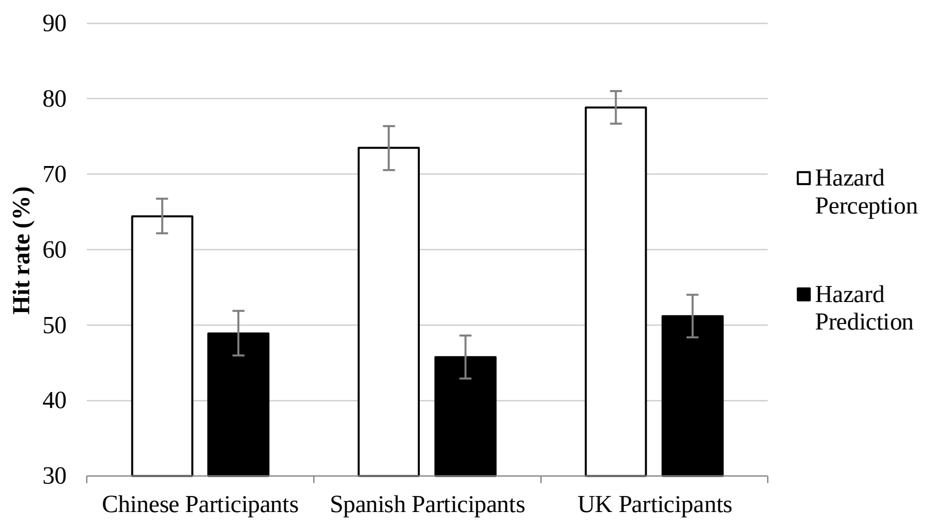


Figure 7.4: The mean percentage of hit rate for the hazard perception and the hazard prediction test across driver's nationality (with standard error bars)

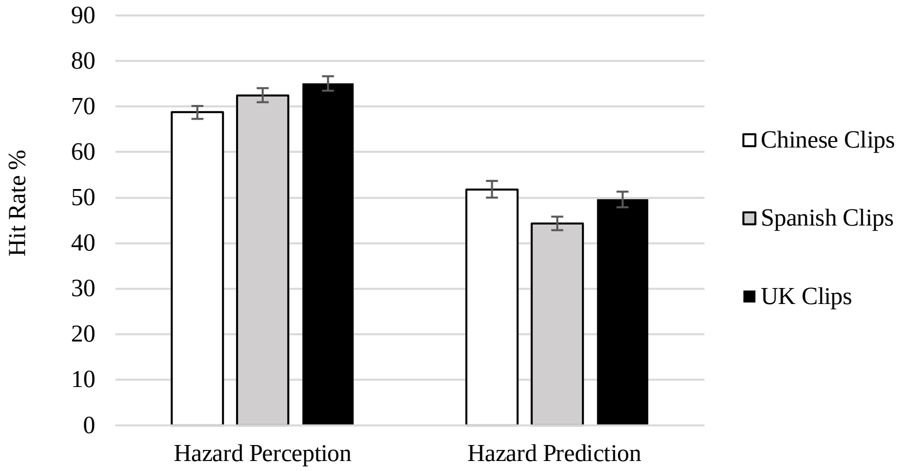


Figure 7.5: The mean percentage of hit rate for the hazard perception and the hazard prediction test across clip type (with standard error bars)

7.7 General discussion

The aim of this study was to assess whether the hazard prediction test is suitable for export to different driving cultures, as the typical hazard perception test methodology may be culturally sensitive, and therefore less suitable for adoption in other counties. Many research groups around the world have investigated hazard perception in their own countries, though vastly differing methodologies making it difficult to compare the validity of the tests across different regions. In this chapter, it was compared both the hazard perception and the hazard prediction test and only the prediction test showed differences in performance between experienced and novice drivers.

Crucially, the clips edited for the hazard prediction test were the same as those used in the hazard perception test in Experiment 4, allowing a direct comparison of the two tests. This is the first time that the hazard perception and hazard prediction tests have been directly compared in a single analysis[6], though this was complicated by the

[6] Since this study was completed, NTU researchers have compared hazard perception and hazard prediction test variants using video clips filmed from fire appliances on blue-light training runs. Once again, the prediction test was found to be the better discriminator of driver groups (Crundall & Kroll, 2018).

fact that the two tests record very different primary measures: response times and percentage accuracy, respectively. However, as the hazard perception test required response times to fall within a temporal scoring window around the appearance of the hazard, the presence or absence of a response allowed the calculation of an accuracy score that could be compared to the hazard prediction test.

While participants found the hazard prediction test much harder than the hazard perception test, the superiority of the prediction test in differentiating between driver groups on the basis of experience was clearly demonstrated. Most importantly, the main effect of experience, with experienced drivers outperforming novices, was present across the participants as a whole, and did not interact with nationality. As the prediction test was designed to remove criterion bias, it was comforting to note that the cultural differences that arose in Experiment 4, which were interpreted as potentially arising from hazard threshold differences, were ameliorated to a large extent in this study.

7.7.1 Do different countries produce different hazards?

In both Experiments 4 and 6, differences were noted in responses for ratings to the clips on the basis of their origin. This is unsurprising, as Beijing, Granada and Nottingham, differ on a great many characteristics. The higher population, congestion and collision rates in China suggest that this should provide the most hazardous stimuli. While the clips were filmed with the same protocol, there were inevitable differences in the visual clutter and frequency of hazardous precursors across the countries. From an experimental design point of view, this was not a great concern. As every participant saw clips from all three countries, we could thus analyse the relative differences between the responses of our participants across the three nationality groups.

The effect of clip origin on ratings closely mirrored the behavioural findings in Experiment 4. Chinese clips were considered most dangerous by the Spanish drivers, but Chinese and UK drivers considered the Chinese and Spanish clips to be equally hazardous. The number of extra responses per minute followed this pattern, with Spanish people responding most frequently when watching the Chinese clips, while Chinese and UK drivers responded equally to Chinese and Spanish clips. The *a priori* hazards in the Chinese clips were more likely to be missed by all drivers. Presumably, this was due to a greater number of precursors resulting in the possibility that participants were looking in the wrong place at the time of hazard onset.

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

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

7.7.2 The limitations of hazard prediction tests

The current study suggested that the hazard prediction test is a better discriminator of driver safety than the hazard perception test, however it is not without its limitations. For instance, it may be argued that the average experienced-driver score of 52.1% accuracy is not very high. We counter, however, that it is the difference between the two groups that is more important, rather than an absolute score. While the difference between experienced and inexperienced drivers was significant (albeit it with a small-to-medium effect size), this could be improved with iteration of the stimuli sets, as would occur in the development of a formal test.

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

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

One final point to note is that the free-response format of the current hazard prediction test does not lend itself to widespread automated testing, due to the lack of immediate feedback, and the possibility of coding errors and subjectivity influencing the scoring.

To counter this criticism, there have already been several attempts to create a multiple-choice hazard prediction test (Lim et al., 2014; Gugliotta et al., 2017; Ventsislavova et al., 2016) which have all been successful in validating this format type, and in finding differences between experienced and novice drivers. This format type is especially beneficial as it offers pragmatic advantages such as objective and quick marking, quick administration and it is time and cost effective.

In conclusion, the hazard prediction test provides a purer, culturally agnostic variant of the traditional hazard perception test, and offers a blueprint for future test development at a global level.

Chapter 8

REFINING THE HAZARD PREDICTION TEST

Experiment 7

Validation of a multiple-choice hazard prediction test

The material covered in this chapter has been adapted from previously published paper in Ventsislavova, P., & Crundall, D. (2018). The hazard prediction test: A comparison of free-response and multiple-choice formats. *Safety Science, 109*, 246-255. https://doi.org/10.1016/j.ssci.2018.06.004 (see Appendix D)

8.1 Summary of previous findings

After several unsuccessful attempts to find differences between experienced and novice drivers with the traditional hazard perception methodology, it was decided to try to apply an alternative version of the HP test. This new hazard prediction test managed to differentiate between different driving groups, with the experienced drivers predicting significantly more hazards than the novices in three different countries (China, Spain and the UK). Furthermore, both versions of the test were compared across China, Spain and the UK and although the prediction test appeared to be harder, it clearly showed superiority in differentiating between the experienced groups. These differences were based on the experience level of participants and not on their nationality, and the effect was not confounded by threshold bias (as with the traditional hazard perception test). This finding suggests that the prediction test is more suitable for export to other countries.

However, a pragmatic limitation of the hazard prediction test is its current free-response format. If participants are required to type their answers, coding error could influence the scoring. Also, responses will require translation which in itself is a lengthy process that needs to be repeated after each experiment. Finally, this format does not allow automated testing or immediate feedback which can be a disadvantage in terms of widespread international testing.

As a result, for the following study a multiple-choice version was created. This alternative format allows quick and objective testing and does not require continuous translation. The decision to move to this format was also motivated by the previous attempts to validate a multiple-choice hazard prediction test, all of which have been successful in several countries (Crundall & Kroll, 2018; Lim et al., 2014; Gugliotta et al., 2017). However, in the present chapter the multiple-choice hazard prediction test will only be applied to a UK sample in order to find out whether it would be able to discriminate between driving groups in the UK. If successful, this new refined format will be exported to a completely new country.

8.2 Introduction

The two previous studies suggested that the hazard prediction test can provide a robust and simpler alternative to the more traditional hazard perception test. However, one of the problems with the version of the test used by many researchers (Castro et al., 2014, 2016; Crundall 2016; Jackson et al., 2009) is that participants give free-response answers which must be hand coded. This introduces the potential for rater error, and renders the test impractical for use on a wide scale. An alternative is to provide the participants with multiple options to choose from following occlusion, instead of inviting a verbal or typed response. This approach simplifies the test further and allows for automatic and unambiguous coding. This variant of the hazard prediction test was first employed by Lim et al. (2014), and was followed by Ventsislavova et al. (2016), but there has never been a formal approach to designing a multiple-choice version of the prediction test, nor a direct comparison between a multiple-choice test variant, and hand-coded free responses. Despite the evidence to suggest that multiple-choice prediction tests can find differences between drivers of differing experience, we do not know whether the experiential effect is enhanced or diluted through the use of multiple options over free responses.

This current study aimed to address both these issues. First, a formal procedure was adopted for creating a multiple-choice prediction test, which was then tested across driver groups of varying experience in search of validity. Secondly, a comparison was made between the MCQ test and the previous free-response variant in Chapter 6, in order to assess whether validity is improved with the simpler response format. In addition, while the hazard perception test has been previously correlated to driving experience (Wetton, Hill & Horswill, 2011) and age (Horswill et al., 2010), the hazard prediction test has not been directly correlated with any of these variables. The prediction test successfully discriminates between experienced and novice drivers, however none of its formats (free-response or multiple-choice) have been previously correlated to age or driving experience. Therefore, it will be assessed whether both formats correlate positively with years of driving, age and annual mileage.

Lim et al. (2014) first used a multiple-choice hazard prediction test with Malaysian and UK drivers. Their results showed that experienced drivers were able to predict more hazards in comparison to the novices (though only for the Malaysian clips). While this result was in contrast to a previously unsuccessful study using response-time hazard perception clips (Lim et al., 2013), they did not directly compare the two datasets across the studies. Furthermore, the limitation of the finding only for the Malaysian clips raises potential concern over the choice of multiple-choice options in the UK clips. More recently, Ventsislavova et al. (2016) created a multiple-choice Spanish prediction test in order to test the differences between experienced and novice drivers, asking both “What was the hazard” and “What happens next?”. They found that experienced drivers were more accurate in identifying and predicting the hazard than novices.

Two recent studies have also compared multiple-choice prediction tests directly to traditional response-time hazard perception tests (Crundall & Kroll, 2018; Malone et al., 2016). Malone et al. (2016) described their two test variants as differing on ecological validity, with the response-time test (similar to the national UK test) representing high ecological validity, while the multiple-choice version represented low ecological validity. Even though they expected that the high ecological validity test would discriminate better between experienced groups, the multiple-choice version of the hazard perception test also yielded a significant difference between experienced and novice drivers. It should be noted however that Malone et al. used hazards that fully materialised in those trials where participants were given multiple-choice options. As participants saw the whole hazard, this test did not isolate the act of prediction, and is therefore less applicable to the current discussion. Crundall and Kroll (2018) compared a response-time HP test and a multiple-choice prediction test across groups of fire appliance drivers (with clips filmed from fire appliances on blue-light training runs). They found the multiple-choice test to be more sensitive to group differences. Multiple-choice formats have also been employed in the training of hazard perception (Cockerton & Isler, 2003; Isler & Cockerton, 2003, Petzoldt, Weiss, Krems, & Bannert, 2013). Petzoldt et al. (2013) compared a paper-based intervention with computer-based training using a multiple-choice format, with the latter being more beneficial for participants to spot earlier critical cues and scan relevant areas in the visual field.

One concern that has been raised regarding the development of multiple-choice tests relates to the plausibility of the distracters (Andrich & Marais, 2014). Poor or implausible distracter options can be easily rejected allowing the participant to choose the correct response by default. For this reason, the guidelines of Worthen et al, (1999) and Haladyna, Downing and Rodriguez (2002) were followed, creating the distracter items using the incorrect free-response answers that were provided by the participants in Experiments 5 and 6.

In order to test this new format, the sample was divided into four groups: learners, novices, moderately experienced drivers and highly experienced drivers. It was predicted that there would be clear differences in prediction accuracy across the driving groups, with the enhanced range of drivers providing more opportunities for differentiation between the groups.

Finally, a matched sub-group of participants from Experiment 5 and 7 were compared in terms of their prediction accuracy with the only difference being the response mode: free-response or selection of an answer from four options. It was hypothesised that both test variants would discriminate between drivers to a certain degree, but no prediction was made about which test would be the better discriminator in the final analysis. No prediction was made about which format will correlate with years of driving, age and annual mileage either.

8.3 Method

8.3.1 Participants

Fifty-one participants were split into four different experienced groups: 12 learner drivers (L), ten novice drivers (N), 15 drivers that have driven for less than six years (E<6y) and 14 drivers that have driven more than six years (E>6y) (see Table 8.1 for demographics). All participants had normal or corrected-to-normal vision and were University students and staff.

Table 8.1: Mean age, mileage and driving experience for the different driving groups

Participants	Mean Age	Mean driving experience since test (in years)	Mean mileage in the previous 12 months
Learner drivers	18.3	0	9
Novice drivers	18.4	11 months	555
E<6 years	21.4	3 years	4933.3
E>6 years	35.9	16 years	7857.1

8.3.2 Design, materials and procedure

The video clips used for this study were the same as in Study 4. However, this time only the clips with the proximal occlusion point were used in order to assure that participants had the best possible chance to predict the hazardous situations. In Chapter 6, the results suggested that performance was best with the proximal occlusion point, although there was no interaction between the experienced groups and occlusion points. Furthermore, for this study, clips were edited to include a multiple-choice question and participants were no longer required to type their responses. The multiple-choice options were created following the guidelines provided by Haladyna et al. (2002). They have suggested more than 30 guidelines on how to create a valid multiple-choice item covering concerns such as content, formatting, style and writing. All guidelines were followed and special emphasis was made on the following five:

1. All options were written to be of similar length. Consistent length ensures that the correct answer cannot be identified from higher levels of description in the option.
2. Future tense was used to design the options, as the question was “What will happen next?”.
3. The location of the right answer (i.e. position 1, 2, 3 or 4) was randomly assigned for each question.
4. Plausible distracters were created using previously collected common errors of participants.
5. The *frequency* method was used to create the distracters (i.e. the most frequently incorrect answers are used for distracter options)

Participants’ incorrect typed answers to the UK video clips in Experiments 5 and 6 were used to create the distracters. In order to identify the most frequent incorrect answers, the *frequency* method was used. This method is an empirical approach and yields items with highly plausible distracters (Owens, Hanna and Coppedge, 1970). Once the answers were grouped, the most common incorrect responses were selected as distracters for each video, thus assuring the plausibility of all options. It should be noted that each distracter was compared to the corresponding video to verify that it indeed represented the hazardous scene.

Four options were created (one correct and three incorrect) for each video (15 in total). Both correct answers and distracter options were written in short, simple sentences and future tense.

This first set of options was piloted across 41 drivers with different experience. To this end, a Google-forms questionnaire was created. The link to the online forms was distributed via social media. Participants were not shown the video clips. Instead, a very general sentence was given as introduction to the scenario, and they were then asked to simply select the answer from the four options that they think correctly predicted what happens next (e.g. “You are driving along a busy urban street with pedestrians and parked vehicles on either side of the road. What happens next?”). If people could guess the correct answer without seeing the clip (i.e. without any clues to the answer, beyond the wording of the answer itself), then the real answer might be too obvious or the distracters too implausible.

Participants’ answers for each scenario were subjected to a 1 x 4 chi-square analysis to identify whether one of the four options evoked a higher response rate than the others. If the Chi Square analysis for a particular scenario was significant, and the correct answer was the source of the effect, this was termed ‘correct bias’. If a distracter option had caused the effect however this was termed ‘incorrect bias’. Where ‘correct bias’ was observed, the correct option for that scenario was changed to make it less obvious. According to the results, four of the items showed no bias, eight of the items showed incorrect bias (a single distracter was chosen more frequently, which was not considered to be an impediment to the test) and three of the items showed correct bias. These three items were revised. The final selection of multiple-choice options were edited into the final frames of the hazard prediction scenarios.

A 1 x 4 between-subjects design was employed, where the independent variable was the level of experience of the participants and the dependent variable was the percentage accuracy of hazards predicted. The procedure was identical to that used in Experiment 6 with the only difference being that this time participants were asked to predict the driving situation by choosing one of four options after each of 15 clips (see Figure 8.1). Participants were aware that only one option would be correct and, if they had been looking in the correct place at the time of occlusion, then they would have seen very brief confirmation that their prediction was correct. Following their selection of an option, participants were asked to rate how hazardous the predicted situation would have been for them, by selecting a point on a 1 to 7 Likert Scale (where 1 is ‘not at all hazardous’ and 7 is ‘extremely hazardous’). All clips appeared in a randomised order. Prior to the actual experiment, all participants saw a practice trial. The experiment took an average of 20 minutes to complete.

All apparatus remained the same as that used in Experiment 6.

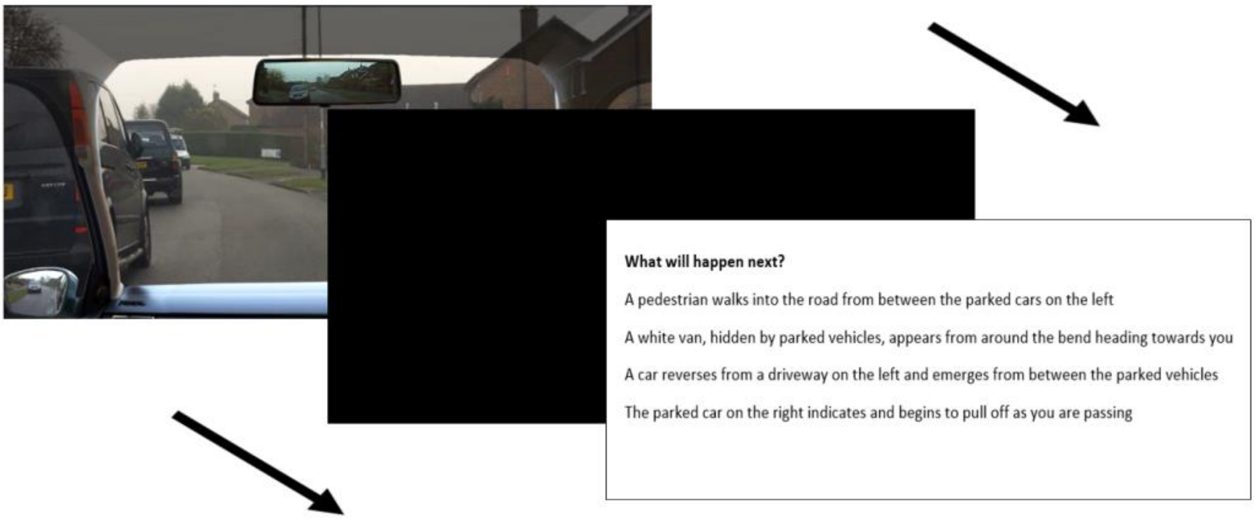


Figure 8.1: Three panels represent the end of a clip: the first panel is the final frame (where the impending hazard is an oncoming white van, hidden by parked vehicles). This is followed by an occlusion panel, before four multiple-choice options are displayed on screen.

8.4 Results

Accuracy was analysed using a 1 x 4 ANOVA, where the between-group factor was driver experience. A main effect was found (see Figure 8.2), with performance improving in line with increasing experience across the four groups (71.7%, 79.3%, 88.9% and 89.5% for learners, novices, moderately experienced and highly experienced drivers, respectively; $F(3,47) = 6.9, MSe = 134.2, p < .001, \eta_p^2 = .31$). Planned contrasts were undertaken to unpack this main effect. The first contrast pooled the learners and novices and compared them to the pooled experienced driver groups (mirroring the analysis in Experiment 5 which just compared inexperienced drivers to experienced drivers). The difference was found to be significant with the experienced drivers out-performing the learner/novice group ($t(47) = 4.2, p < 0.001$). Repeated contrasts (learners vs. novices, novices vs. moderately-experienced drivers, and moderately-experienced vs. highly-experienced drivers) were also conducted to identify whether the test was sensitive to more subtle changes in experience, but the only significant comparison was found between novices and moderately-experienced drivers ($t(47) = 2.0$, $p < 0.05$). This suggests a clear experiential boundary in the ability of the test to discriminate between groups at the transition between novice and moderately-experienced drivers.

Ratings were also analysed and a main effect was found for driving experience across the four groups ($F(3,47) = 2.9$, $MSe = .85$, $p < .05$, $\eta_p^2 = .16$). While learners gave the highest risk ratings for the clips (4.5) and the highly-experienced group gave the lowest ratings (3.7), the experiential effect was not sufficient for any of the planned contrasts (L/N vs. all experienced drivers; L vs. N; N vs. E<6yr; E<6yr vs. E>6yrs) to reach significance (see Figure 8.2).

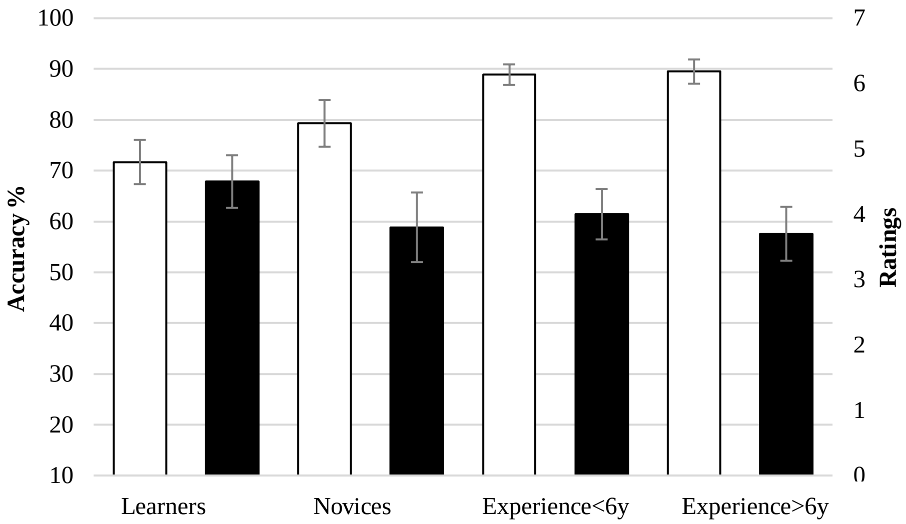


Figure 8.2: Percentage accuracy (white bars) and ratings (from 1-7; black bars) across all experienced groups: learners, novices, experienced drivers with less than six years of driving experience, and experienced drivers with more than six years of driving experience (standard error bars are added).

8.5 Comparison of the multiple-choice test with the free-response test

A third additional analysis was conducted in order to directly compare performance on the new multiple-choice test with performance data from the free-response test used in Experiment 5. For that purpose, all the participants that took part in the proximal occlusion condition of Experiment 6 (N=20; ten novices and ten experienced) were matched to participants with the most similar driving experience (in terms of passing the driving test and annual mileage) from the current experiment (N=20; ten novices and ten experienced). This matching process was done without sight of the actual performance data. The average difference between matched pairs in terms of mileage was 76 miles for the novices and 400 miles for the experienced drivers, while the average difference in years' experience since passing the driving test was of 0.4 years for the novice drivers and 0.2 for the experienced (see Table 8.2).

Table 8.2: Mean mileage, driving experience and mean difference between experience and mileage for the matched groups

	Free response		Multiple-Choice			
Participants	Mean experience (years)	Mean mileage	Mean experience (years)	Mean Mileage	Difference experience (years)	Difference miles
Novice drivers	0.4	276.4	0	352	0.4	76
Experienced drivers	3.1	3300	3.4	3700	0.2	400

A 2 x 2 between-groups ANOVA was conducted on the two data sets. A marginal difference was found between the tests ($F(1,36) = 3.5$, $MSe = 152.8$, $p=.07$, $\eta_p^2 =.09$), with performance on the multiple-choice test being ostensibly higher than that on the free-response version. The main effect of driving experience remained with experienced drivers outperforming novices (82.3 vs. 69.7; $F(1,36) = 10.5$, $MSe = 152.8$, $p<.01$, $\eta_p^2 =.23$), although there was no interaction between the test type and the experienced groups (see Figure 8.3). No differences were found for the ratings. The hazards were not rated as more hazardous in terms of test type. These results do not provide any evidence to suggest that the tests differ regarding their ability to discriminate on the basis of driving experience, and that the type of question does not influence perceived hazardousness.

Pearson correlations were conducted to assess the relationship between the accuracy of responses, annual mileage and years of driving (post-license) for both types of test. Significant correlations were found for accuracy and years of driving (r(18)=0.51, p <.05), and accuracy and miles driven (r(18)=0.56, p <.05), but only with the data from the multiple-choice test. Neither of these correlations was significant for the free-response data from Experiment 5 (See Table 8.3).

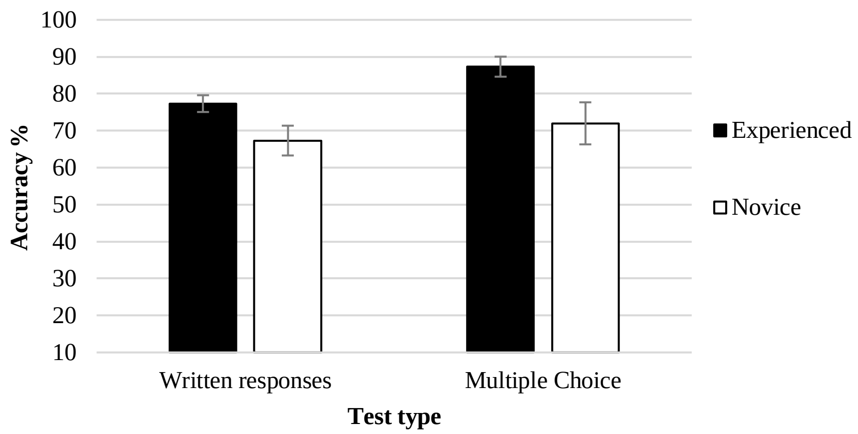


Figure 8.3: Percentage accuracy of correct responses across experience groups and test type (with standard error bars added)

Table 8.3: Pearson correlations for accuracy vs. years of driving and accuracy vs. miles for both types of tests (multiple-choice and free-response type)

	Multiple-Choice test			Free-Response test		
	Accuracy	Years	Miles	Accuracy	Years	Miles
Accuracy	–			–		
Years	.512*	–		.382	–	
Miles	.562**	.697**	–	.395	.531*	–

*** p < .05, ** p < .01**

8.6 Discussion

The multiple-choice hazard prediction test was able to discriminate between the driving groups on the basis of experience. Experienced drivers (E<6y and E>6y) outperformed the learner and novice drivers in predicting hazards. Planned contrasts suggested that the threshold for performance differences lies between the novice and moderately-experienced group, suggestive of a step change in the ability to predict hazards following one year of post-license driving experience. This is in keeping with the literature that consistently reports novice drivers, especially within the first 12 months of post-license driving, to be overrepresented in crashes. This suggests that safety-relevant skills are still developing in this first year of independent driving (Foss et al., 2011, McCartt et al., 2003; Williams & Tefft, 2014; Pradhan & Crundall, 2017).

While several studies have reported finding significant differences in accuracy between novice and experienced drivers when using multiple-choice formats (Lim et al., 2014, Ventsislavova et al., 2016), others have failed to find these differences (Malone & Brünken, 2013). However, when Malone and Brünken (2016) compared the traditional reaction time paradigm with the multiple-choice format, they found that experienced drivers outperformed novices in both test types. Following these results, they concluded that the ability of a test to discriminate between driver groups is more likely to be due to the scenario type (i.e. the particular videos used) rather than the assessment methodology employed.

On the basis of the current results, and those of other researchers reported in the literature, it appears that the multiple-choice format is no less discriminative than other test formats, yet pragmatically it offers a range of other advantages. For instance, the multiple-choice format allows quick administration and objective scoring of the answers. In contrast, a free response prediction test can take much longer to administer, and, without the development of natural language processing algorithms, must be scored offline by hand. Not only is this time consuming and risks errors during subjective coding, but it also means that participants cannot receive an immediate score for their performance. For research purposes, participants do not need to be given a score at the end of the test (though they typically prefer this). However, if the prediction format was to be used in driver training or licensing procedures, then the immediate feedback that can be provided with a multiple-choice format is a necessity.

The multiple-choice format could also encourage drivers to think and consider about a wider range of possible situations that could possibly happen. Typically, the hazard perception test would only require drivers to focus and consequently press for a single hazard, ignoring possible alternative situations. For instance, if a pedestrian steps out from behind a bus, the driver might think “there is a pedestrian” and press for it. However, in a real life scenario the pedestrian might wait and the bus could pull off instead. Anticipating just one hazard in the test may hinder drivers from considering alternative possibilities of how the situation might develop in the real world. By providing more options they become aware that a single situation could have many possible outcomes. This will require a more in-depth situation awareness where drivers would also need to decide which is the correct possibility. This can be especially beneficial for driver training and licencing procedure.

There are potential concerns however regarding the use of multiple-choice formats. Malone and Brünken (2016) argued that providing multiple options for drivers to choose between might simply tap into memory processes rather than driving skill. This was particularly pertinent for their study, which only presented the options after a hazard clip had been viewed in its entirety. In this case, drivers might have to choose an option that refers to a hazard that they had seen tens of seconds ago. While such a short temporal gap might not be considered to place a huge strain on memory, the point remains valid: any gap between the target event and the probe question might produce results confounded by memory processes. However, in the current study, the clip occludes just as the hazard begins to develop, and the options are immediately provided. The safest drivers will be aware of this hazard at the point of occlusion, and thus this answer should be readily available in working memory.

It is possible that the multiple-choice options prime participants to potential hazards other than the actual answer for each clip. This might explain why the multiple-choice version of the test is slightly easier than the free-response. However, this does not impact on the ability of the test to differentiate between experienced and novice drivers. Also, this is beneficial in terms of hazard training. For example, if a driver approaches a bus, they might be expecting a pedestrian to appear on the basis of the experience with the hazard perception test. In turn, if they have undertaken the hazard prediction test, the may be more likely to anticipate up to four potential hazards (three distracters and the correct answer).

One further concern with the multiple-choice format is the generation of incorrect options that provide plausible alternatives to the correct answer. If the distracter options are too unlikely, then the viewer may be able to guess the correct answer by rejecting the implausible options. For this reason, the distracters were developed using the responses given by participants in Experiments 5 and 6. We recommend this as a method for generating plausible and realistic distracters. While this process is time consuming, it helps ensure that the correct answer does not stand out.

When overall accuracy rates for the two tests are examined, the multiple-choice test is ostensibly easier (though this effect did not reach conventional levels of statistical significance; $p = 0.07$). Such an effect can be understood in terms of the opportunity for participants to guess, or infer, the correct answer out of the multiple-choice options. Arguably, an act of inference may still tap into hazard prediction skill (as participants may choose the most likely response on the basis of the evidence they had gathered up to the point of occlusion, even if they were not looking in exactly the right place at the right time). Pure guesses are potentially more problematic, but as the ostensible increase in accuracy did not negatively impact on the ability of the test to discriminate between our driver groups, it can be argued that it is of little consequence.

The ability of the multiple-choice test to differentiate between the driver groups suggests that it might have similar validity to that of the free-response test. Considering the correlation analyses, one might be tempted to argue that the multiple-choice test is even more sensitive to driving experience than the free response test, with years of driving (post-license) and mileage only correlating with the former but not the latter. The reason for this could be that some drivers with good experience may have not had the vocabulary to describe what they saw or the act of turning their prediction into a verbal code caused a drop in score. On the basis of the current results, it appears that the multiple-choice test is as good as the free- response test at discriminating between the driver groups. Given the positive correlations with measures of experience, and the pragmatic advantages that a multiple-choice format offers (objective marking, immediate feedback, quick administration, online testing across nations etc.), it can be argued that providing participants with options to choose between should be the preferred format.

In conclusion, the results suggest that a multiple-choice format of the hazard prediction test can provide an efficient and effective tool for discriminating between driver groups on the basis of driving experience. It is particularly sensitive to the boundary between novice and moderately-experienced drivers, which accords with the literature that identifies the first 12 months post-license as being particularly problematic for young drivers. With the growing evidence-base for hazard prediction tests as a valid method of discriminating driver groups, this study has demonstrated how such a test could be designed for widespread automated deployment, thus providing a valid alternative to the current UK hazard perception test that still relies on confounded response-time measures. This new hazard prediction test features real hazard footage, including wider range of hazardous situations such as overtaking and undertaking hazards. In order to include such hazards, the design of the traditional hazard perception test was improved to include mirror information, which allowed the anticipation of these hazards. This feature is especially beneficial for developing countries, where these types of situations are observed more often. The multiple-choice version is also beneficial for cross-cultural studies as it removes the need for translating participant responses, and subsequent hand coding. This allows quick administration, automated coding and immediate feedback.

The final challenge for this thesis is to apply this refined variant of the hazard prediction test to a brand new country.

Chapter 9

THE HAZARD PREDICTION TEST

Experiment 8

The Hazard Prediction test applied to the Israeli driving context

9.1 Summary of previous findings

The main motivation of this thesis was to examine whether the traditional hazard perception test is suitable for international export. The HP test has been frequently found to differentiate between experienced and novice drivers, and it seemed a logical step to consider exporting the HP methodology to other countries which could benefit from it. As mentioned in Chapter 1, several European Governments are considering the possibility of implementing the HP test as an official driving test in their respective countries (e.g Spain, the Netherlands), however, creating and adapting a hazard perception test suitable for any driving context is not an easy task. For instance, the potential for increased overtaking hazards requires the addition of mirror information. Furthermore, when the traditional HP methodology was tested overseas, it did not seem to repeat the same success as in the UK, in regards to finding the experiential effect.

When Chinese, Spanish and UK drivers were tested in the cross-cultural study in Chapter 5, the results showed that the HP test differentiated drivers based on the cultural background of the participants and not on their experience levels. Previous studies conducted in Malaysia showed similar results (Lim et al., 2013). Therefore, an alternative version of the HP test was considered: the hazard prediction test. Although this test was first created in 2009 (Jackson et al., 2009; but with a pedigree dating much further back: McKenna and Crick 1997), a new and improved version was developed for this thesis featuring video clips filmed in China, Spain and the UK. The crucial difference between the two tests, is that the latter variant replaces potentially cofounded response-time measures, with measures of accuracy to the question “what happens next?” following occlusion at the point of hazard onset. In Chapter 7, a new cohort of Chinese, Spanish and UK drivers were tested using this new version of the hazard prediction test and it successfully discriminated between the experienced and novice drivers. No differences were found for nationality this time.

The hazard prediction test also appeared to be successful in differentiating between driver groups in a multiple-choice format. Although in Chapter 8 the multiple-choice format was applied only to a UK sample, previous studies conducted overseas (Lim et al., 2014; Ventsislavova et al., 2016) have featured Spanish and Malaysian drivers and have been successful in differentiating between experience groups. Nonetheless, the results of the study conducted in Chapter 8 showed that only the multiple-choice format was significantly correlated to mileage and years of driving when compared to the free-response version, which suggests that, beyond the practical advantages of scoring multiple-choice questions, this format might better reflect driving experience.

The results so far have suggested that the hazard prediction test is more effective as an overseas measure of driver safety, and have provided a potential blueprint for how to design such a test (e.g. with mirrors, using multiple-choice questions, etc.). However, all of these findings need to be confirmed with a new hazard prediction test, designed for a completely different country, following the guidelines suggested in previous chapters. After testing across Europe and Asia, for this last study the hazard prediction test was applied in a country from the Middle East: Israel.

9.2 Introduction

9.2.1 Features of the hazard prediction test and evidence of its effectiveness

The most remarkable feature of the hazard prediction test is its ability to differentiate drivers based on their driving experience and not on their cultural background. This is an extremely important feature as it removes threshold bias and other biases motived by the social and cultural rules of each country (Lim et al., 2014). Participants are not asked to decide whether a situation is hazardous or not but to predict a future event on the basis of clues in the current environment. Furthermore, each test should be adapted to the specific cultural context it is aimed for, however its ability to measure the prediction/perception ability should not be influenced by cultural bias.

Several studies have discussed the influence criterion bias has on hazard perception (Egea-Caparros et al. 2016; Horswill & McKenna, 2004; Horswill and Wallis, 2007; Underwood et al., 2013). Wallis and Horswill (2007) reported that novices required a higher threshold to consider a situation as hazardous and Egea-Capparos et al. (2016) reported that a delay in detecting hazards is linked to a stricter criterion. Conversely, Ventsislavova et al., (2016) did not find differences for criterion between experience and novice drivers.

As neither criterion nor sensitivity have been linked directly to accuracy in prediction, in this study a model is proposed including both variables together with driving experience. It is expected that the model will significantly predict accuracy as more experienced drivers are more sensitive to hazards with less conservative criterion for considering a situation as hazardous. However, if criterion does not significantly predict accuracy this would support the notion that hazard appraisal does not impact on how accurate one is in predicting hazards.

The second important feature of the hazard prediction test is that the new stimuli contain mirror information which not only reflect a more realistic driving environment (and the demands of a more realistic environment) but also provide cues that help to predict overtaking hazards. Overtaking and undertaking hazards are not included in the national UK hazard perception test as the viewer does not have access to any precursors to such hazards. Nonetheless, overtaking and undertaking hazards do pose a problem on UK roads, and even more so in developing countries; the provision of mirror information allows these hazards to be included. Drivers also need to pay attention to mirrors when overtaking or changing lanes, and mirror information in hazard perception clips can be very useful in assessing whether drivers check these areas of potential danger (Shahar et al., 2012; Zhang et al., 2016).

The third important feature is that the latest variant of the hazard prediction test is both time and cost effective, which would be especially beneficial for developing countries. The original hazard prediction test was developed as a free-response test, which took time to code and had the potential problem of rater error. For that reason, a multiple-choice version was created which provides automatic and unambiguous coding. Several studies have already adopted this methodology (Crundall & Kroll, 2018; Lim et al., 2014; Ventsislavova et al., 2016), however the study conducted in Chapter 8 was the first to compare both a free-response vs. a multiple-choice version. Such a comparison was necessary, as it was possible that the MCQ variant increased the salience of the correct answer. In Chapter 8, we found that both versions successfully differentiated between the experienced groups but only the multiple-choice version correlated significantly with driving experience. Thus the MCQ format may not only be more efficient, it may also better reflect driving experience.

Considering the previous success of the hazard prediction test in differentiating between experienced and novice drivers within different countries (Crundall, 2016; Lim et al, 2014; see also Experiment 6), the final study of this thesis will focus on creating a completely new hazard prediction test featuring a completely different cultural driving context. This is necessary to demonstrate that the blueprint for a culturally unbiased hazard prediction is not specific to those countries it has been developed on (i.e. China, Spain), but that the application of these test guidelines in any country can produce a valid test of driver's hazard prediction skills.

9.2.2 Guidelines followed to create the test

In order to develop a prediction test suitable for the Israeli driving context and ensure its validity and successful differentiation between experienced and novice drivers, the guidelines that were created and examined in previous chapters of this thesis were followed.

9.2.2.1 Mirrors

Although, there was no evidence to suggest that mirror information enhances hazard perception performance in Experiment 1, neither was there any evidence that the inclusion of mirrors damaged the effectiveness of the test. Mirrors were therefore included to create a context that reflects real driving, and to allow the inclusion of overtaking and undertaking hazards. While filming the footage used for the current Israeli study, it was observed that, similar to China (see Chapter 5), the Israeli driving context contained a wide range of both overtaking and undertaking hazards. This fact makes inclusion of mirrors highly necessary in order to avoid excluding such hazards, which are representative of the Israeli driving context.

9.2.2.2 Car overlay

Regardless of the lack of evidence that the transparent overlay is more beneficial in terms of hazard detection, the current clips were designed to contain a transparent overlay for obvious reasons (see Chapter 2). Many hazards and precursors can be partially occluded by the A-pillars, which provides an unfair hindrance. In the real world such partial occlusion can be mitigated by small head movements and stereopsis. Accordingly, the transparent overlay was used. This still partially obscures some precursors and hazards, but should the driver choose to focus on an occluded object, she will still be able to process it (in the same way that a driver might move their head slightly to focus on a potential hazard behind an A-pillar on a real road).

9.2.2.3 Non-hazardous trials

The prediction test, up to this point, has avowedly avoided any concerns of appraisal bias or hazard criterion. In this final test, it was decided to add in an element of appraisal to assess how this interacts with the prediction task. To this end, non-hazardous trials were included in the current study. These were situations which did not end into a materialised hazard (see Table 9.1). Prior to being asked “What happens next?” participants were asked “Did you see a hazard?”. Inclusion of non-target trials in this way allows a more extensive analysis where the Signal Detection Theory (SDT) could be applied to look at the sensitivity to hazards and criterion bias of participants that has not been possible so far in this thesis. This represents a divergence from the current studies contained in this thesis, but is not without precedence (Ventsislavova et al., 2016). The SDT allows measurement of two important aspects of the decision-making process: sensitivity to the signal which reflects the intensity of the stimulus (measured by the number of correct identifications and false alarms) and the criterion bias that guides our decisions (measured by the tendency to report an event as either hazardous or not). In the hazard perception field, increased sensitivity to the signal translates into better ability to detect hazards and this sensitivity is compared to our threshold bias. While a liberal criterion shows the tendency to report each event as hazardous and a conservative criterion requires a higher threshold for considering a situation as hazardous. Wallis and Horswill (2007) however, used a fuzzy signal detection theory arguing that the traditional SDT is not suitable for tasks such as hazard perception. The fact that potential hazardous events that have not yet become hazardous could still be labelled as such, makes an assessment of a binary true state difficult. The fuzzy signal detection theory allows calculation of continuous outcomes instead of discreet levels, which in turn allows the evaluation of potentially hazardous scenarios. Wallis and Horswill (2007) assessed two models. The first one proposed that novices are less sensitive to hazards than experienced drivers, while the second one stated that novices required a higher threshold to consider a situation as hazardous. Their findings supported the second model, showing that novices were as sensitive to the hazards as the experienced drivers, however required a higher threshold of danger in order to classify an event as hazardous. In a later study, Ventsislavova et al., (2016) argued that it is possible to apply the classical SDT if speeded responses are removed and a simple binary probe question is adopted instead (e.g. Is there a hazard? Yes-No). They explored both hazardous and quasi-hazardous situations and contrary to Wallis and Horswill (2007) found that both non-offenders and experienced drivers were more sensitive to hazards than offenders and novices. However, no differences were found for criterion.

The above studies applied different paradigms which might explain the contradictive findings. Furthermore, some studies conducted in Israel reported that novices find potential hazards particularly difficult to identify (Borowsky et al., 2009; Borowsky & Oron-Gilad, 2013). In order to find out whether experienced drivers will be more sensitive to the non-hazardous trials than novices, the present study replicated the method of Ventsislavova et al. (2016) using the classical SDT and a binary probe question (Is there a hazard?). The aim was to examine whether Israeli experienced and novice drivers show differences in their sensitivity to both hazardous and quasi-hazardous situations, and whether there will be differences in criterion for both types of situations.

9.2.2.4 Occlusion point

To ensure that enough information is provided to allow drivers predict the hazardous situations, clips were cut according to the proximal occlusion points. Previous studies (Crundall, 2016) and the results of the study in Chapter 6, suggested that cutting the clip too early may diminish the link between precursors and hazards, potentially removing the distinction between novice and experienced drivers. Proximal occlusion points also produce the highest overall accuracy, which may be important in encouraging ongoing voluntary participation (especially if a test is provided online). If drivers note a mismatch between their ability to perform the test, and their self-belief in their own driving skill, this will reduce their faith in the validity of the test, and may lead to early withdrawal.

9.2.2.5 Multiple-choice format

Finally, results in Experiment 7 showed that only the multiple-choice version of the hazard prediction test correlates with driving experience, suggesting it to be a better measure for the underlying hazard skill. The further advantages of automatic and unbiased scoring, led to an MCQ format being chosen for the current test.

9.2.3 Israel

Israel already has a strong record of traffic psychology research. In recent years there have been studies in hazard perception (e.g. Hoffman & Rosenbloom, 2016), young drivers (Ben Ari, Skvirsky, Greenbury & Prato, 2018; Korn, Weiss & Rosenbloom, 2017), risky driving (Taubman-Ben-Ari, Mikulincer & Gillath, 2004) and driving behaviour among the ultraorthodox population (Guggenheim & Taubman – Ben-Ari, 2015, Rosenbloom et al. 2004; Rosenbloom, Shahar, & Perlman, 2008).

Israel is considered a familiaristic and collectivistic culture (Kagitcibasi, Ataca & Diri, 2010). Several authors have argued that driving behaviour of young drivers is especially influenced by the norms set by their family and close friends (Hartos, Eitel, Haynie & Simons-Morton, 2000; Taubman-Ben Ari & Katz-Ben-Ami, 2012; Guggenheim & Taubman-Ben Ari, 2015). For instance, young drivers tend to report being more committed to safe driving when parents are role models for safe driving, than those who are less committed. However, parental support is not noticed for the ultraorthodox young men, as there are strict restrictions regarding driving in their community. Eight percent of the Israeli population is comprised of the ultraorthodox Jews. These communities restrict driving for those young men who have yet to marry, and sanction those who decide to learn to drive. However, many young males still desire to obtain a driving licence, most of the time hiding it from their parents and community. Therefore, driving under such circumstances of constant socio-cultural tension brings an additional risk and affects the driving of these young males (Guggenheim & Taubman – Ben-Ari, 2015). Although there is no official statistics that account for accidents in this community, the combination of prohibition, the lack of family support and limited practice could enhance risky behaviour.

The other major population in Israel is Arab (approx. 20% according to the Central Bureau of Statistics, 2012), with a culture and lifestyle that differ considerably from those of the Jewish population. Towns populated with Arab residents report higher transport-related fatality rates compared to those mostly populated by Jewish people (Baron-Epela, Obid, Fertiga & Gitelman, 2015). This has been attributed to the lower level of road infrastructure and more violations of traffic laws in the Arab community (Gitelman, Dain, Levi, & Eizenman, 2003). According to the Central Bureau of Statistics (2016) Arabs are overrepresented in traffic accidents, accounting for over a third of all fatalities (35%).

In 2016 there were 12,430 road crashes in which 372 people died (Central Bureau of Statistics, 2017). To put this into context, that is one fatality for every 23,440 residents (compared to one fatality per 36,842 residents in the UK). Vehicles (33%) and pedestrians (34%) account for the majority of the Israeli traffic accidents. During the period 2000-2015, the largest decrease in fatal injuries was observed for the 18-20 year olds (68.8%), followed by the 15-17 years olds (52%) and 0-14 years olds (48%) (Road

Safety Annual Report, 2017). Most of the collisions are reported in Tel Aviv, due to the fact that it is the most congested city in Israel, especially during rush hours, as vehicles are accessing and leaving the city every day. Moreover, the most common factor in traffic collisions is speeding (just over 10%), yet despite this, 20% of Israeli drivers do not believe that speeding increases the risk of traffic crashes (Elias, 2018).

Driving lessons in Israel can begin at 16 years and five months, however the road test can only be held once the learner has turned 16 years and nine months (with a minimum of 28 on-road driving lessons). In order to obtain a driving licence, learners undertake both a theory and an on-road driving test, and an experienced driver is required to accompany them for the first three months after licencing.

Besides the cultural diversity and population in such a small country, another factor that was found to contribute to the increase of fatality rates are terror attacks. Terror attacks seem to influence crash rates (Stecklov & Goldstein, 2010). For instance, while the immediate aftermath of a terrorist attack may result in a temporary lull in slight collisions, perhaps due to increased security presence, on the third day after a terrorist attack Stecklov and Goldstein (2004) noted a 35% increase in fatal collisions (which they interpret in terms of psychosocial stress).

All these aspects make Israel a context worthy to explore and, although this chapter will not focus on the differences between populations within the country, it will be interesting to assess whether the hazard prediction test can differentiate between driver groups in such a country where cultural influences differ significantly to those of the UK.

9.2.4 Hazard perception in Israel

Unlike the hazard prediction test, the more typical hazard perception test has already been studied by several research groups within the Israeli context. Some of the studies have solely focused on examining the hazard perception skill within different driving groups (Rosenbloom, Perlman & Pereg, 2011; Hoffman & Rosenbloom, 2016), while others have also added hazard categorisation, risk estimation and hazard fixation (Borowsky et al., 2010; Borowsky, Oron-Gilad, 2013). Rosenbloom et al. (2015) even developed a hazard perception test suitable for pedestrians with ostensible benefits as a training system to help improve pedestrians' safety.

Rosenbloom et al. (2011) tested the hazard perception skill of motorcyclist vs. car drivers using car-based hazard clips and a 5-point scoring system based on that used in the UK national test. They found the motorcyclists to outperform the car drivers, while crash-involved drivers also scored lower on the HP test than crash-free drivers. Another interesting study by Hoffman and Rosenbloom (2016) where the HP test was used, reported that experienced drivers showed higher anxiety under implicit threat when compared to novices, however they did not find differences between experienced and novice drivers.

Borrowsky et al. (2010) studied hazard perception in novice, experienced and elderly drivers featuring potential and materialised hazards. They found that novice drivers are worse at identifying potential hazards than the experienced drivers (however they did not find differences for materialised hazards), which supports the notion that novices have not yet developed their situation awareness, and ability to predict hazards. Potential hazards are more difficult to identify as they are not as obvious as materialised hazards and precursors to these hazards have been found to be more difficult to spot (Crundall et al., 2012). This supports the argument that potential hazards, or precursors, are better indicators of driving experience. Another study by Borowsky et al. (2009) that has considered potential hazards (i.e. precursors that do not ultimately lead to hazards), found that experienced drivers classified hazardous situations according to similarities in the traffic environment (urban settings), while novices classified the same situations according to the hazards (a pedestrian). This might suggest that with age and experience drivers tend to focus more on the global aspect of the situation rather than the local aspects (hazard), which in turn might lead to a better precursor detection (as they know what hazards might appear in such environments). Therefore, novices find difficulties in anticipating hidden hazards. Similar results were found comparing taxi drivers to novices, with the former being more sensitive to hidden hazards (Borowsky & Oron Gilad, 2013).

Some of the above findings are consistent with previous literature that failed to find differences between experienced groups in the detection of materialised hazards (Chapman & Underwood, 1998; Crundall et al., 2003; Sagberg & Bjørnskau, 2006;

Borowsky et al., 2010; Underwood et al., 2013), and support other studies that have looked at potential hazards (Vlakveld, 2014) or precursors (e.g. Crundall, 2016). For that reason, it would be interesting to test the hazard prediction skills of this population and to assess whether experienced drivers will perform better at identifying precursors that will help predict the hazardous situation.

9.2.5 Overview of the study

The aim of this study was to create a hazard prediction test featuring hazardous driving situations recorded in the Israeli driving context, and test whether they can differentiate between experienced and novice drivers on the basis of drivers selecting the correct option when asked "What happens next?" following occlusion. In addition, prior to the multiple-choice question, drivers were asked whether they thought a hazardous event was about to occur. This allows data on hazard appraisal to be captured at the same time as hazard prediction, decoupling the hazardous nature of the post-occlusion event from the actual prediction.

It was expected that experienced drivers would outperform novices in prediction accuracy and that the hazardous situations would be more easily predicted than the non-hazardous. It was also possible that experienced drivers' appraisal of the hazardousness of a predicted event relates differently to their predictive accuracy compared to that of novice drivers (e.g. criterion bias might be evident).

It was also predicted that sensitivity, criterion and driving experience will be significant predictors for accuracy in prediction, although considering the mixed results regarding the relation between sensitivity, criterion and accuracy in prediction, the relation can go either way.

To this end, 13 hazardous situations and 13 non-hazardous situations were selected from footage filmed in Tel Aviv. These clips were then converted into a multiple-choice hazard prediction test. Participants were asked first whether they have seen a hazard or not, followed by four possible options to choose from. Only the hazardous trials contained multiple-choice options.

9.3 Method

9.3.1 Participants

In total 81 participants took part in this experiment, however seven were removed due to not completing the experiment according to the instructions. These seven participants spent less than five seconds on the video page when they were expected to remain an average between 14 seconds and 44 seconds (depending on the duration of each video, the average duration was approx. 21 seconds). Forty-three of the participants were experienced drivers with an average of 14 years of driving experience (average mileage of 6858) and 31 were novices (including 19 learners) with 1.5 months of driving experience (average mileage of 619.2). The mean age of the experienced drivers was 33.5 years old and of the novices 18.8 years old.

All participants were native Israeli drivers that had mainly driven in Israel. None of the participants had to undertake a hazard perception or hazard prediction test to obtain their driving licence, therefore this sample was not familiar with the HP test (contrary to the UK drivers).

Participants were contacted via social media where the study was advertised offering a link to the hazard prediction test. Thus, it was possible to reach a wider and more representative sample than typical opportunity sampling within a campus community. Participants were asked to click on the link and undertake the test following the instructions. Finally, they were asked to provide an email address in order to claim their 48 Israeli Shekels voucher (equivalent £10).

9.3.2 Design, apparatus and materials

A between-groups design was used, where the primary factor was the driving experience and the dependent variables were the prediction accuracy and whether participants believed there was an imminent hazard.

To create the hazard prediction stimuli filming was undertaken in Tel Aviv (Israel). The footage was recorded with a mini HD video camera (the procedure of recording was identical to the previous studies; see Chapter 2). In total, 13 hazardous and 13 non-hazardous videos made the final cut of stimuli. Each hazardous clip was occluded prior to the hazardous situation fully materialising. There was always, however, a precursor that would help participants predict how the hazardous situation was going to develop. For example, in one clip the camera car is driving on a three-lane carriageway. An urban bus from the right intends to join the lane. The clip occludes when the bus has started slightly to turn left and has emitted a couple of flashes of the indicator. Regarding the non-hazardous video clips, although there was always a source of potential hazard prior to occlusion, none of these precursors led to an actual hazard. For example, in one non-hazardous clip the camera car is driving along a dual carriageway and there is a cyclist in front at the point of occlusion. Had the clip played in full, participants would have seen that the cyclist continues her trajectory without causing any danger, and thus not creating a hazardous event. All hazards were captured naturalistically. A full description of the hazards can be seen in Table 9.1.

The stimuli were presented in a randomised order to participants via Qualtrics online (www.qualtrics.com). They were asked to undertake the experiment using a computer or laptop, but to avoid using an IPad or a mobile phone (due to the poor visibility of the video clips).

9.3.3 Test development

Prior to the actual experiment, a pilot study was undertaken in order to test the video clips and generate plausible alternatives for the multiple-choice questions. Two experts in traffic and transport psychology selected the initial set of hazardous videos and occlusion points (identical criteria as with the other experiments was followed to select the clips, please refer back to Chapter 2). Four participants (that did not participate in the final experiment) were asked to watch the initial set of 50 videos and predict the hazardous situations. Two were non-drivers and two experienced drivers (three years and nine years of driving experience, respectively). The non-drivers found it particularly difficult to predict hazards when mirrors were required or when the precursor was an indicator to turn. The scores of the pilot participants were 30.8 %, 38.5%, 46% and 61% of accuracy for non-drivers and experienced driver (three and nine years), respectively. Clips that were predicted by all four participants and those that appeared to be too difficult (zero score) were removed.” The initial set of stimuli was composed of 15 hazardous videos, however two were removed as they appeared to be extremely difficult to predict even for our two experienced drivers. The final set was composed of 13 hazardous videos, with each hazard successfully predicted by at least one of the participants.

After the clips were selected, an online Google Form questionnaire was created where participants were asked to watch the clips and write down as many possible answers as they could think of, in regard to how the driving situations might develop (they were asked to provide a minimum of four options). Twelve participants, all Israeli drivers with more than 3 years of driving experience, completed the online questionnaire. They were recruited online via social media platforms and were provided with a link to the Google Form questionnaire. Their answers were then subjected to a frequency analysis where semantically similar answers were grouped together. The three most frequent incorrect alternatives were used for each video clip, along with a correct answer (the wording of which was inspired by participants' responses).

Once the final set of videos was selected, and the multiple-choice options created, the final test was developed using Qualtrics. The experiment (instructions, consent form, alternatives etc.) was translated from English into Hebrew following the forward backward translation procedure (following the guidelines of International Test Commission; ITC, 2010). The translation into Hebrew was performed by a team consisting of two bilingual experts with a high level of expertise in both Hebrew and English, and expertise in traffic regulations and driving habits.

Participants were asked to read through the consent form and sign it if they agreed to take part in the study (by selecting a tick box in the Qualtrics form). Prior to the first session, participants viewed a practice trial, where they had the opportunity to familiarise themselves with the experiment. They received feedback on their performance only during the practice; that is, following their selection of a multiple-choice option, the clip was replayed in full, showing the fully materialised hazard.

As the experiment was created using an online tool, videos were programmed to start automatically and the progress bar (which typically appears on each video) was removed. This was necessary, as participants would have been able to see when the video was about to finish, thus priming them that the occlusion point was approaching. Participants were also instructed to avoid clicking anywhere on the page while undertaking the experiment (as this would inadvertently pause the video). They saw 13 hazardous and 13 non-hazardous video clips and the first question they were required to answer after each video was whether they had seen an upcoming hazard by ticking yes or no. Following this, they were asked to predict how the traffic situation was going to develop by choosing one of the four options provided after each video. Only the hazardous trials contained multiple-choice options.

Table 9.1: A description of the a priori hazardous and non-hazardous traffic situations selected for each clip

N	Type of clip	Content	Hazard	Duration of the clips (ms)
1	Hazard	You are driving along a narrow residential road, with cars parked on the right side of the street. One of the parked cars turns on its brake lights and another car emerges from the left side of the street. The clip occludes at the moment when the front part of the emerging car is partially visible.	A car emerging from the left	16800
2	Hazard	While driving you are approaching an underpass. You enter and there is a congestion of cars on both sides. The car on the right signals and cuts your path. The clip occludes with the first flash of the intermittent.	A car on the left side cuts your path	27866
3	Hazard	You are driving on an urban dual carriageway road and there is an additional lane where there are parked cars. On the right mirror a taxi is visible and it undertakes your car just immediately before colliding with one of the parked cars. The clip occludes at the moment when the taxi is no longer visible at the right mirror.	A taxi undertakes form the right	40666
4	Hazard	You are waiting due to red traffic lights. The road ahead diverse and there is a congestion of cars on your left side. On your right mirror is visible a van that is approaching towards the diversion. It quickly undertakes your car and squeezes between into your lane. The clip occludes at the moment when the van is no longer visible on the right mirror.	A van undertakes and cuts your path	22300
5	Hazard	You are driving on a narrow urban road and there is a truck in front of your car. There are parked cars on the right side of the road and a pedestrian is approaching towards the street. The clip occludes at the moment when the pedestrian looks at your car.	A pedestrian steps into the road from the right	16166
6	Hazard	While you are driving along a congested urban road the car in front diverse into the left lane and then suddenly brakes and cuts your path. The clip occludes after the car’s intermittent emits a couple of flashes.	The car on the right cuts your path	21166
7	Hazard	You are driving along the middle lane of a motorway while the right lane is heavily	A bus from the right cuts my path	18333

		congested. One of the urban buses from the right lane cuts yours and other vehicles' path. The clip occludes when the bus has slightly started to turn left and have emitted a couple of flashes.		
8	Hazard	You are driving along a residential area with pedestrians crossing the street. You are approaching a roundabout where there is a zebra crossing. Suddenly a cyclist appears from the left and crosses the street. The clip occludes when the cyclist is visible on the right window of the car.	A cyclist from the left crosses the street	19500
9	Hazard	While driving along a narrow road, it becomes visible that the car in front has its emergency lights on, while still moving. The car then turns left into a side street and starts reversing back into the main road. The clip occludes at the moment when the car starts to reverse.	A car reversing back into to the main street	13000
10	Hazard	You are driving along a congested urban road, all traffic lights are green. On the right-hand side, there are cars waiting to join the main road. A motorbike appears from the right and approaches the main road without stopping at the red lights. The clip occludes at the moment when the motorbike is about to join your lane.	A motorbike joins the main street form the right	19700
11	Hazard	You are driving on a three-way carriageway. An urban bus from the right lane intends to join your lane. The clip occludes when the bus has started slightly to turn left and have emitted a couple of flashes.	A bus cuts your path from the right	14600
12	Hazard	You are waiting while the traffic lights turn green. You start driving slowly as the traffic is congested. On the left mirror is visible a motorbike which approaches very fast and cuts your path. The clip occludes when the motorbike is no longer visible at the mirror and is visible on your left window.	A motorbike cuts your path from the left	19233
13	Hazard	You are driving on a narrow road with parked cars on your right side. One of the parked cars is trying to join your path. The clip occludes at the moment when the brake lights of the car are visible.	A parked car from the right tries to join your way	16200
1	No Hazard	You are driving along a narrow street and the traffic lights turn red	-	31800
2	No Hazard	You are driving along a motorway and the two-lane narrows into one lane	-	19733
3	No Hazard	You are driving on a busy urban road and a bus is waiting on your right	-	22200
4	No Hazard	You are driving along a narrow dual road and there is a cyclist in front of you	-	21300
5	No Hazard	You are driving along an urban road and the traffic lights ahead turn red	-	17800
6	No Hazard	You are driving along a busy urban road with pedestrians and cars on both sides	-	23100
7	No Hazard	You are driving along a busy road with cars constantly undertaking you	-	21100
8	No Hazard	You are driving along a motorway and there is a motorcyclist in front of you	-	17300
9	No Hazard	While driving the car in front of you brakes but then continues its way and joins the roundabout	-	21100
10	No Hazard	You are driving on a three-lane motorway with a busy left lane	-	25000

11	No Hazard	You are driving on a busy motorway with cars constantly braking in front of you	-	16900
12	No Hazard	You are driving along a five lane motorway and traffic circulates around you	-	23066
13	No Hazard	You are driving along an urban road and there is a lorry in front of you	-	21033

9.4 Results

9.4.1 Is there a hazard?

An initial independent sample t-test was conducted to compare the percentage of participants' hits across driving experience (i.e. clips that participants correctly reported whether or not there was an upcoming hazard). Mean hits for hazard detection were almost identical for experienced and novice drivers (43% and 44%, respectively; (t(72) =.07, *p*=.94).

These hits were then broken down into hits for clips that led to hazards, and non-hazardous clips. A 2 x 2 repeated measures ANOVA was conducted to compare accuracy for hazard and non-hazardous clips across the two experienced groups. A main effect was found for the type of clip, with non-hazardous clips receiving significantly more correct responses than the hazardous videos (53% vs. 43%, respectively), ($F(1, 72)=6.05$, MSe = .04, p <.01, η_p^2 =.07) (see Figure 9.1). However, there was no significant difference between the driver groups ($F(1, 72)=.35$, MSe = .04, p =.55, η_p^2 =.00), and neither there was a significant interaction ($F(1, 72)=.23$, MSe = .04, p=.62, η_p^2 =.00).

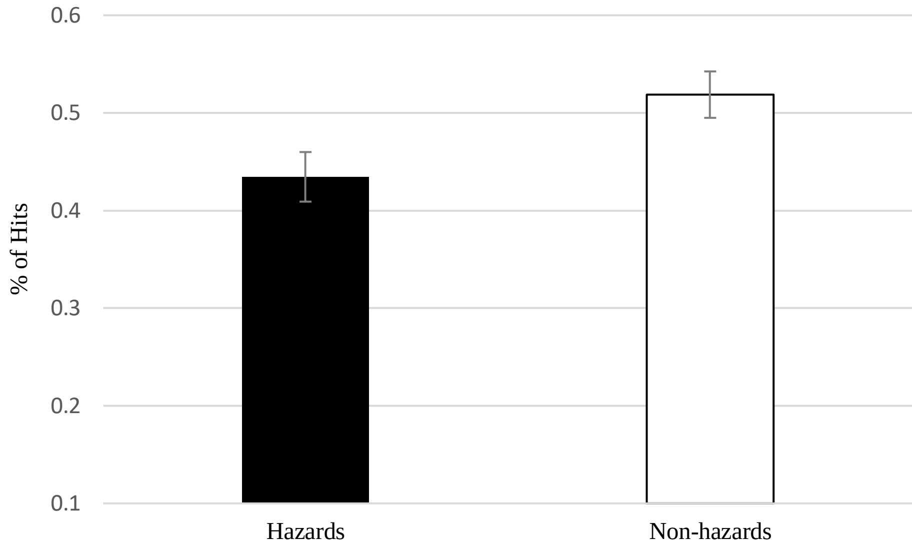


Figure 9.1: Percentage of hits for the questions "Is there a hazard?" across clip type with standard error bars.

The binary nature of this initial question also allowed Signal Detection Theory measures to be calculated. Sensitivity (d') was measured by calculating how well participants can discriminate between a hazard-present and a hazard-absent trial (hits and false alarms, using a binary classification of response). Hits were identified as those trials where there was indeed a hazard and participants responded affirmatively. Misses were identified as those trials where there was a hazard, but participants did not believe there to be a hazard. False alarms were those trials where there was not a hazard, however participants reported they have seen an imminent hazard. Finally, correct rejections were those trials where there was not a hazard, with participants responding that they had not seen an upcoming hazard. A t-test on d' failed to show any differences between driver groups (t(72) = -.52, *p*=.60).

A similar t-test was conducted to compare drivers' criterion c, which is the tendency to report everything as either hazardous or non-hazardous. The response bias was measured using criterion (c) rather than index β, as it has been argued that is more appropriate for vigilance-like tasks (Wallis & Horswill, 2007; See et al., 1997). No main effect was found for response bias. Experienced drivers did not have a tendency to report more (or fewer) hazards than the novices (t(72) = -.24, *p*=.81)

9.4.2 Accuracy on the multiple-choice prediction question

Prediction accuracy for the multiple-choice questions was calculated. An independent-samples t-test, comparing percentage accuracy across the driving groups, showed a clearly significant difference between experienced and novice drivers ($t(72) = -2.48$, $p<.01$). Experienced drivers outperformed novices in correctly predicting the hazardous driving situation (34.3% vs 26.6%, respectively) see Figure 9.2.

A t-test was also conducted to compare the mean accuracy scores of each group of drivers to mean chance expectancy (25%). This outcome was significant for the experienced group ($t(42) = 4.46$, $p<.001$), but not for the novices ($t(30) = .74$, $p=.46$).

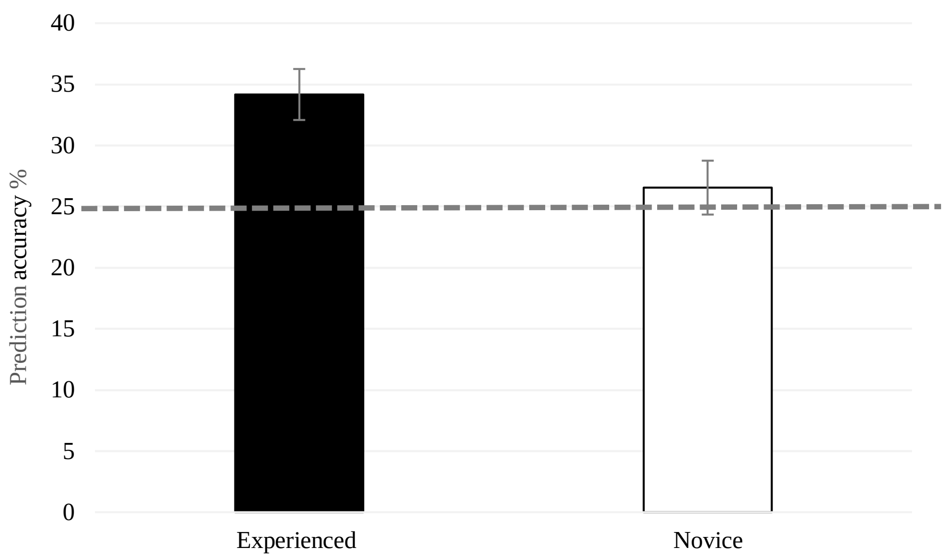


Figure 9.2: Percentage of prediction accuracy across driving groups with standard error bars. The dashed line represents the mean chance expectancy

A multiple regression was carried out to investigate whether driving experience (number of months since passing the driving test) and hits from the first question (correctly saying that a hazardous event is about to occur) could significantly predict accuracy in prediction. The original model contained age as a third predictor, however due to multicollinearity between age and driving experience (VIF=10.9; r=.95) age was removed from the model. According to the results, 9.8% of the variance can be explained by both predictors (driving experience and hits). The model was a significant predictor of performance, $F(2,73) = 4.95$, $p < .01$. Both predictors contributed to the model: experience ($\beta = .03$, $p < .01$) and hazard hits ($\beta = .14$, $p < .05$). The final predictive model was: Accuracy score = 21.553 + (.03*Experience) + (.14*Hits).

Another regression was conducted where sensitivity (d'), criterion (c) and driving experience were set as predictors, with accuracy as the outcome. Both sensitivity ($\beta = 5.90$, $p < .05$) and driving experience ($\beta = .03$, $p < .05$) appeared to be significant predictors for accuracy. In turn, criterion was not a significant predictor ($\beta = -1.9$, $p = .43$).

The overall model was significant $F(3,73) = 3.8$, $p < .05$ and explained 10% of the variance. The final predictive model was: Accuracy score = 29.5 + (.03*Experience) + (5.90*Sensitivity) + (-1.93*Criterion).

9.5 Discussion

This is the first time a hazard prediction test was specifically developed for the Israeli driving context, and the results appear promising. Although, no differences were found for hazard detection (which is consistent with the previous literature), this newly developed test was successful in differentiating between experienced and novice drivers for accuracy in prediction, as predicted.

Although, this study focused on the predictive element of hazard perception, sensitivity to hazards and criterion bias were also examined. Experienced drivers did not appear to be better at discriminating between hazardous or non-hazardous situations (sensitivity), neither were there differences between the groups in relation to the threshold bias (experienced drivers did not seem to be more, or less, liberal in their responses). This was surprising as the present results contradict previous studies that have found differences for at least one of the parameters. For instance, both Wallis and Horswill (2007) and Crundall et al., (2003) argued for criterion differences between expert and novice drivers (albeit in different directions). In both of these previous studies, the participants witnessed fully-materialised hazards. In the current study, however, before one can appraise the upcoming event as a hazard, one must first successfully predict it. Novices' lower prediction accuracy in the current test is likely to have reduced their likelihood of appraising an imminent event as hazardous (i.e. if they have not predicted that anything will happen, they will not consider the clip to contain anything hazardous).

Ventsislavova et al., (2016) found differences between experienced groups, with experienced and non-offender drivers showing greater sensitivity to hazards than novices and offenders. Experienced and non-offender drivers were more likely to report hazards than the other two groups. In turn, they did not find differences for criterion bias. It should be noted though that Ventsislavova et al. (2016) applied the hazard prediction test measuring accuracy in prediction, while Horswill and Wallis (2007) used response latencies. One possibility is that the inconsistent results could be due to differences in the main variable and the methodology used to measure it. However, the present study also applied the hazard prediction test and unlike Ventsislavova et al. (2016) did not find differences for sensitivity between the experienced groups.

Overall, participants seemed to be more biased toward reporting situations as non-hazardous, as they showed more conservative attitude (Criterion C=0.15, which reflects a tendency to not report hazards). Perhaps, the Israeli drivers required a higher threshold for the hazardous situations similar to the Chinese drivers (see Chapter 5). It has been found that threshold bias influences hazard perception accuracy (Horswill & McKenna, 2004). However, it is difficult to conclude whether it is a threshold bias issue, as only 14.1% of the participants reported there was no hazard, but then went on to successfully predict the future situation on the multiple-choice question. In turn, 58.3% of the participants reported that they do not consider the future situation as hazardous and also failed to predict it correctly, and only 16.8% responded to both questions correctly. If participants did not report the situation as hazardous but were still able to predict it, than we can safely assume that they indeed require higher threshold bias. Results are unclear and further investigation is needed in order to find out whether this is the case with the Israeli drivers.

Egea-Caparros et al. (2016) looked at response latencies similar to Wallis and Horswill (2007) and reported a negative correlation between the sensitivity and the criterion bias meaning that a higher threshold bias (more conservative) translates into less sensitivity to hazards. This relation seems reasonable as drivers that do not consider a certain situation as hazardous are less likely to report it as such. Egea-Caparros et al. (2016) also reported a significant correlation between criterion and response latency, similar to other studies (Horswill & Wallis, 2007; Underwood et al., 2013). The greater the delay in detecting the hazard, the stricter the criterion (not considering the situation as hazardous enough). Egea-Capparos et al., (2016) argued that criterion bias could be the cause of differences in latency of hazard responses and results so far have supported this idea. However, this also puts into doubt the ecological validity of the hazard perception test. The level of hazardousness of the different driving contexts could influence the level of hazard acceptance. More cluttered environments where several potential or materialised hazards can occur at the same time, require our attention to focus on several targets. This increased exposure might desensitise drivers to hazards and increase their threshold bias (Lim et al., 2013).

Conversely, in this study there was no correlation between prediction accuracy and criterion ($r = -.086$, $p = .233$). However, it was found that sensitivity to hazards is a significant predictor for accuracy in multiple-choice prediction questions, but not the criterion. This result is important for two reasons. First, it is sensible that being able to correctly detect a possible hazardous situations means that one is more aware of such situations and are more likely to be able to predict how the situation is going to develop. Secondly, the fact that the criterion was not a significant predictor of accuracy shows that in this test the threshold bias is not a factor for accuracy prediction, which supports the idea that the hazard prediction test is culturally agnostic. We have observed that drivers from countries with more hazardous driving environment show a stricter criterion for reporting a situation as hazardous, meaning that criterion is influenced by cultural differences. The fact that criterion is not significantly predicting accuracy may suggest that these cultural factors do not impact accuracy. This is of a high importance considering the problems found to export the traditional hazard perception methodology overseas due to the threshold bias issues (Lim et al., 2013).

Although, the classical hazard perception test is not directly assessing a real driving behaviour during a certain situation, it still asks drivers to report whether they have seen a hazard. If the driver does not press for a particular hazard, it could mean that they did not see the hazard or that they did not consider it hazardous enough. Consequently, their behaviour might not be cautious enough. In turn, the hazard prediction test asks drivers to predict how a hazardous situation is going to develop independently of whether they consider it hazardous or not. At least, this clearly tells us whether they are aware of the situation independent of whether they considered it hazardous or not.

It has been already argued in previous chapters (see Chapter 8) that the hazard prediction test seems to be much harder than other variants of the hazard perception test. Similar results were obtained in Chapter 7 (when both the hazard perception and the hazard prediction tests were compared). The hazard prediction test appeared to be much harder but, again, successfully discriminated between experienced and novice drivers. It should be noted that in this study the scores obtained on the test were slightly lower. It might be argued that performance reached a floor effect, as the test appeared to be too hard. Therefore, a one-sample t-test was conducted to compare the mean accuracy scores of each group of drivers to chance (25%). The outcome was significant only for the experienced drivers (i.e. only the experienced drivers scored significantly higher than chance). Although, the test appeared to be hard, it still differentiated between the experienced groups, and it seems that experienced drivers were not randomly guessing the correct answer. However, for the novices, predicting the situations seemed to require a greater amount of effort. Even though, it is expected that novices will perform worse than the experienced drivers, it would be more comforting if novices could also score above chance. There could be two reasons behind novices' extremely low scores. First, during the pilot session, clips were presented only to two non-drivers and the responses were generated only by experienced drivers (to assure the plausibility of the distracters). Perhaps, showing the videos to a younger, more inexperienced population would have yielded different results. Second, the hazard prediction test was set using Qualtrics . This is a broadly used tool, suitable for creating research surveys that can be administrated online. Access to surveys created using Qualtrics is easy and quick, thus allowing a greater amount of data to be collected in a short amount of time from everywhere in the world. Since the sample in this study was comprised of Israeli drivers, Qualtrics appeared to be a good option for international testing. The only caveat was that using video clips instead of questionnaires brings some additional issues. In order to avoid participants being aware of when the video is about to stop, progress bars and controls were made invisible. Also, videos were programmed to start automatically, and stop after a precise period of time, to avoid participants clicking and stopping them. This was of high importance as participants could see each video only once, and immediately following occlusion the page would automatically lead to the multiple-choice questions. Unfortunately, it was possible to inadvertently pause the clip by clicking on the screen; as the presentation timing was independent of the actual playback of the clip, any pause in playback would mean that the test would progress to the first question before participants had seen every frame of the clip. It was not possible to control whether participants clicked on the screen and stopped the clips by accident or intention (although they received instructions to avoid doing this). Perhaps some of the participants stopped the videos and missed the precursor information. Although, it was not possible to know whether participants were clicking to stop the videos, the time each participant spent on each page was tracked. Those who *clearly* spent less than the prescribed time on the video page were removed. However, the tracked time that participants spent watching a particular clip was not sufficiently accurate to determine whether they had missed the final few frames due to a brief pause during playback. This could have been the case of learner and novices who in their attempt to predict correctly the situation may have paused the videos (gaining time) and thus missed the final couple of frames. This could also explain their low performance and the lack of differences for hazard detection.

In order to solve the above issues our research group *Transport Research in Psychology (TRiP)* created a website platform www.testmydriving.com to reach the general public and give them the opportunity to test their hazard perception and prediction skills. This website platform is the result of many years of research where a variety of hazardous videos were created and set into hazard perception and prediction tests. Participants from all over the world will be able to access these videos for free and participate in our research while practicing their skills. As the cross-cultural line of research is one of our main pillars, investing in a proper platform that will provide accurate data and allow quick and easy access, was a logical step for our research group.

This study was the last experiment of this thesis and the aim was to design a hazard prediction test suitable for the Israeli context on the basis of the findings of all previous studies contained in this thesis. As this thesis has focused mainly on investigation a suitable methodology to export overseas, applying a newly developed prediction test in a setting with such a diverse cultural profile was a big challenge. This first attempt seemed to be promising and in view of the consistent results so far (see Chapters 6, 7 and 8; Castro et al., 2014; Crundall, 2016; Jackson et al., 2009; Lim et al., 2014; Ventsislavova et al., 2016), this test could be considered as a valid option for international export. Since it has been argued that the traditional hazard perception test has poor face validity (Groeger, 2000), and with some studies finding the basic experiential effect (Wallis & Horswill, 2007; Smith et al., 2009; Wetton et al., 2010) and others failing (see Chapters 3,4 and 5; Chapman & Underwood, 1998; Sagberg & Bjørnskau, 2006; Borowsky et al., 2010), an alternative methodology was sorely needed. Certainly, the newly developed test requires more investigation as these have been the first steps into countries overseas. However, this test has proven that performance is not confounded by the threshold bias (which is of a high importance for countries with more hazardous environment where drivers are desensitized to hazards) and it requires both early detection and future prediction (which seems to be a better predictor of experience). This is the first time a hazard prediction methodology has been successfully validated across several countries.

Chapter 10

GENERAL DISCUSSION

10.1 Brief overview of the thesis

The motivation behind this thesis was to examine whether the hazard perception methodology could be validated for international export. The idea to implement the HP test in other countries was supported by the recommendations of the UN to transfer the successful practices of high-income countries to low-and-middle income countries. Furthermore, by its success in both the UK and Australia, where the test has demonstrated its ability to differentiate between those drivers who are more prone to be involved into traffic accidents and those who are collision-free (e. g. Horswill et al., 2010; Wells et al., 2008). Some studies have even demonstrated the predictive validity of this test by reporting that drivers who score higher are less prone to be involved in future traffic collisions than those who obtain lower scores (Drummond, 2000). Considering that traffic collisions are responsible for 1.3 million deaths, and are the primary cause of death for young people (WHO, 2015), it is a logical step to investigate whether the typical UK hazard perception test was fit to export to countries with higher traffic-related fatalities. To this end the traditional methodology was refined and then tested overseas. The traditional response-time methodology was found to be lacking, but a variant – the hazard prediction methodology – appeared to fare better.

In Chapter 1, the need for global road safety initiatives proposed by the UN was discussed. We are now approaching the deadline for the Decade of Action for Road Safety 2011-2020 and, unfortunately, not all of the goals have been met. Although developed countries such as the UK have managed to reduce their fatality rates up to 45% between 2006 and 2015 (UK Department for Transport, 2016), other countries found reaching the same level more challenging (e. g. countries in Asia, Africa and South America, WHO, 2015). This process could be lengthy and difficult, as even the most successful initiatives still need to be examined and adapted to each country.

This first chapter also provided an extensive literature review on hazard perception, covering its history, diagnostic validity, strengths and weaknesses (e.g. definition, methodology), and its different variants. This thesis questioned the inconsistencies in the research findings concerning hazard perception tests, and possible reasons for mixed evidence were discussed. Specific focus was given to understanding why some studies find significant differences between groups of drivers on the basis of experience or collision history (e.g. Boufous et al., 2011; Wetton et al., 2011), but others do not (e.g. Sagberg & Bjørnskau, 2006; Underwood et al., 2013). It has been suggested that the lack of consensus of a clear definition of what constitutes a hazard (or even what defines hazard perception as a skill), along with differences in methodology across the multitude of research groups, could have caused the inconsistent results (Pradhan & Crundall, 2017). As a result, it was concluded that prior to proposing the HP test for export, its transferability to other countries should be examined.

Chapter 2 focused specifically on describing the methodology applied to design the hazard perception test and the process of filming the driving footage. The decision to include a methodology chapter in this thesis was motivated by the urgent need to consider a common methodology underlying the hazard perception test. Therefore, Chapter 2 described in detail the design of the footage, stimuli, apparatus and procedure for each experiment. Far too often, research papers do not provide sufficient detail regarding the preparation and presentation of their stimuli, which results in many research groups adopting slightly different practices. This causes problems when considering mixed evidence in the literature, as it is difficult to pinpoint the reason why one test might work and another might not.

In the following section, the main findings of the experimental chapters will be summarised before discussing the theoretical and practical implications, as well as limitations and suggestions for future research. The chapter will finish with a conclusion focused on the main research question.

10.2 Summary of the main finding of each study

Overall, this thesis has been successful in its attempt to validate a variant of the hazard perception methodology for international export. This newly validated hazard prediction methodology was tested in four different countries and in each one experienced drivers outperformed novices in predicting the hazardous situations (while the traditional hazard perception methodology only discriminated between nationalities). In addition, two versions of the prediction test were compared: a free-response and a multiple-choice version. Both versions discriminated between the experienced groups, however only the multiple-choice version correlated with mileage and years of driving experience. Finally, it was concluded that the multiple-choice version of the hazard prediction test is a better option for international export as it provides multiple pragmatic advantages such as objective and quick marking, quick administration and it is time and cost-effective.

Following, the main findings of each experiment will be summarised throughout this chapter.

10.2.1 Experiment 1: Design of the hazard perception video clips. Does adding mirror information to the traditional hazard perception clips improve driving performance?

The first study of this thesis, described in Chapter 3, was motivated by the need of a robust stimulus design. This study examined the possibility of improving the traditional hazard perception design, which merely contains the forward view from a moving vehicle, by including mirror information. Several studies have included mirror information previously (Borowsky et al., 2010; Crundall et al., 2012; Shahar et al., 2010). It has been suggested that providing only forward view information creates unrealistic expectations that hazards always appear in front (Alberti et al., 2014; Allen et al., 2005; Shahar et al., 2010). Research in eye movements has also demonstrated that mirrors produce differences in response between different age groups. Young drivers were observed to use left-side and rear-view mirrors more before lane changing in comparison to older drivers. (Lavalliere et al., 2011). In turn, left and right mirrors were inspected significantly more by experienced and expert drivers in comparison to novices (Konstantopoulos et al., 2010; Underwood et al., 2002). Accordingly, the provision of information is likely to produce more realistic scanning patterns that may systematically change the relationship between a driver's level of experience and their ability to spot hazards.

Providing mirror information also allows the use of overtaking hazards. These types of hazards have been ignored by the official UK hazard perception test (as with a single forward view, there are no precursors available that allow safer drivers to response sooner than less safe drivers), although they have been identified as one of the most dangerous scenarios (Rameau et al., 2016). Overtaking hazards were considered potentially important for some international contexts, based on reported experiences filming hazard clips in Malaysia (Lim et al., 2013; 2014). Lim et al. had to remove all overtaking hazards from their (single-forward view) hazard test, which they felt unfairly represented the type and frequency of hazards.

In addition, research that has looked at the impact mirrors have on driving performance has either measured reaction times or eye movements, but not both. Experiential differences were not reported either. As a result, the research question in Experiment 1 was whether the inclusion of mirror information would enhance hazard perception performance and show differences between the driver groups based on experience (in regards to behavioural responses and eye tracking measures).

To this end, a completely new hazard perception test was developed, with new footage recorded and a new design specification. The new clips included additional rear-facing camera footage to provide side and rear-view mirror information. This allowed the use of a wider range of hazardous situations (e.g. overtaking hazards).
As no other study has assessed the impact of providing mirror information on hazard perception performance, three different types of video clips were designed and compared across experienced and novice drivers. The first type of clip was similar to those of the traditional hazard perception test (a single forward view). The second type contained a graphic overlay of a car interior, but no mirror information, while the third clip type had the three additional camera streams edited into the mirror-placeholders in the graphic overlay. While the initial premise was to compare mirror information with a traditional forward view, it considered helpful to place this mirror information within a context (i.e. the graphic overlay of the car interior). However, upon initial design of

the car overlay it was also apparent that it could make hazards feel closer to the viewer. To assess whether the addition of the car overlay itself had any impact on hazard perception performance, this intermediate condition was created.

The results suggested that, contrary to the previous literature that examined at least one of the behavioural measures (e.g. Chapman et al., 2007; Crundall & Underwood, 1998; Crundall et al., 2012; 2003), no differences were found for both eye movements and reaction time regarding driving experience or clip type. Experienced drivers did not react faster to hazards, nor fixated them quicker than the novices. Moreover, it was expected that mirrors would enhance the differences in hazard perception performance between experienced and novice drivers (as experienced drivers have a better knowledge of when to look at mirrors and what to do with the information therein). However, no differences or interactions were observed between the three conditions (even with a few specific overtaking hazards that were included). The mirror condition did not aid for faster detection of the overtaking hazards in comparison to the other two conditions. Interestingly, the car layout by itself seemed to hinder performance especially for novices. Although no significant results were found between the experienced groups, it should be noted that participants did make use of the mirrors with the rear-view and right mirrors receiving the highest amount of fixations.

This was not the first study that failed to find the experiential effect (e.g. Crundall et al., 1999; Groeger et al., 1998; Sagberg & Bjørnskau, 2006; Yeng & Wong, 2015), however it was surprising that no differences were found between the three design conditions. Without the experiential effect, it was impossible to conclude that the mirror information improved the quality of the test in differentiating between safe and less-safe drivers. Nonetheless, there was no evidence that the inclusion of mirrors hindered performance either, and it was found that drivers do make use of the information provided. For this reason, a decision was made to keep the mirror design for the studies in this thesis, as it allows the inclusion of overtaking hazards and reflects what we find in the real world.

Finally, several possibilities were discussed to address the results of Experiment 1, such as the low sample power and the probability that UK learners are too trained on the HP test as they were actively practicing it during the testing period (although not necessarily transferring these skills in real life, as novice drivers are still over-represented in crash statistics). There were less than 20 participants for each design condition (with ten or fewer per cell when divided by experience) and all were UK drivers. Some of the novices were learners, actively practicing the online versions of the hazard perception test for their driving test. In order to address these limitations, a more sensitive measure was suggested focusing on detection of hazardous precursors. Other research has supported the suggestion that precursors are more sensitive to the differences between experienced and novice drivers than the actual hazards (Crundall & Kroll, 2018).

10.2.2 Experiments 2 and 3: Does the design of the car overlay impact hazard perception performance? Comparing transparent vs. opaque overlays.

It was surprising that the results in Experiment 1 showed a lack of experiential difference. Possible explanations for these results include a low sample size and the familiarity of UK participants with the hazard perception test. However, both possibilities do not provide explanation of why mirrors did not influence response times to hazards, either slowing them due to drivers focusing elsewhere when the hazards occur, or improving them as found by Shahar et al. (2010). While drivers did make use of the mirrors, it appears their fixations on these areas were infrequent and not necessarily related to an active search for precursors. Nonetheless, the mirrors allowed the use of overtaking hazards (which these drivers were previously unaware of, and therefore did not search actively for), and possibly created a more realistic or immersive environment.

Building upon the findings of Experiment 1, Experiment 3 sought differences between experienced and novice drivers once again. In addition, Experiments 2 and 3 raised a fresh question regarding the design of the video clips. If future clips were to contain mirror information there is a question about how the graphic overlay of the car interior should be represented. In Experiment 1 a semi-transparent overlay was chosen (at least from the A-pillars upwards) in an attempt to compensate for the ability of drivers to ‘see through’ real A-pillars using stereopsis or small head movements. It was noted however that this could have created an odd impression on the viewers that may have diminished immersion, and it remained possible that a completely opaque overlay would have been superior. Certainly, several studies have reported that occluded visibility could lead to collision (Cheng et al., 2016; Summerskill & Marshall, 2015; Yashuda & Ohama, 2012) and that stereoscopic systems enhance higher task performance (Chen, Oden, Kenny & Merritt, 2010). Crundall and Kroll (2018) adopted a transparent car overlay design for their study featuring fire-appliance drivers, but they did not directly compare both types (transparent vs. opaque) of overlay with hazard perception video clips.

Thus, in Experiments 2 and 3 the newly developed design featuring mirror information was tested once more, directly addressing the potential impact of the opacity of the overlay. Experiment 2 simply required participants to watch the clips and provide ratings for the hazardousness, realism and immersion of both types of design. The following experiment (Experiment 3 in this thesis) however required novice and experienced drivers to actively search for hazards within the same clips while their eye movements, RTs and ratings were recorded. To combat power concerns regarding experience, the sample size was doubled compared to Experiment 1. First, it was hypothesised that the transparent design will be rated as more realistic and less hazardous, as it imitates stereopsis. Second, that the transparent car overlay would benefit faster hazard detection for both groups. Third, that both eye movements and reaction time would yield the basic experiential effect, which is characteristic for the hazard perception test. Finally, it was predicted that experienced drivers would fixate precursors to hazards more than the novices.

Unexpectedly, Experiment 3 did not yield significant differences between experienced and novice drivers, mirroring the results of Experiment 1. Experienced drivers did not fixate or react faster to hazards than the novices. Fixation on the car frame and dashboard were also examined. No differences were found between the types of car overlay or experience. It was hypothesised that the opacity of the overlay could influence responses to hazards. Specifically, the transparent overlay was less likely to obscure precursors. This may have improved hazard responses (at least in experienced drivers) and may have increased their hazardousness ratings. Regarding questions of realism, taken at face value, the transparent overlay may have reduced realism ratings (as A-frames are not transparent in the real world). Alternatively, the visual affordances offered by the transparent overlay may have increased the perceived behavioural realism of the clip. Equally, however, it was possible that drivers would not notice any difference between the clip and it would not impact on ratings of realism.

Regarding the mirrors, differences were observed between the groups, with experienced drivers looking significantly more at the right mirror while novices spent more time gazing at the rear-view mirror. This suggests once again that mirrors have been consulted by participants.

Even though sample size was augmented, no experiential effect was found. It has already been argued in the previous study that learners have become very well trained on the hazard perception test. In a further attempt to elicit the experiential differences, fixations on hazardous precursors were also calculated. It was expected that experienced drivers would attend more to precursors compared to the novices, even though this more subtle measure might not carry through to experiential differences in actual hazard responses. Only marginal evidence for a difference was found, with a trend for experienced drivers to spend more time fixating the precursors. Previous research supports the notion that being able to predict a future hazardous situation is evidence of advanced situation awareness (Jackson et al., 2009; McKenna et al., 2006). This means that drivers who know what to expect and are able to identify clues in the environment even prior to the hazardous situation would direct their attention and gaze towards the critical points, thus increasing the likelihood that they are looking in the right place in the visual scene when the hazard triggers. In addition, it has been argued that such hazard prediction is a purer and more robust differentiator of driver groups on the basis of safety or experience (Castro et al., 2016; Crundall, 2016; Jackson et al., 2009). The current results suggest that we came closest to finding an experiential difference when looking at precursor processing, rather than using a simple response time measure to actual hazards.

No differences were observed for hazardousness, realism and immersion ratings between the transparent and opaque car layout for both experiments. Transparent clips were not rated as more hazardous, realistic or immersive than the solid ones and these results were replicated in both experiments in Chapter 4. There were no differences in the ratings in terms of experience either. It should be noted, however that both designs were judged as being realistic, with mean ratings higher than the mid-point of the scale for both experiments. Regarding the immersion, only the transparent design was rated higher than the mid-point of the scale. In addition to this, when participants were asked which type of overlay they preferred, they opted for the opaque when they just had to watch the video clips (experiment 2) and for the transparent while actively searching for hazards (experiment 3).

Although, the results from Experiments 1, 2 and 3 did not provide enough evidence in favour of the newly proposed design, it was decided to keep this design for pragmatic reasons (e.g. mirrors allow overtaking hazards which are prevalent in other countries). Also, in Experiment 1, novices appeared to be faster (although not significantly) when fixating the hazards only during the forward view condition. This could be due to a practice effect, which overestimates the true skill of novices (Alberti et al., 2014). Considering that the validity of the hazard perception test has been questioned and that it has mainly been studied within the UK and Australia, further study was deemed necessary to test its validity overseas. Experiment 4 attempted to do this and extend the research conducted so far through the inclusion of two additional countries: China and Spain.

10.2.3 Experiment 4: Testing the Hazard perception test in three different countries: China, Spain and the UK.

The lack of experiential differences in the previous two studies questioned the validity of the hazard perception test. In Experiment 4, therefore, the aim was to study differences in performance between experienced groups in international samples. As it was suggested that UK novices might be too trained on the hazard perception test, in order to verify this, drivers from different cultural background were tested.

The hazard perception test has already been studied in many different countries, however, with wildly different design principles, and largely inconsistent results (e.g.; Horswill, et al., 2015; Lim et al., 2013; Sagberg & Bjørnskau, 2006; Wallice & Horswill 2007; Yeung & Wong, 2015,). These inconsistencies might be due to the fact that different research groups have used different methodologies, as there is no official protocol on how to develop a hazard perception test. In addition, each country has different driving contexts, with some of the hazards being potentially culturally-specific to the region. Some hazards might not be suitable for a hazard perception test as they are too abrupt and obvious which does not allow a clear discrimination between safe and unsafe drivers. Therefore, in Experiment 4 it was examined whether the hazard perception methodology could successfully discriminate between the novice and experienced drivers across three countries: China, Spain and the UK. An identical methodology was used for all three countries.

Experiential differences were expected again, although Experiments 1 and 3 suggested that it might be hardest to find in the UK sample. Unfortunately, the results revealed no experiential differences for any of the countries. Although this was surprising, there are many studies from different countries that have failed to find differences between driving groups (e.g. Norway: Sagberg & Bjørnskau, 2006; Malaysia: Lim et al., 2013; China: Yeung & Wong, 2015).

Several potential issues regarding the null results were considered. First, it was considered whether the original scoring windows were too lax, possibly allowing Type 1 errors to add noise to the results. Accordingly, the original scoring windows were revised and the data reanalysed. Results were identical to those obtained with the original scoring windows. Second, this study was submitted for publication to a scientific journal and one of the reviewers argued for a different statistical approach to ANOVA[7]. Consequently, all the data were re-analysed using multilevel logistic regression and Bayesian logistic regression, yet the results were again identical to the original analysis (see Appendix C for the results section of the published paper that uses multilevel modelling). Differences between the experienced groups were still not found. These results throw into doubt the efficacy of the hazard perception test (at least as operationalised in these studies) to evoke differences between driver groups.

There was, however, evidence that the test was sensitive to driving cultures. Chinese drivers appeared to be the slowest when detecting hazards and UK drivers the fastest. Also, Chinese drivers were observed to provide fewer responses overall in comparison to the UK and Spanish drivers. This could be due to the fact that the Chinese driving context is more hazardous than the Spanish and UK driving context (Chinese clips

[7] It should be noted that the present study paired with Study 6 was published in Accident Analysis and Prevention in 2019 featuring Multilevel Modelling Statistics (See Appendix C)

evoked the highest number of extra hazard responses) and Chinese drivers could be desensitized to the types of hazards prevalent in Spain and the UK. This result points towards the threshold bias problem suggested previously. Chinese drivers have possibly developed a higher threshold bias for reporting hazards. The criterion bias problem was also discussed by Lim et al., (2013) within the Malaysian sample. Malaysian drivers made fewer hazard responses in comparison to the UK drivers. These findings do not support the target goal of developing a culturally-agnostic test, but instead suggest that we have developed a culturally-sensitive test.

Contrary to the expectations, the basic experiential effect was not replicated even within populations that are unfamiliar with the hazard perception test. This was the third attempt to identify experiential differences and, at this point in the research, it felt that an alternative direction was required. Therefore, a new alternative methodology was developed for Experiment 5 based on the hazard prediction test. This test variant has already been investigated in different countries such as Malaysia (Lim et al., 2014), UK (Jackson et al., 2009; Crundall 2016), Spain (Castro, 2014; 2016; Gugliotta et al., 2017), and Finland (Lehtonen et al., 2016) with consistent and successful results. Due to its consistency and some indications of external validity, it seemed plausible to consider this new version as an alternative option and test whether experiential differences will be found this time.

10.2.4 Experiment 5: Free-response hazard prediction test: occlusion points

Experiment 4 concluded that the traditional hazard perception methodology is not suitable for international export, as it appeared to be culturally sensitive, rather than culturally agnostic. There was also the slight problem that the test had failed to differentiate between driver groups across three separate experiments. The hazard prediction test was considered a viable alternative due to its growing evidence based (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009; Ventsislavova et al., 2016). Furthermore, there were several issues surrounding the hazard perception test, such as the lack of robust definition of what constitutes a hazard, the fact that the traditional hazard perception test has been built on pragmatics rather than theory, and the lack of guidelines on how to define a scoring window (Pradhan & Crundall, 2017).

It has been argued that the hazard prediction test is free of threshold bias (Lim et al., 2014). The response-time methodology is replaced by accuracy of prediction where drivers are asked to predict upcoming situations independently of whether they consider such events hazardous or not. Also, this test focuses on drivers' perception of precursors to hazards, which is a likely pre-requisite for hazard detection. This means that drivers should know where to direct their attention prior to the hazardous situation, and presumably experienced drivers are better able to identify such cues (Crundall et al., 2012). It was hypothesised that this occlusion methodology is more likely to discriminate between safe and less safe drivers.

The first aim of Experiment 5 was to trial the hazard prediction methodology and also identify where the occlusion points should be set (to avoid repeating methodological issues). Occlusion points have already been investigated by Crundall (2016) and although there was not a significant difference between the different occlusion points, temporally distant occlusion points (i.e. earlier in time from the actual hazard) showed a decline in prediction accuracy, while more temporally proximal occlusion points (closer in time to the hazard) showed an increase in prediction accuracy. The aim was to test whether better results (i.e. with significant experiential effects) could be found with the same set of clips, but with the hazard occluded at the point of onset.

Following data collection and analysis, the experiential hypothesis was verified for the first time in this thesis, with significant differences found between experienced and novice drivers. Experienced drivers were more accurate than the novices in predicting the hazardous situations. The same selection of video clips was used for this study as in the previous three studies. However, when occlusion was added to them, they were able to discriminate between the driving groups. Moreover, the results found by Crundall (2016) in regards to the occlusion points were replicated. The temporally proximal occlusion point increased the possibility of prediction accuracy and the more temporally distal one degraded prediction accuracy. Although there was no interaction between experience and occlusion points (again, following Crundall, 2016), the data suggest that early occlusion might reduce the link between precursor and hazard.

The hypothesis that precursors would be better discriminators of experience than the actual hazards was supported. Experienced drivers were better able to anticipate cues from the environment, presumably because they have had many years' experience in seeking, processing and prioritising precursors (Groeger, 2000). Such experience leads to more accurate anticipation of future hazardous situations, especially when there is feature overlap between test stimuli and situations that have been previously encountered.

The decision to attempt to search for group differences with an alternative methodology was supported by the results of this study. The hazard prediction test appeared to be more sensitive to driving experience and it removed several methodological issues related to the traditional hazard perception methodology (scoring window concerns, potential criterion bias issues, etc.). Therefore, the next step was to return to China and Spain and apply the new hazard prediction test in order to assess whether it will be a better option for an international export.

10.2.5 Experiment 6: Testing the Hazard Prediction test in three different countries: China, Spain and the UK

The aim of Experiment 5 was to assess the hazard prediction methodology featuring the same hazardous clips as those in Experiments 1-4. Several important outcomes were observed in the previous study. For the first time significant differences were found between the experienced groups, with experienced drivers predicting more hazardous situations than novices. Another important point was that results pointed towards a robust methodology, as the present findings replicated those of Crundall (2016), using a completely different set of stimuli in terms of hazards and design (Crundall, 2016, used different clips without mirror information). These results suggested that the hazard prediction test could be a more feasible option for international testing.
Beyond the experiential effect, several other benefits were noted:

1. Mitigation of criterion bias - participants are asked to predict the traffic situation independently of whether they consider it hazardous or not (Lim et al., 2014). It should be noted that if the null results in Experiment 4 were due to culturally-induced criterion bias, this new test should be a better discriminator.

2. Removal of temporal scoring windows – this removed the problem of early responses falling outside a scoring window, Type 1 errors occurring inside the scoring window, and problems associated with missing response time data.
3. Language problems – Defining a 'hazard' for participants is difficult enough in the UK even after 16 years of public awareness. This problem is likely exacerbated when translating instructions into Chinese and Spanish (despite following best practice for translating). 'Hazards' no longer need to be discussed in relation to the prediction test.
4. The new mirror design allowed a greater range of hazardous situations (overtaking hazards, especially representative for the developing countries). This allows occlusion to occur when hazards move from the mirrors into the blind spot, or as they just enter the forward view. Without mirror information, these hazards would have previously had no precursors, and could not have been used in the study.
5. In addition to the prediction test showing differences between driver groups in the UK contexts, studies have replicated the effects in several other countries, supporting the export potential of this test variant (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009; Lim et al., 2014; Lehtonen et al., 2017; Ventsislavova et al., 2016).

All of the above points appeared to be essential not only from the methodological perspective, but also regarding the diversity of each driving context. The Chinese, Spanish and UK contexts differ in their hazardousness and characteristics, which requires an unbiased and homogeneous methodology (including a hazard typology). For this reason, and in order to assess the hypothesis that the hazard prediction methodology would be a better option for international testing, the test was constructed using the same hazardous clips as those in Experiment 4 across China, Spain and the UK. This would ensure that any group differences were evoked by a change in the methodology rather than a change in stimuli.

Following completion of data collection in all three countries with new cohorts of participants, the analyses reveal a clear difference in the performance accuracy of experienced and novice drivers across all countries. Furthermore, unlike the hazard perception test, there were no differences between the nationalities. This supported the evidence from smaller previous studies (e.g. Lim et al., 2014) that this test is culturally agnostic.

As with Experiment 4, this study was also subjected to Multilevel Modelling Analysis (at the request of a reviewer) and the subsequent results did not show any major differences to those analysed with the ANOVA. The new analysis supported once again that there were differences between the experienced groups, but not between the different nationalities (see Appendix C). The only difference was that the new analysis did not reveal a main effect for clip origin (which was only observed with ANOVA). The original results showed that Spanish clips were the most difficult ones to predict for all participants. The new analysis however did not find a similar pattern, although there was an interaction between clip origin and nationality, which suggested that performance was superior when the clip origin was consistent with participants' nationality. At initial glance, such results could be seen to argue against the premise that the prediction test is culturally agnostic, however this test asks participants to identify precursors in order to be able to predict a future traffic situation. It is expected that participants will be more accurate at identifying precursors in a familiar environment as they know where to look and what cues to search for (Groeger, 2000). In addition, the familiarity effect should not influence the main purpose of the test, as the differences between the experienced groups persisted.

The results from the cross-cultural hazard prediction and hazard perception tests were also compared. Hit rates obtained from the hazard perception test and accuracy from the hazard prediction test were compared in the same analysis. A main effect was found for test-type. Participants scored higher on the hazard perception test, while the hazard prediction test appeared to be much harder. The significant interaction for test-type and experience showed that only the hazard prediction test was able to discriminate between the driving groups.

At this stage, it was argued with more certainty that the hazard prediction test is a more valid option for global testing. The target was to create a test suitable for export to different cultural contexts, and the results suggested that this aim was possible. However, the actual form of the test (free-response) was not suitable for widespread automated testing. In order to address this, a multiple-choice version of the hazard prediction test was created for experiment 7.

10.2.6 Experiment 7: Validation of a multiple-choice hazard prediction test

Experiments 5 and 6 provided evidence that the hazard prediction test is a more valid option for export to other countries. However, as the intention was to adapt this measure to different driving contexts of both developed and developing countries, the present free-response format had some pragmatic disadvantages. This format is potentially sensitive to coding error and subjective bias, and does not allow automatic scoring and immediate feedback. Furthermore, the translation of typed responses was a difficult and lengthy process. For this reason, an alternative response method was explored in Experiment 7: a multiple-choice answer format. The existing literature contains a few successful reports of hazard prediction tests using multiple-choice options (Crundall & Kroll, 2018; Gugliotta et al., 2017; Isler & Cockerton, 2003; Lim et al., 2014; Petzoldt et al., 2013). However, all of the above studies followed their own guidelines, without offering a formal procedure to follow, and none of them had actually compared the performance on a multiple-choice format to that of a free-response (e.g. Jackson et al., 2009). The aim of this study was to create and validate a new multiple-choice version of the hazard prediction test, following specific guidelines and examine it across different levels of experience. It was expected that both versions would discriminate between the experienced groups, though no directional prediction was made regarding the impact of response format on group performance.

This new multiple-choice format successfully discriminated between the different levels of experience. Accuracy in prediction improved in line with increasing experience. More importantly, the test appeared to be sensitive to the experiential boundary between novice (< 1 year) and moderately-experienced drivers (> 1 year and < than six years). This finding suggested a step change in the ability to predict hazards following the first year of post-license driving experience. This ability of the multiple-choice test to detect that boundary is in line with crash statistics and the hazard perception literature. Novice drives are still overrepresented in crashes during the first year of their post-license driving and it has been argued that the hazard perception skill is still developing within the first year of experience (Foss et al., 2011, McCartt et al., 2003; Pradhan & Crundall, 2017; Williams & Tefft, 2014). At this stage, it can be argued that the hazard prediction skill follows a similar pattern of development.

Both formats were also compared and performance on the multiple-choice test was ostensibly higher (although not significantly so) and no interaction was found for experience and test. According to these results, the response format had no significant impact on the experiential differences. However, a significant correlation between accuracy and miles driven, and accuracy and years of driving was only found for the multiple-choice format, suggesting that the multiple-choice format might better tap into underlying driving experience.

This new hazard prediction test also has the potential to become a good training tool due to the possibility of immediate feedback, while providing drivers with a variety of potential outcomes to consider. While typical hazard perception tests only evoke a speeded response to a single actualised event, the prediction test encourages reflection on up to four potential outcomes from a scenario.

Regardless of the potential training benefits, evidence was found that the hazard prediction test discriminates between experienced and novice drivers once again, even when an alternative format was tested. Given the correlations between performance on the multiple-choice test with measures of exposure (miles driven, years of experience), and the pragmatic benefits offered by this methodology, this format was retained for the final study.

The next step was to apply this new version overseas within a completely different context and find out whether this version would be the final blueprint for international testing.

10.2.7 Experiment 8: The Hazard Prediction test applied to the Israeli driving context

The results of the studies conducted in this thesis showed that in order to export the hazard perception methodology, it is essential to consider the cultural driving context. The mere application of the traditional hazard perception methodology without a proper validation can create important confounds related specifically to the impact that the driving context has on hazard perception. These confounds appear related to the threshold bias as the HP test appeared to be sensitive to the cultural differences, yet not the experiential ones. It was not clear whether the lack of response to the hazards is due to the participants not perceiving these hazards or not considering them hazardous enough. For this reason, the alternative version of the hazard perception test, the hazard prediction test, seemed to be a better option for international testing as it removes the threshold bias confound.

Since the hazard perception test successfully discriminated between the experienced and novice drivers in Experiments 5, 6 and 7, the last study of this thesis aimed to apply this test to a completely different driving context. This final study intended to demonstrate that this newly developed test could be applied to a new country, where we have never conducted HP research before.

The newly developed recipe on how to create a hazard prediction test was followed and a new element was added to it: non-hazardous trials. The addition of such trials would make possible the assessment of the appraisal element by looking at both sensitivity and criterion bias with signal detection theory. In summary, the newly developed test included:

1. Mirror information, which allowed overtaking hazards.
2. Transparent car overlay which allowed visibility through the A-pillars.
3. Non-hazardous trials, which allow SDT analysis.
4. Proximal occlusion points as these were found to produce the highest overall accuracy.
5. Multiple-choice format questions due to numerous pragmatic reasons (testing, scoring, feedback) and its positive relationship with driving experience.

The context chosen for this last study was Tel Aviv, Israel. Several research groups have undertaken research with the traditional HP methodology, but with inconclusive results. Several studies failed to find differences between the experienced groups (Borowsky et al., 2010; Borowsky & Oron-Gilad, 2013; Hoffman & Rosenbloom, 2016). As a result, the aim was to find out whether the prediction test will repeat its success and differentiate between the driving groups. Participants were asked to first identify whether there was a hazard (yes or no) and then try to predict this hazard (selecting from one of four options). The SDT analysis was conducted to test sensitivity to whether there was an upcoming hazard. Results showed no differences between the experienced and novice drivers in terms of reporting hazards. However, differences were found for clip type with the non-hazardous trials being easier to identify for both groups. As the criterion bias has been found to create an important confound related to hazard perception, driver's criterion was compared. No differences were found between the groups, which at first might have appeared surprising. However, this group was not compared to other nationalities, meaning that it was only possible to compare experienced groups. Clearly, the prediction test is relatively insensitive to criterion bias, as none of the groups showed tendency to report the situations as either hazardous or non-hazardous.

The important finding was that there were differences for prediction accuracy, with experienced drivers outperforming novices. Once again, this new version managed to yield differences for experience. Furthermore, a regression was conducted with criterion and sensitivity as predictors, and accuracy as an outcome. Only sensitivity was a significant predictor for accuracy. This outcome suggested that correctly acknowledging an upcoming hazard does relate to prediction accuracy. However, the tendency of reporting everything as either hazardous or not (which is related to the threshold bias) does not. This is important, as generally the Israeli drivers showed a tendency to not report hazards, which might suggest that they also have higher threshold for hazardous evidence, similar to the Chinese drivers from Experiment 4. Although, further research is needed for more conclusive results, the finding that criterion is not related to prediction accuracy supports the notion that the prediction test is culturally agnostic. This is important, as other authors (Egea-Capparos et al., 2016) argued that criterion could be the reason behind the differences in response latency. Thus, once again, evidence was found that the prediction test is suitable for international export due to its ability to discriminate between experienced groups in different countries. This is promising as the prediction test is the first hazard test validated in several countries across the world.

10. 3 Overview of the results

10.3.1 Areas in which this thesis was successful

The principal aim of the thesis was to examine whether the traditional methodology of the hazard perception test is suitable for international testing. Such research was necessary as crash rates still represent the first cause of death for young people (WHO, 2015). International validation is especially needed for the developing countries, as they account 90% of the world's road traffic deaths. Although, the traditional hazard perception methodology failed to find the basic experiential effect within all three countries (China, Spain and even the UK), the results of this study provided a crucial pointer into a new direction. The traditional hazard perception test appeared to be culturally sensitive (as it only discriminated between nationalities) and appeared susceptible to threshold bias. Therefore, an alternative version was needed, a version which would be culturally-agnostic and able to detect differences between the experienced groups. As a result, in Experiment 5 a new test was selected based on previous successful attempts to apply this new methodology overseas (e. g. Lim et al., 2014; Castro et al., 2016). This new hazard prediction test successfully discriminated between experienced groups in three different countries, including the UK, while the hazard perception test repeatedly failed to do so in the previous three studies. It can be argued that the test showed good external validity as 300 participants were tested. The experienced drivers outperformed novices in prediction accuracy and no differences between nationalities were found this time. This suggested that this test is culturally-agnostic as its ability to differentiate between driver groups is based on the experience level and not the cultural background of the participants. Although, the first attempt, featuring the traditional hazard perception methodology was not successful, in Experiment 8 the hazard prediction test was again successful in discriminating between experienced and novice drivers in a completely different context-Israel. Thus, it can be concluded that the aim was achieved and a methodology which is able to differentiate between safe and unsafe drivers was validated.

The second aim of this thesis was to create valid methodological guidelines based on empirical testing for developing a robust hazard perception or prediction test which can be applied in different countries. Due to the lack of a theoretical basis underlying the hazard perception test it is difficult to provide a robust explanation of how the act of detecting a hazard occurs. Pradhan and Crundall (2017) argued that hazard perception is a complex process consisting of several sub-processes (hazard searching, hazard salience, hazard evidence, hazards appraisal, hazard response etc.), thus the correct term for this process should be *hazard avoidance*. The authors suggested that, to provide a better understanding of the hazard avoidance process, each one of these sub-processes should be addressed separately. Due to the lack of a general theoretical framework, different research groups have designed their own tests, often modifying the traditional methodology and these inconsistencies could have been the reason for the mixed results (e.g. Groeger, 2000; Underwood et al., 2013; Horswill & McKenna, 1994; Wetton et al., 2011).

The suggestion to isolate the different sub-processes of hazard avoidance has been followed with the hazard prediction test, as it focuses solely on the prediction sub-process, which is also known as "projection in the future" according to Endsley's Situation Awareness model (1995). In order to be able to predict a hazard prior to its materialisation, there is a need of *hazard evidence*, typically found embodied in a variety of hazardous precursors. What was learned from the results of Experiment 5 is that the relation between precursor and hazard is crucial. It was found that the amount of information shown immediately prior to occlusion appeared to be critical for participants to be able to successfully predict the hazardous situation. From a methodological perspective, the occlusion point is vital to the success of the prediction test, removing the visual scene at the point of hazard onset and testing what visual information had already been gathered (Jackson et al., 2009). Those drivers who have correctly prioritised all the potential precursors are more likely to be looking in the right place at the right time, and can therefore select the correct answer (Crundall & Kroll, 2018). Interestingly, however, Experiment 5 suggests that there is a relatively large window during which one can decide to occlude the clip and still maintain a distinction between driver groups. This means that the methodology can be adopted without developers being too concerned over the precise points of occlusion.

In order to create a more pragmatic version that could be applied in different countries, in Experiment 7 the test was converted into a multiple-choice prediction test. This version was also subjected to methodological analysis and a specific method was used to create the distracters following the guidelines of Haladyna et al., (2002). Although, previous studies have criticised the multiple-choice format (e. g. Malone & Brünken, 2015), the results of Experiment 7 showed that this more pragmatic format is still able to discriminate successfully between the driving groups. Indeed, some correlations with driver characteristics only held for the multiple-choice version. The findings also suggested that the ability to predict hazards is not fully developed until the first year of post-license driving experience, which is consistent with the hazard perception literature (e. g. McCartt et al., 2003). As a result, the second aim of this thesis was also achieved, as the results of the aforementioned studies provided an insight into the main methodological features that contribute to the validity of the test. It should be noted that the aim was not to develop a theory, but robust methodological guidelines that could be followed by others to create a test that works (please, refer back to Chapter 9, section 9.2.2 Guidelines followed to create the test).

10.3.2 Areas that need further research

A test is as only as good as the stimuli that comprise it. In experiments 1 to 3 it was hypothesised that adding mirror information to the video clips would improve realism and immersion, and potentially widen the performance gap between the experience groups. Furthermore, the inclusion of mirrors allowed a wider range of hazards, such as overtaking hazards, which are very common in developing countries. It was expected that this new mirror design would encourage better performance compared to the traditional forward view design, as it provides more environmental information and it is more realistic. However, no differences were found between the traditional design of the video clips and those containing mirror information. Mirrors did not improve or detract from hazard detection for any of the experienced groups when analysing reaction times, hit rates or fixation on hazards. This was unexpected, but looking closely at the data, mirrors seemed to be more useful for detection of precursors prior to the materialisation of the hazards and not during the actual hazards. Several studies in the field argued that precursors are better discriminators than the actual hazards (e.g. Crundall & Kroll, 2018; Crundall et al., 2012). Therefore, it was suggested that future research should look at fixations on precursors.

As learner drivers typically seek out hazard perception training (with a wealth of online and DVD-ROM training packages available), it is understandable if they now are relatively quick at responding to hazards. This may partially explain the lack of differences for reaction time responses. It was also expected that fixations to hazards would show these differences, or at least differences between the experienced groups (e.g Chapman & Underwood, 1998; Crundall et al., 2012). The differences between the driving groups when looking at fixations to the precursors were only marginal. Perhaps novices have improved in the underlying visual skills to detect the hazards that they respond to so quickly? If that were the case however, then our novices should have also been good at the hazard prediction test. Hopefully, future research will identify exactly how visual skills and hazard perception ability develop during the learning process, perhaps by following a relatively large cohort through their driving lessons, monitoring eye movements and hazard perception skill at several stages.

One further issue with the current research is that, although isolation of the prediction element of hazard avoidance is a worthy goal, this ignores many of the other processes involved in safely avoiding a collision. While the prediction test ostensibly removes or mitigates criterion bias, the notion of a hazard threshold is one worthy of further study. Though it fell outside the scope of the current thesis, it would be interesting to chart the shift in criterion in experienced drivers. This may have particular benefits for understanding situations where drivers identify a potential hazard, but then do nothing to avoid the hazard until the last moment. At such points, drivers may engage in an exaggerated display of emergency avoidance in order to demonstrate the level of danger that another road user has provoked. This socially-motivated behaviour differs from that of the rational driver that is assumed in most of the hazard perception literature, but may reflect a more realistic use of hazard perception skills on the road in some instances. In other words, good hazard perception skill may tempt drivers to behave unsafely when they perceive other drivers to cause hazards.

Finally, the focus on hazard perception skill has ignored those hazards that are created purely by the wilful violations of the driver in question. Running amber lights, driving above the speed limit, tailgating, and changing lanes without the necessary space or time to do so, all raise the risk that drivers face. These are typically addressed by questionnaires (e.g. the Driver Behavioural Questionnaire, Parker et al., 1995; the

Driving Behaviour Inventory, Glendon et al., 1993). While there have been some attempts to behaviourally assess risk taking (e.g. the Vienna Risk Test, Hergovich et al. 2007), this particular sub-field is far behind that of hazard perception in assessing behaviour in an automated and engaging way. This sub-field would benefit greatly from the involvement of hazard perception researchers.

10. 4 Theoretical implications

This thesis has provided evidence that has contributed to a better understanding of the traditional hazard perception framework and the hazard avoidance process. Although providing a theoretical explanation of the hazard perception process is outside of the scope of this thesis, the results still contribute towards the confirmation of the framework proposed by Pradhan and Crundall (2017). Pradhan and Crundall suggested that the process of hazard perception is much more complex than the mere *perception* of hazards. The whole process consists of several sub-processes that provide a clearer framework of the act of avoiding a collision. As a result, Pradhan and Crundall (2017) defined this overall process as *hazard avoidance* (see Figure 10.1 a, b).

These sub-processes have already been described in Chapter 1, therefore in this sub-section, the focus will be on discussing how the outcomes of the different studies comprising this thesis support Pradhan and Crundall’s model. The process prior to the actual hazard onset is essential. Drivers’ ability to predict a possible hazard should lead to a much faster and more appropriate response in order to avoid a collision. We use the term *respond* instead of *react,* as *responding* to hazards is a conscious process in which the driver deliberately makes their decision. In turn, *reacting* to hazards is a quick unplanned action taken, at short notice, to avoid a collision. This is possibly the crucial difference between the experienced and novice drivers. Experienced drivers are more likely to spot the precursors that lead to hazards early, due to their exposure to a variety of hazardous situations, which have been stored in memory. Successful prediction primes fast identification once the hazard is triggered, and provides the greatest time to prepare a response. On the contrary, novices have not yet had the chance to experience a wide range of hazardous scenarios and cannot rely on stored instances or templates in their long-term memory. Thus, while they may be very fast at reacting to abrupt hazards where no precursor is available (after all, youth is on their side), they are not as good at predicting hazardous situations (which requires situation awareness). This was reflected by the results in several studies in this thesis. In Experiments 1 to 4 participants were asked to react to the hazards and no differences were found between the driving groups. However, when participants were asked to predict hazards in Experiments 4, 5, 6 and 7 the experienced drivers clearly outperformed novices. These findings supported the idea that the process prior to the hazard onset (see Figure 10.1) better reflects the remaining differences between safe and unsafe UK drivers (Crundall, 2016; Crundall & Kroll, 2018). If researchers just focus on defining the process of hazard perception as it is highlighted in Figure 10.1 a, then we will miss an essential part of the hazard avoidance process which is related to the understanding of the situation (Figure 10.1.b).

According to Endsley's Situation Awareness model, the hazard prediction process is represented within Level 3, following perception (Level 1) and comprehension (Level 2). The results of this thesis support the importance of these three levels showing that drivers are able to project in the future and anticipate other drivers' behaviour (Level 3) only when cues from the surrounding environment (Level 1) have been extrapolated in a meaningful way (Level 2). In other words, a driver should know where to look for potential hazards, searching for cues/precursors that will help them predict what will happen next on the driving scene. In the case of novice drivers, they might still be able to perceive these precursors, however without yet understanding the potential danger of the situation. This has been clearly demonstrated by the results in this thesis. Only when drivers were required to spot such precursors and understand their importance in order to predict the hazardous situation, differences between experienced and novice drivers were reported. In the current hazard framework, this process is iterative and likely to be done in parallel. Once perception and comprehension lead to a prediction, this location may be visually checked more frequently, with subsequent information iterating the perception – comprehension – prediction loop. Indeed, it is likely that the driver has several predictions at any one moment of what might happen, and these must be prioritised to ensure that drivers check most frequently on the most relevant precursors.

Another important point which has provided some confusion in the hazard perception literature is that hazard perception and risk perception are sometimes discussed interchangeably (Egea-Capparos, 2012). They are however two different processes. Hazard perception refers to the objective perception of hazards (the fact that there is a car in front) and risk perception or *risk estimation* refers to the subjective estimation of the amount of danger that a certain situation poses. The confusion surrounding these two concepts has created an important confound related to the traditional method of hazard perception scoring. This confound is especially evident when the hazard perception test is applied in contexts which are too cluttered and hazardous. This type of context may contribute to cultural desensitisation to hazards. This possibility was suggested by Lim et al. (2013, 2014) when they found that Malaysian drivers showed higher threshold bias when they were asked to identify UK hazards. It was unclear whether Malaysian drivers recorded fewer responses to the pre-defined hazards due to not seeing those hazards, or because they did not consider these situations hazardous enough. Thus, while there is a clear delimitation between hazard perception and risk estimation in theory, in practice it is difficult to separate these concepts, especially when using the traditional hazard perception methodology. We learned this from the results of Experiment 4, where the Chinese drivers showed similar behaviour to that of the Malaysian drivers. They produced significantly fewer responses to all clips (especially the UK clips) in comparison to the Spanish and UK drivers. Interestingly, there were no differences between the driving groups. The fact that the hazard perception test was sensitive to the cultural differences but did not discriminate between experienced and novice drivers, supported the notion that this test is not culturally-agnostic. In order to create a valid test that could be applied in any driving context, it is necessary to assure that this test measures what it is expected to measure. With no clear definition of what hazard perception is (as a construct), it is difficult to create a valid tool. For that reason, the suggestion of Pradhan and Crundall (2017) to study every sub-process of hazard avoidance separately seems reasonable.

In order to address the issue that the traditional hazard perception methodology is culturally sensitive, a new alternative and culturally-agnostic tool was needed. In addition, this tool was also expected to be able to measure the process prior to the actual onset of the hazard, as this pre-hazard window was argued to be more discriminative than the actual hazard window. As a result, the hazard prediction test, originally developed by Jackson et al. (2009), was refined and used for Experiments 5-8. This test did not discriminate between different nationalities, but did differentiate between the different driving groups. The hazard prediction process, in which the drivers should be completely aware of the situation in order to be able to predict how it is going to develop, has been clearly isolated. This is evident by the ability of prioritising precursors which may become potential hazards. Participants are just asked to predict the hazard and show an evidence that they are aware of what is occurring, independently of whether they consider such situation as hazardous or not. They are not merely asked to react to hazards but to be aware of those even prior to their occurrence. As this ability requires higher experiential skills, it is understandable why there are differences for prediction accuracy and not for response times.

This suggests that the training that learner drivers seek out in the UK might have helped them process a hazard once it has onset, perhaps by providing a library of hazardous events in memory (and this might be why they are relatively good in terms of response times). However, they are not being explicitly trained to identify precursors or predict hazards in advance. The current results argue that, if our novices have benefitted from commercially-available training packages, this may have only benefitted post-onset processes.

There is evidence that both novice and experience drivers can benefit from HP training (Horswill, et al., 2010b). Training in HP has been widely investigated using commentary training (Castro et al., 2016; Crundall et al., 2010; Isler et al., 2008), hybrid training tests combining tutorials on HP (Meir et al., 2014), 'What happens next?' exercises with added commentaries (Wetton et al., 2013), error-based feedback training programs (Unverricht, Samuel & Yamani, 2018) and a 'Multi-Skill Program for Training Younger Drivers on Higher Cognitive Skills' (Yamani, Samuel, Knodler & Fisher, 2016) among others. The above studies reported that commentary training improves hazard perception and especially if provided in a hybrid package combined with other exercises. Furthermore, Yamani et al. (2016) demonstrated that their training program (Multi-Skill Program on Higher Cognitive Skills) improves younger drivers' performance within a short span of time. However, no coherent understanding of which is the best training approach has been offered. Therefore, and in order to know what works best, it is essential that HP training is adapted specifically to the needs of a particular driving cohort (e.g. fire-appliance HP test, Crundall & Kroll, 2018) and tests used for training purposes are previously validated (Helman, Palmer, Delmonte & Buttress, 2012).

Lastly, in Experiment 8 a clear separation was found between prediction accuracy and criterion bias. This points towards a robust internal validity of the prediction test, as it does not confound hazard appraisal with hazard prediction. As the results of this thesis have provided evidence to suggest that the prediction test shows both internal and external validity, there are enough reasons to consider it as the perfect blueprint for international testing.

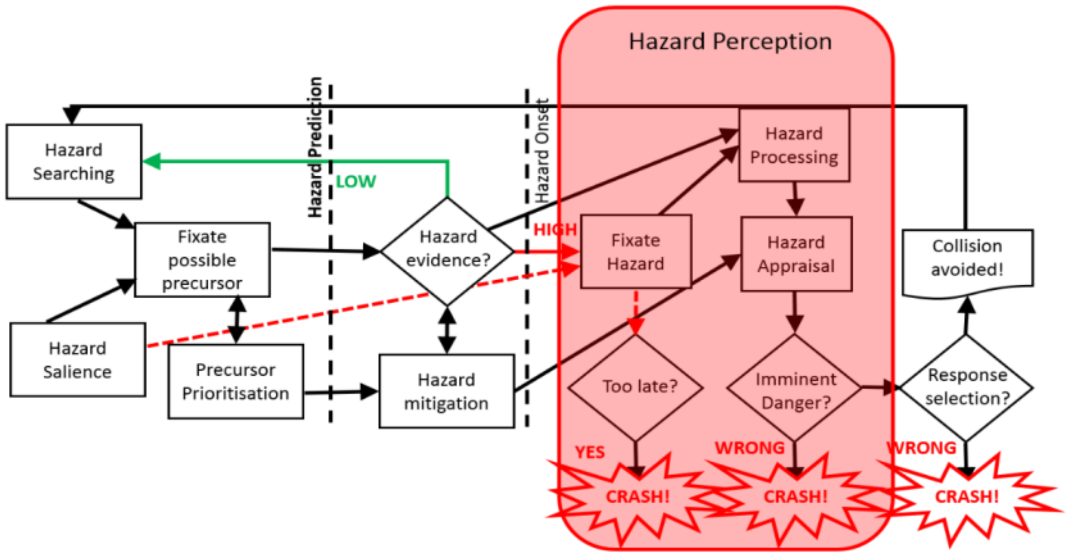


Figure 10.1 a: Sub-processes comprising the hazard avoidance process. The processes involved in the hazard perception process are highlighted in red (hazard fixation, processing and appraisal).

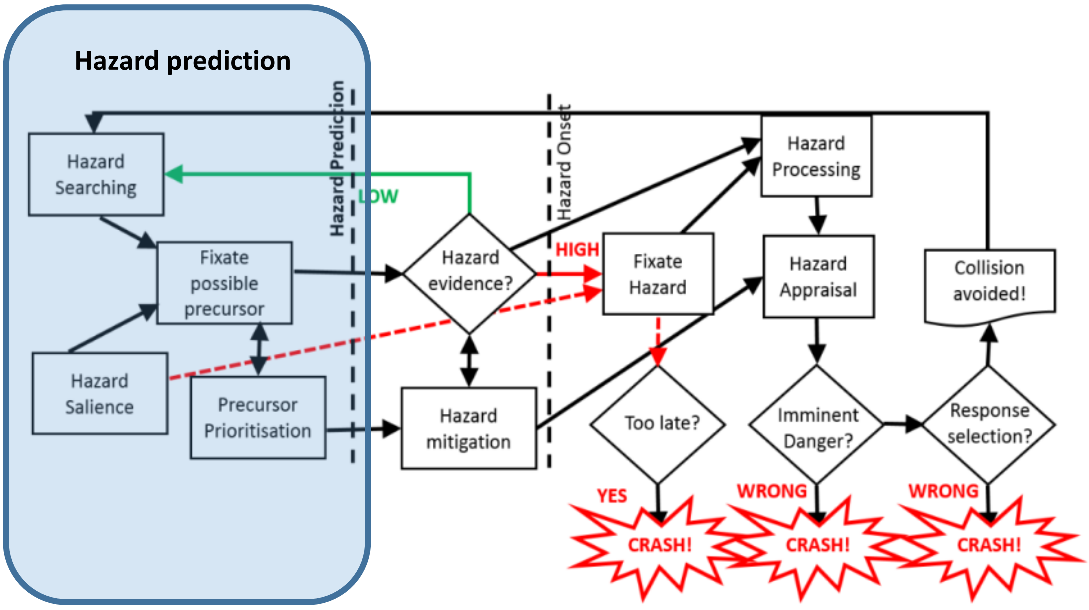


Figure 10.1 b: Sub-processes comprising the hazard avoidance process. The processes involved in the hazard prediction process are highlighted in blue (hazard searching and salience, precursor prioritisation).

10. 5 Practical implications

There are numerous practical implications related to the findings of this thesis. For instance, the newly developed hazard prediction test could serve as both a diagnostic and training tool. The results of Experiment 6 and 8 clearly showed that the test is able to differentiate between safe and unsafe drivers (at least based on experience), independently of the cultural background and driving context. This suggests that this test could be successfully applied anywhere in the world and could contribute to the licensing procedure in many countries. The results of this thesis point towards such a possibility, as the prediction test was validated in different driving contexts with different cultural backgrounds. Such relatively large-scale validation (at least from the perspective of a doctoral project) should provide at least some evidence that the prediction test is suitable to be considered as part of the official driving test.

Teamed with the traditional hazard perception test, the hazard prediction test could be very useful in terms of testing young drivers' prediction skills (which, similarly to the hazard perception skill, have also been found to develop within the first year of driving experience, according to the results of Experiment 7) and situation awareness.

This test also has great potential as a training tool. As drivers are asked to predict a hazardous situation, they might consider several possible outcomes that would enrich their hazard memory. Indeed, the multiple-choice format forces them to actively consider the possibility of four potential outcomes. Furthermore, the multiple-choice hazard prediction test allows immediate feedback of performance, and drivers could also receive guidance on where to direct their gaze in search of possible precursors even prior to the hazardous situation. Thus, drivers will learn not only to react in time to hazards, but to actively search for their precursors.

It should be noted that this test is not only intended for young drivers. Research has demonstrated that experienced and professional drivers such as fire-appliance drivers (Crundall & Kroll, 2018) or police drivers (Crundall et al., 2003) could also benefit from professional training. Currently, our research group is also developing a hazard prediction test which would evaluate the skills of professional bus drivers (in collaboration with two national operators), following the guidelines proposed in this thesis.

In addition, the pragmatic and culturally-agnostic features of the hazard prediction test will allow global training which in turn could reduce collisions around the world. We have already filmed footage in Greece (including hazards from the Greek islands) and the next step is to test the hazard prediction skills of the Greek drivers and non-Greek tourists visiting the Greek Islands. The Greek driving context could be especially challenging for tourists as, according to the European Transport Safety Council (2017), Greece has the highest number of deaths from single vehicle road collisions.

Finally, the research in this thesis opens the possibility to explore the use of modern technology and adapt the prediction test into 360-degree environments presented through VR headsets. Experiments 3 and 4 did not show that mirrors benefit hazard perception performance over a single screen, therefore a move to presenting hazards in a 360° environment could evoke more ecologically-valid driver behaviour. Our research group is currently exploring such possibilities by creating 360-degree hazard clips where drivers can search for hazards in a completely immersive environment. The aim of this project is to identify what benefits a 360° environment provides for an HP testing and training tool, above and beyond a single-screen HP test (for both video-based and CGI content).

10.5.1 International perspective

The hazard perception test has already been implemented as a part of the official driving tests of some states of Australia, Great Britain and the Netherlands (although the Netherlands only uses still images). Other countries have shown interest in the HP test (e.g. Spain, Germany, Israel, Singapore, Malaysia, Canada, Hong Kong, China, Japan and New Zealand). So why has the hazard perception test not yet been implemented as an official part of the driving test in all these countries? Adoption of such a test is not a trivial task. Important factors such as legal framework, infrastructure, alignment of the current crash data, training and education etc. need to be considered. Developing countries can find it challenging to achieve the standards of more developed countries in implementing the required changes. Also, the hazard perception test should adequately fit the licensing system of each country. The task becomes even more challenging if we think in terms of creating a generic set of hazard clips that will be suitable for different driving environments. The results of this thesis showed that each country has a specific hazardous driving environment where both legal and social rules play an important role. In order to implement a standardized hazard perception testing it will be necessary to create a specific set of clips that represent the driving environment of each country and at the same time represent a wide range of hazardous driving situations. Furthermore, it will be necessary to validate and adapt the methodology (design, instructions, format etc.) to ensure that the test will discriminate between experienced and novice drivers. This thesis has offered specific guidelines paving the way towards a standardized methodology. This process has already started as several governments (e.g. Spain, Canada, Germany, Lithuania) has explicitly shown interest in implementing the hazard perception test as part of their official driving test. Prof. David

Crundall and myself are currently collaborating with the research groups of University of Granada (Spain) and Vytautas Magnus University, (Lithuania). We are participating as international experts in their projects, funded by the Spanish and Lithuanian governments, offering our expertise and assistance following the guidelines exposed in this thesis for international hazard perception testing.

10.6 Challenges of cross-cultural research

It has been argued that culture is not relevant for risk perception and it is the characteristics of the hazard instead that impact it (e.g.Sjöberg, 2000). Some authors have been more specific pointing out that it is not culture per se, but the traffic-safety culture that impacts on risk perception (Rundmo, Granskaya & Klempre, 2012). A recent study using a driving simulator compared hazard perception performance in German and Chinese drivers supporting the idea that hazard avoidance behaviour is sensitive to the cultural background (Chinese drivers reacted slower to the hazards and time-to-collision scenarios; Wang, Cheng, Li, André & Jiang, 2019). The results of this thesis support the notion that each cultural driving context produce specific types of hazards. Both UK and Spanish drivers rated Chinese hazards as the most hazardous while all three nationalities (Chinese, UK and Spanish) rated the UK hazards as the least hazardous. It should be noted, however that this thesis only compared hazard perception performance across countries (focusing solely on the methodology) and did not delve further into examining specific differences within the cultural driving context. Future research could focus on such cultural differences.

10.6.1 How does the differences in traffic culture impact driving behaviour?

The vast majority of cross-cultural research in the driving field has been conducted in the developed countries, yet 90% of the world road fatalities occur in the developing countries (Nantulya & Reich, 2002; Peden et al., 2004; Toroyan, 2009). Even within Europe, driving behaviour in the South-Eastern part has been studied relatively less. Southern/Middle Eastern European countries report higher levels of aggressive driving and driving errors, while Western/Northern European countries report higher scores on ordinary violations (Özkan, Lajunen, Chliaoutakis, Parker, & Summala, 2006). Drivers from the Western countries also reported higher optimistic bias towards risk assessment (Bränström, Kristjansson & Ullen 2006; DeJoy 1989) in comparison to the those in low-income countries who are exposed to a greater risk (Lund & Rundmo 2009; Nordfjærn & Rundmo 2009). This trend to report results only from the Western countries could bias our understanding of driving behaviour. Driving is not independent of the traffic safety culture and the socio-economic status of the country within which it is taking place, and expecting standardised results from one country to another is unrealistic. For example, traffic regulations and their enforcement vary between low-to-middle and high income countries and the tendency to not enforce traffic regulations in low-to-middle income countries impacts on drivers' decisions regarding risk-taking (Nordfjærna, Şimşekoğlu & Rundmo, 2014).

Culture in general, and traffic safety culture in particular, impacts on protective road traffic behaviour (Nordfjærn, Jørgensen & Rundmo (2011) as risk exposure influences risk perception (Boholm, 1998). Differences in hazardousness of different driving environments influences risk priorities. Nordfjærn, Jørgensen & Rundmo (2011) compared risk perceptions, attitudes towards traffic safety and driver behaviour across multiple countries and found considerable differences in attitudes towards risk taking, especially between low-income and high-income countries. Individuals in low-income countries in Sub-Saharan Africa were more sensitive towards risk and also reported a higher willingness to take risk than those from Northern European countries who reported the safest road traffic attitudes. In terms of driving style, Southern Europe and the Middle East drivers reported more aggressive violations and errors than drivers from Western Europe (Özkan, Lajunen, Parker, Sümer, Summala, 2011). These results could be due to less developed infrastructure, less respect for traffic rules and higher levels of driver stress.

While risk perception and attitudes towards risk have been widely studied and compared across countries, only a few studies have conducted such comparisons on hazard perception performance (e.g Lim et al., 2013; 2014; Wetton et al., 2010). However, both hazard perception and hazard prediction performance have been studied independently in many different driving contexts although with mixed results (e. g.

Norwegian HP test: Sagberg & Bjørnskau, 2006; Israeli HP test: Parmet, Meir & Borowsky et al, 2014; Singaporean HP test: Yeung & Wong, 2015; German HP test: Malone & Brünken, 2015; Australian HP test: Wetton et al., 2011; UK HP test: Crundall et al., 2012). When Lim et al., compared hazard perception performance of UK and Malaysian drivers, they found that UK drivers outperformed Malaysians in detecting hazards. Similarly, UK participants were better at predicting hazardous situations when compared to the other group. However, this group also appeared more sensitive to unfamiliar hazard types. In turn, both UK and Malaysian drivers reported similar visual strategies contrary to the behavioural results which suggest that Malaysian drivers might require a higher threshold to consider a situation as hazardous. The Malaysian driving environment is more hazardous overall and this may have impacted the sensitivity of the Malaysian drivers.

Wetton et al., (2010) also compared performance on both UK and Australian driving footage but did not find differences for hazard origin (although they only compared performance of Australian novice/experienced drivers). Their results suggested generalisability between countries which are highly similar in terms of culture and road laws. Yet, studies conducted in developing countries (Malaysia, China) have found such transferability more challenging.

10.6.2 Validation of a methodology instead of a theoretical construct

The majority of cross-cultural studies in the driving field have applied self-report measures to test and validate different theoretical constructs such as attitudes towards risk, driver distraction, sensation seeking etc. One of the most popular tools that measures crash-related behaviours is the Driver Behaviour Questionnaire (DBQ) which has been validated in many countries such as Finland, Great Britain, Greece, Iran, The Netherlands and Turkey (Özkan, Lajunen, Chliaoutakis, Parker & Summala, 2006); Bulgaria, Romania and Serbia (Stanojević, Lajunen, Jovanović, Sârbescu, Kostadinov, 2018) and across samples of young drivers in Ireland and Finland (Mattsson, Fearghal, Lajunen, Gormley, Summala, 2015) to name but a few. The DBQ measures crash-related behaviour, looking at intentional violations and cognitive failures which are further categorised in other sub-behaviours such as slips and lapses or those that involve aggressions towards other road users. The validation of such measures typically involves dismantling a certain theoretical construct into sub-factors in order to create a structural model that can be related to a certain behaviour. Relationships between factors and whether the grouping of these factors can predict a certain behaviour are further explored and confirmed. This is not a trivial task and multiple studies have been conducted in order to validate the most stable factor solution specifically for the DBQ (e.g. Rowe, Roman, McKenna, Barker & Poulter, 2015; Özkan, Lajunen, & Summala, 2006; Warner, Özkan, Lajunen, & Tzamalouka, 2011). Furthermore, additional factors such as gender, socio-economic status and cultural differences have been found to impact on the factor structure of the DBQ (e.g. Blockey & Hartley, 1995; Reason, Manstead, Stradling, Baxter & Campbell 1990; Özkan, Lajunen, Chliaoutakis, Parker, & Summala, 2006).

Contrary to the DBQ, the hazard perception test has been built on pragmatics and has not been validated in terms of factor structure. Neither is there a valid theory that explains the act of perceiving hazards to aid the development of a theoretical construct. Hazard perception has been widely studied but none of the studies has offered a coherent theoretical model. There have been attempts to link HP to Endsley's SA model, though this has not lead to a deeper understanding of the underlying sub-processes. This theoretical caveat has not been an obstacle for obtaining valid results however. Many studies have demonstrated that the underlying methodology (the RT paradigm) of the HP test is able to successfully discriminate between experienced and novice drivers (Horwsill & McKenna, 2004; Wallis & Horswill, 2007). However, the attempts to export this same methodology in different countries have reported mixed results (e.g. Sagberg & Bjørnskau, 2006).

The focus of this PhD thesis was also primarily based on the effectiveness of the methodology. Therefore, when HP cultural sensitivity has been discussed throughout the cross-cultural studies, each point has only referred to the methodology and not the theoretical construct. This PhD sought to validate a paradigm that will be able to discriminate specifically between experienced and novice drivers regardless of country. The results showed that the cultural sensitivity of the HP test may mask the experiential differences that might otherwise be found. Hence, an RT paradigm might not be suitable for international export. As a result, a second attempt was made to validate an alternative methodology that appeared to be culturally agnostic and successfully discriminated between different levels of experience.

10.6.3 The challenges of cultural equivalence

Cross-cultural validation represents a challenge for researchers but it is highly necessary for providing important answers to the differences in the outcomes between different countries. Flaherty et al. (1988) proposed five major dimensions for cross-cultural equivalence of new instruments (a) content equivalence (b) semantic equivalence (c) technical equivalence (d) conceptual equivalence, and (e) criterion equivalence.

10.6.3.1 Content equivalence

Content equivalence refers to the challenge of ensuring that people from different cultures are presented with equivalent content. When considering the translation of questionnaire items, this primarily refers to ensuring that participants in different countries are presented with identical survey items. The difficulty in the current thesis is that we are capturing naturalistic hazards, so it is impossible to predetermine content equivalence in this strict way. However, the primary aim of the hazard perception test is to assess drivers' abilities to spot hazards that are relevant to the potential collisions they might face in their countries. This allows for the possibility that hazards might differ across countries in considerably ways, while still arguing for content equivalence at a functional level.

Nonetheless, there were clear overlaps in hazard content across countries. Vehicles emerging from side roads and errant pedestrians were two hazards common to all three countries. Others were more specific however. For instance, motorcycles behave with more degrees of freedom in China. Providing these hazards were representative of real driving in the country, these were consider to represent a valid hazard.

10.6.3.2 Semantic equivalence

The second challenge is the language or semantic equivalence of the test content. If we consider this in terms of questionnaire items, then it refers to the requirement that not only should items contain the same content, but that content should be understood to be equivalent. For instance, if one English questionnaire item contains an idiom, the content may be translated perfectly, but the meaning may not as the readers from the other culture may not recognise the idiom. To place this within a driving context, flashing headlights might indicate that another driving is willing to let you out of a junction in the UK. Flashing headlights might mean something completely different in China. Again, however, one can argue that the behaviours captured in these naturalistic clips represent culturally-specific behaviour that should still allow novice and experienced drivers to be discriminated, at least when testing participants with experience of that country. If Chinese drivers had not understood the actions of some drivers in the UK clips, this would not detract from the value of the test if Chinese experienced and novice drivers were still differentiated on Chinese clips.

While content and sematic equivalence are interesting concepts in relation to the hazard clips, the instructions also have to have both equivalences. It is vital that participants undertake the test in their own language as it has been observed that they tend to subconsciously adapt their answers to fit with the culture associated with the language (Harzing, 2005), but most importantly, many Chinese participants did not speak any English. An important aspect of adapting a study for a new language process is forward-backward translation as instructions and items in the source and target language should have equivalent meanings. During the translation of the instructions special emphasis was made on the definition of a hazard and a hazardous situation. It has already been discussed in Chapter 1 that the lack of a coherent definition of what a hazard is (even in the original language) may create important confounds. Notably, even when providing a generally accepted definition of a hazard, participants still adapt this definition to their own cultural representation of what constitutes a hazard. Therefore, an expert panel consisting of bilingual translators and experts of the local culture were involved in the process. Despite all the efforts, it cannot be guaranteed that participants have had an unbiased representation of what constitutes a hazard.

Finally, some translation problems may arise when there is not an equivalent translation for a specific word from one language to another. An illustrative example is the word “undertake”. Only the English language has a specific word for undertaking. Chinese and Spanish drivers would just use the word ‘overtaking’, though they might preface this with the word ‘illegal’ when referring to (what UK drivers would call) undertaking. Videos from all four countries contained undertaking hazards, which required the phrase ‘illegal overtaking’ in the other languages. This is a case where semantic equivalency necessitates a reduction in content equivalency.

10.6.3.3 Technical equivalence and conceptual equivalence

The cross-cultural studies in this thesis were specifically concerned with testing the ability of the HP test to achieve comparable results in different cultures. Comparisons of hazard tests from around the world is difficult, because they have been constructed and presented in different ways. The current thesis employed the same technical protocol across all countries to ensure that, though the content might differ, the test remained the same in all other ways (e.g. car-overlay, mirror information, clip length, etc.).

Conceptual equivalence is a prerequisite for cross-cultural comparison, however in this case, we have not focused on the theoretical construct but on validating a methodology that captures the HP skill in an unbiased way. It is difficult to claim conceptual equivalence when there is no accepted theory of hazard perception. As this thesis has focused upon assessing and refining the methodologies, it perhaps makes more sense to talk in terms of criterion equivalence (i.e. the equivalence of the recorded measures) rather than conceptual equivalence.

10.6.3.4. Criterion equivalence

Criterion equivalence required that the interpretation of the measures remains identical across cultural groups. A valid measure should report equivalent results between different cohorts independently of their origin or characteristics. The main problem that has been encountered in this thesis was the sensitivity of the traditional

HP methodology towards cultural differences which unfortunately masked the experiential effect. The response time measure does not have criterion equivalence across the cultures (as systematically-biased hazard perception thresholds may differ across cultures, which then confound HP RTs). The prediction methodology focused on a purer measure (hazard prediction), which therefore allowed for greater criterion equivalence across cultures.

10.6.3.5. Sample equivalence

While not one of the five equivalencies mentioned by Flaherty (1988), sample equivalence proved to be an important factor that required consideration from the initial stages of the research. Participants should be representative of their culture but at the same time equivalent on non-cultural demographic variables. In driving research, years of driving experience and mileage are used as proxy for different levels of HP practice. All four countries examined in this thesis (China, Spain, UK and Israel) provide a unique traffic culture and different levels of safety. The World Health Organisation (2015) estimates the road fatalities of China, Spain, the UK and Israel to be 18.8, 3.7, 2.9 and 4.2 deaths per 100,000 population, respectively. The difference between the safety records of the UK, Spain and Israel is considerably less when compared to the Chinese safety record. China has experienced a significant increase in motor vehicles and road construction in the last three decades. These abrupt changes in infrastructure have negatively correlated with the adaptation of the road users to the new conditions. Although the safety record has slightly improved, China still reports high accident rates usually between vehicle and pedestrian (Report on the Development of Road Traffic Safety, 2017). In turn, the UK is one of the top ten safest countries in the world where the accessibility to modern vehicles, respect for rules and safe infrastructure has helped to reach low accident rates and a safe driving environment. Finally, Spain and Israel are the closest to each other in terms of accident rates, although Spain reports a slightly lower crash index. While climate and infrastructure could be quite similar, Israel is under tension circumstances caused by the possibility of terror attacks which bring additional risk and affects the driving context. Another interesting fact is that many ultraorthodox Jewish communities restrict driving for young men who have yet to marry. However, these young drivers still opt for a driving licence although hiding it from their parents and community. Such circumstances do not permit a sufficient driving exposure and create additional socio-cultural tension. China reports a similar problem for driving exposure (although for different reasons). With over 300 million registered vehicles, the country faces a considerable congestion and pollution problems which required the implementation of policies such as the even-odd license plate policy or end-number policy (allowing cars to drive only on certain days). Beijing has also taken new measures to control the number of cars without local licence. Many Chinese drivers wait for years to obtain a licence plate not being allowed to drive their own car. These policies make a breach between China and the other three countries in terms of driving frequency and experience as many drivers had their driving licencing for years but have not had the sufficient driving exposure.

Such differences in the traffic environment and culture have required considerable effort in terms of data equivalence. The first important decision was related to the experience criteria. The above examples clearly show that having a driving licence for a few years does not necessarily reflect driving exposure. Therefore, average mileage driven in the last two years was considered an important factor for deciding the level of experience together with the years since passing the driving test (please, refer back to Chapter 5, point 5.3.1 for more detailed description). Equivalence in procedures of collecting data in cross-cultural studies is necessary although most of these studies are non-equivalent to some degree. Recruitment strategies such as convenient sample and random population with no inclusion criteria identified prior to the selection, are often used. Access to a large cohort of drivers with different levels of experience appeared to be easier in some countries than others. For example, learner drivers in Spain could be accessed via specific driving schools where they prepare for their theory test. This is not the case for the UK or China where participants have been recruited via online platforms or in occasions via snowball sampling. To ensure equivalent and representative samples at least in terms of driving exposure, education, social status and sex have not been controlled (although all the participants had basic education). In turn, age appeared to be a tricky variable due to the differences in minimum age in which a person may obtain a licence. In Both China and Spain the minimum age is 18, while in UK and Israel is 17 and 16 years old, respectively. This explains to some extend why some of the experienced groups had equivalent age to some of the novices. It should be noted that some of the Chinese drivers who were classed as novices due to the lack of driving exposure, were age 20 or above.

In conclusion, a hazard prediction paradigm that discriminates between experienced and novice drivers has been successfully validated although these initial studies are just a first step towards a series of studies that seek to test the transferability of the hazard perception test and more specifically the hazard perception skill across different countries. This thesis has only validated a paradigm that test a specific sub-process of the hazard perception skill-hazard prediction and results seems promising. It should be noted, however that this thesis only compared hazard perception and prediction performance across countries (focusing solely on the methodology) and did not delve further into examining specific differences within the cultural driving context. Future research should focus on such cultural differences.

10. 7 Limitations

10.7.1 Stimuli design

Although, the studies conducted in this thesis have provided some insights into the sub-processes of hazard perception, the studies were primarily applied in nature. The hazard perception test has been built on pragmatics, and it became clear that the lack of theoretical framework has prevented any substantial advances. Unfortunately, testing particular theories was outside the scope of this thesis. The driving force behind this thesis was equally pragmatic: to create a test which will show good validity regardless of cultural context. In order to create a realistic environment and stimuli that will reflect real world situations, all footage contained non-staged hazards. It has already been argued that staged hazards could be biased due to the pre-selection of these situations by experienced researchers or instructors (Crundall et al., 2003). Nonetheless, even naturally occurring hazards have to be selected from the raw footage, and this was typically done with a small group of traffic psychologists. It is possible that bias crept into the selection process. Therefore, a qualitative approach such as focus group (comprised of participants with different levels of driving experience) could be very beneficial in terms of selecting the final set of stimuli. Such qualitative approach will allow discussions where novel aspects of stimuli design and hazardous situations can emerge as well as deeper understating of the existing issues.

In order to make clips as realistic as possible, mirror information was included to the forward view footage with a semi-transparent car overlay. The decision to keep this design for the rest of the experiments was based only on the initial hypothesis that hazard perception will improve. Experiments 1 to 3 failed to provide such evidence and this design was not found to enhance hazard perception performance. Further research is needed for more conclusive results, as there are very few studies that have looked at clip design, and what research does exist has not been directly related to hazard perception videos (Cheng et al., 2016; Yashuda & Ohama, 2012) or differences between experienced and novice drivers (Shahar et al., 2010). Also, it is of a high importance to gain a better understating of why such realistic design was not encouraging better hazard perception performance or greater differences between the driving groups. It has already been argued that the hazard perception ability could be dependent on the type of hazards (Crundall, et al., 2012) and it has been found that potential (or latent) hazards are the most difficult to identify (Borrowsky et al., 2010). Therefore, further research will add valuable knowledge in relation to the stimuli comprising the hazard perception test.

10.7.2 Sampling limitations

For Experiment 1, 57 participants were recruited (approx. 20 participants in each of the three conditions) and for Experiment 3, 42 participants took part). In Experiment 4 however, the sample was augmented to 150 participants (though still only N=25 per group). The aim of Experiment 4 however, was not to test stimuli design, but experiential differences, yet these were not found either.

Alternatively, it could be the case that the experiential gap was not large enough, as in Experiment 2 experienced drivers had an average of three years of driving experience. Although, it has been argued that the critical crash period is within the first 12 months of license (Deery, 1999; Foss et al., 2011, McCartt et al., 2003; McKnight

& McKnight, 2003; Williams & Tefft, 2014), a bigger experiential gap might have yielded significant results. It should be noted that in some experiments the average age of the experienced groups is equivalent to those of the inexperienced drivers in other experiments. Therefore, driving groups have been divided according to years of experience and mileage. Specifically, in China there were drivers who had passed their test some years ago and have not had exposure to driving to improve their skills. An older driver who has passed their test a few years ago without exposure cannot be classed as experienced drivers, therefore, a common sense approach was to consider driving years and mileage than age. Furthermore, McCartt et al., (2009) concluded after an extensive review that the effect of experience was stronger than the effect of age (and crash risk was significantly higher during the first 500 miles driven but did not affect crash likelihood after the first 3500 miles).

Another important point worth mentioning is the difficulty to homogenize samples with different cultural background. Differences are related to both the licensing procedure and cultural background. With such a great number of legal and social differences, it is not easy to classify drivers using identical criteria. For instance, in China, novice drivers could wait many years until they obtain permission to own a car, even though they have obtained a driving licence. 'Five years' experience' since the driving test is unlikely to reflect five years of every-day driving. This could create an important confound if not considered in advance. Also, only the UK drivers are familiar with the hazard perception test which could have created an important confound related to their performance.

10.7.3 Qualtrics software

It has already been demonstrated the importance the occlusion points have for prediction accuracy. This became evident with Crundall's study in 2016 and was supported by the results in Experiment 5 of this thesis. The type of occlusion point is related to the amount of information we receive related to the precursors. Being able to correctly identify a precursor is the only way to successfully predict a future hazardous situation.

The data in Experiment 8 was collected via Qualtrics which is a widely used online platform for collecting survey data. The use of this online tool to show videos to participants and test their prediction accuracy introduced certain noise in the results. Clips could be accidentally paused by participants should they inadvertently click on the screen during playback, and the page would automatically transfer to the next page which contained the multiple-choice question after the duration of the video. Unfortunately, this would mean that they would not see the final frames of the clip and thus be handicapped in identifying the precursor to the hazard. During the data cleaning process, the amount of time each participant spent viewing each clip was assessed, and those trials where the duration was obviously shorter than the length of each particular hazard clip were removed. However, it is possible that a small number of clips were terminated when the video had almost played through. In such events, it would have been difficult to distinguish these errors due to the accuracy of the timings provided by Qualtrics.

Nonetheless, differences between the groups were still found which argues in support of participants seeing the whole video. A floor effect was reached however, and it was not clear whether this effect was due the MC-test being too hard or to the above-mentioned problems. Fortunately, we are currently developing our own online testing portal (testmydriving.com) and we are exploring the option of desktop apps to deliver training and testing tools.

10.8 Reflections on the research process

A perfect PhD thesis is an unlikely outcome of three years of research. Neither is it particularly desirable. I have learned more from my mistakes than I have from my successes, and without those errors I would not have developed as much as I have. It is important to reflect on those difficult, ambiguous, or unsuccessful parts of the research process, and to identify those gaps that I would rectify if I had three more years of study. This section details some of these reflections.

10.8.1 Was the hazard perception test a straw-man hypothesis?

It has been emphasised on many occasions throughout this thesis how challenging it is to create a hazard perception test that would provide consistent and ecologically valid results. Although in Chapter 1 important flaws of the traditional HP methodology were discussed, the early experiments of this thesis still opted for this methodology. As the final blueprint proposed for international export was not the traditional hazard perception methodology, but a variant –hazard prediction- it might not have seemed logical to insist on the traditional methodology during the early experiments. The progression from experiments 1 to 4 might have not appeared smooth in terms of incremental changes in the results but the clear and logical sequence from experiment 5 onwards reflected the maturation of the research of this PhD thesis. The early studies were undertaken with a view to refining the methodology, and they certainly assisted in the development of the later, more coherent approach.

Indeed, the first three experiments that applied RTs could be considered as 'straw-man' hypotheses, as there is literature to suggest that prediction tests are good alternatives to traditional HP tests. However there is a greater body of literature in support of the traditional methodology using RTs. While there was obvious suspicion that the RT methodology might not work, I entered the research process with an open mind. Given the innovations that I added to the hazard materials (mirror information, the car overlay) it was always possible that RTs could have produced the sought-after effect. Furthermore, the traditional HP methodology is still the dominant measure used in the HP research and it seemed an obvious place to start.

Though one study did exist to suggest prediction would be a better test of cross-cultural hazard awareness than a traditional HP test (Lim et al., 2013), it was not sensible to base a thesis on the conclusions of a single study. Further research was deemed necessary where RTs have been approached in an unbiased and agnostic manner. While it might seem that the HP test was set up as a straw-main hypothesis, this was not the intention and I made every effort to find an effect with the typical methodology.

10.8.2 Alternative explanations for the null effects associated with the hazard perception methodology

We still do not know the exact reason why the traditional methodology reports mixed results (and this question was outside the scope of this thesis). A possibility is that not all of the hazards selected for the final test were good at discriminating between the experienced groups. It could be argued that I should have performed a clip-by-clip analysis for experiments 1 and 2 to identify which clips were better at discriminating between the driver groups. Nonetheless, the positive effects of experiments 5, 6, and 7 have shown that the same clips (with the addition of an occlusion) can produce an effective test. Retrospective clip-by-clip analyses are therefore redundant.

One analysis on a subset of clips was conducted in experiment 4 to isolate overtaking hazards, given the novelty of these situations. This analysis does not provide enough insight in identifying specific hazard features that impact the discriminability of the clips and therefore was not included in the main hypotheses. Nonetheless, clip-by-clip analyses will certainly feature in my future research.

On a related note, clip selection is still open to bias even with naturalistic footage. It is not a trivial task to select valid hazardous situations when no consensus has been reached about what constitutes a hazard. The definition followed for hazard selection for all of the studies in this thesis is "*an object or event that would require an evasive manoeuvre, including sudden braking or swerving to avoid a potential collision*" Hazards were classed as something that encroached into your path and had the possibility of a collision if you do not change your driving behaviour (e.g. braking). This definition was selected as it does not leave space for ambiguity in terms of what the driver should be monitoring for. However, the original definition by McKenna and Crick (1991) was quite different defining hazards as "potentially threatening events". It is debatable which definition is better, as potential hazards might not require an evasive manoeuvre (potential hazards have often been defined as quasi-hazards) and might not be considered as hazards by those participants coming from highly hazardous driving contexts. Although, the above definition was strictly followed, this does not provide a guarantee that the process was free of biased selection that benefitted experienced drivers. Nonetheless, at least clips reflected realistic situations from the real world.

Another important point is the experience gap between the cohorts of novices and experienced drivers selected for this thesis. Some of the studies in this thesis classed drivers with an average of three years of driving experience as experienced drivers (e.g. experiments 3 and 5). This cohort was particularly young with an average age of 21.5 which was equivalent of the age of the inexperienced drivers in experiment 1 and the UK inexperienced drivers in experiment 4. This might have affected the results of the above experiments as compared to the samples of other studies, the experiential gap is not very large. As an example, Wallis and Horswill (2007) selected experienced drivers with an average of ten years and 8000 km while their novices were defined as those who had been driving for four years or less on a provisional or open license; Crundall et al. (2012) reported an experienced group with mean driving experience of 16.4 years (M=33.0 years old) and a learner group with mean years of driving experience of 1.5 months (M=20.3 years old); Wetton et al. (2010) assessed novices with less than three years of driving experience, mean km of 7267.29 (m=21.25 years old) and experienced drivers had an average of ten years of driving experience, mean km of 13.043 (M=40.58 years old).[8] The above experiential gaps could be considered equivalent to those in experiments 1 (see Chapter 3, 3.2.1 Participants); experiment 4 (see Chapter 5, Table 5.1); experiment 6 (see Chapter7, Table7.1); experiment 7 (see Chapter 8, 8.3.1 *Participants*) and experiment 8 (see Chapter 9, 9.3.1 Participants). However, the experiential gap in experiments 3 was much lower which might have contributed to the null results. But if this was the case, experiments 1 and 4 should have reported experiential differences and experiment 5 should have failed to find differences between the driver groups. Other studies have found differences for hazard prediction with much smaller gaps between participant groups. Crundall (2016) had less than three years gap between his driver groups.

When considering the samples from other countries, I was faced with a problem. In China, the number of years since passing a driving test does not have as direct a relationship with driving experience as it might in other countries. Restrictive laws and the high costs associated with driving mean that ostensibly experienced drivers (who

[8] Note that not all of the selected references reported age, however this selection of studies was random in order to avoid bias.

passed a test some years ago) may have had little actual exposure. For this reason it was necessary to incorporate mileage into the definition of experience.

Finally, the definition of the scoring windows is another controversial topic within the hazard perception literature, and may contribute to the generation of null effects. There are no official guidelines on how to define a hazard window or when onset and offset should occur. Typically, the onset starts at the moment the potential hazard begins to materialise while the offset is defined just prior to the moment when an evasive manoeuvre will no longer avoid a collision. However, these definitions are still vague (each hazard is different) and have not been officially validated (as with the occlusion points for the prediction paradigm). The possibility of inadequately defined scoring windows was considered in experiment 4. More conservative scoring windows were chosen by removing a certain number of frames for both the onset and offset. However, each hazard had to be considered individually, therefore a fixed numbers of frames could not be applied across clips. Although no differences were reported between the results with the new scoring windows, it cannot be guaranteed that more substantial changes to the hazard window would not produce significant effects in the early studies. The lack of clear guidelines on how to define a hazard window that would allow a standardised methodological approach and more consistent results clearly affects research in hazard perception. For that reason, a new prediction paradigm was proposed in experiment 5.

10.8.3 Does HP practice mean that learners are better at HP tests than experienced drivers?

Throughout chapters 3 and 4, I speculated that a possible reason for the null results within the UK sample was that learners could be now too trained on the hazard perception test. This has only been discussed as a possibility and has not been measured in this thesis. UK learner drivers were likely engaged in active hazard perception practice during the period within which I was collecting data, and may therefore have performed extremely well. This possibility needs further investigation as to the author's knowledge there are no studies that have attempted to assess naturally occurring practice levels with hazard perception tests.

Most of the experienced UK drivers who took part in the different experiments of this thesis are young enough to have also been required to undertake the hazard perception test before obtaining their licenses. Despite this it is possible that their explicit practice and training benefits waned over time, while the learners and novices retained benefits from their more recent exposure. Similarly, test familiarity may have possibly provided an advantage for UK participants when compared to drivers from other countries, as UK drivers tend to demonstrate better performance on the HP test (e.g. Lim et al., 2013; Experiment 4 of this thesis).

In summary, practice effects have only been offered as speculation, and this possibility has not been tested, but neither have I relied on it to fully explain the lack of differences. Future research should look at this possibility.

10.8.4 Size effects and marginal differences

Even though the failure to find differences for HP accuracy between the experienced groups is not without a precedent, the vast majority of HP literature supports differences between experienced and novice drivers. Most of these studies have reported a medium to large effect sizes for experienced drivers outperforming novices on the HP test (for reference, Isler et al., 2009; Johnston & Scialfa, 2016; Vlakveld 2014; Wetton et al., 2010; Wetton et al., 2011). Regarding hazard prediction, literature is still growing but with consistent results and many studies reporting also medium to large size effect (for reference, Castro et al., 2016; Jackson et al., 2009; Ventsislavova et al., 2016). Similar effect sizes (medium to large) were found for differences in experience in experiment 5 (η_p^2=0.11), experiment 7 (η_p^2=0.3) and experiment 8 (Cohen's *d*=0.6), with the exception of experiment 6 where the effect size was small (η_p^2=.05).

Another statistical concern were marginal differences. Although, it has been argued against describing non-significant effects as marginally significant due to risk of false-positives (see Olsson-Collentine et al., 2019; Pritschet et al., 2016) two marginally significant results have been reported in this thesis. The first one has been reported in experiment 3 (please, refer back to *4.4.2.1.4 Precursor analyses*, page 121) where a marginal difference has been found between the experienced group for fixations on precursors (p=.07). There was 'marginal evidence' that experienced drivers fixated precursors more than the novices. This marginal significance was reported as it was the first time there was some evidence that precursors could discriminate better than RTs between the experienced groups in this thesis. The consistent failure to find experiential differences in experiments 1 and 3 did not provide clarity in terms of the quality of the hazards. However, this marginal effect coupled with some other significant effects such as spread of horizontal search at least suggested that the clips contained qualities that are sensitive to driving experience. In defence of 'marginal significances', the formative stages of a thesis need to identify the most appropriate avenues for further study, and these hints at significance can be very useful in this regard. This was later borne out as the prediction paradigm appeared to be a better discriminator between experienced and novice drivers which suggests that this initial marginal evidence was pointing in the correct direction.

The second marginal significance reported referred to the comparison between the free-response and multiple-choice prediction test (please, refer back to 8.5 Comparison of the multiple-choice test with the free-response test, page 202). The performance on the multiple-choice was marginally higher (p=.07), but coupled with the significant correlation with annual mileage and years of driving (post-license), I concluded that this was the better format. At all times I have attempted to be clear about the marginal nature of these effects, and have used them to point towards future directions for research.

10.8.5 Contribution to theory

The reason behind the mixed evidence surrounding RT paradigm is still unknown, but what we did learn from the results of this thesis is that the prediction methodology is a more robust discriminator of experience. The skill to identify precursors in order to be able to predict how a hazardous driving situation is going to develop clearly discriminates between experienced and novice drivers. This result has been consistent through different cultural driving contexts, different sets of hazards and different levels of driver experience. The prediction task only requires drivers to predict what happens next without involving hazard or skill appraisal, thus permitting the assessment of a very specific sub-process (part of the broader hazard avoidance process). Assessing each sub-process separately could end the uncertainty surrounding the mixed evidence. Perhaps, this is the reason behind the inconsistent results: the attempt to assess a complex skill (hazard *perception*) with a blunt measure (RTs) may lead to multiple confounds. Instead, perhaps we are required to isolate each sub-process involved. Specific tasks could be designed for this end. A reductionist task that involves monitoring precursors and understanding how drivers classify hazards can help create a battery of tasks that measure the process of choosing an appropriate response to different hazard types (e.g. latent hazards, environmental hazards etc.) or even risk-taking response to overtaking hazards, amber lights etc. However, this battery of tasks will be limited by the available resources and learner drivers' willingness to engage in such a lengthy testing schedule. Furthermore, the cross-cultural studies highlighted the importance of considering separately each driving culture and environment. Hazards, and both legal and social norms, could be very different from culture to culture which clearly impacts on the types of hazards that could be found. In order to better understand each driving environment and the behaviour of the road users, a qualitative approach that looks into the type of hazards and whether these are representative of the cultural driving context, could be very beneficial. For example, a focus group where each hazard could be discussed individually would provide a more in-depth understating of its characteristics.

It should be noted, however, that each task or tool should be empirically validated in each cultural context it aims to be applied. It is well-known that a particular tool can perform adequately in a certain context and show different results in another due to cultural and language differences. Tools and tests are important for supporting decision-making not only in the driving context but in many other areas such as educational, clinical, organisational contexts etc. Therefore, conclusions about linguistically and culturally-diverse populations should only be made after an adequate adaptation of the tool (International Test Commission, 2010).

Finally, in the early studies many changes have been implemented in the analyses from one experiment to another (specifically experiments 1, 3 and 4). These changes were implemented in order to test the design and international validity of the traditional HP methodology and were not related to the creation of theoretical framework. On occasion, minor analyses have been conducted only in support of the major analyses (with no differences in the results) and therefore they have not been extensively described in the methods (e.g. the overtaking-clip analysis in experiment 4, changes in the scoring windows in experiment 4). These minor analyses should not be viewed as main aims of any of the studies and therefore should not be expected to follow-up from the previous results. While it is obviously preferable to have decided on all of the analyses prior to looking at the data, the formative nature of the earlier studies on the develop of this thesis benefited from these additional analyses.

The studies in this thesis adhered to the situation awareness theoretical framework of Endsley (1995) and supported the new framework proposed by Pradhan and Crundall (2017). The early studies were undertaken with a view to refining a methodology, and they certainly assisted in the development of the later, more coherent approach. The clear and logical sequence from experiment 5 onwards reflects the maturation of the research of this PhD thesis.

10. 9 Conclusion

The main finding of this thesis was that the newly developed hazard prediction methodology proved to be a better option for international export. The initial studies featuring the traditional hazard perception methodology provided an insight of why several studies failed to find the basic experiential effect (especially when applied overseas). Thus, it was possible to develop and validate new methodological guidelines, combining knowledge from previous theoretical and applied areas. As a result, the newly developed methodology successfully discriminated between experienced and novice drivers in different countries.

Although, this was just a first step towards a global validation, the prediction test has the potential to be the perfect blueprint for international testing and training. This opens multiple directions that generate future research, such as the possibility of a fuller hazard typology, a methodological protocol on how to build a hazard prediction test (an ISO would be ideal), global training, and the possibility of convincing even the UK Government that hazard prediction might be the next step forward for the official UK driving test.

References

Ābele, L., Haustein, S., Møller, M., & Martinussen, L. M. (2018). Consistency between subjectively and objectively measured hazard perception skills among young male drivers. *Accident Analysis and Prevention*, *118*, 214-220.

Adminaite, D., Jost, G., Stipdonk, H., & Ward, H. (2017). Ranking EU Progress on Road Safety. 11th Road Safety Performance Index Report.

Alberti, C. F., Shahar, A., & Crundall, D. (2014). Are experienced drivers more likely than novice drivers to benefit from driving simulations with a wide field of view? *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 124-132.

Allen, R. W., Park, G., Cook, M., Fiorentino, D., & Viirre, E. (2005). Novice driver training results and experience with a PC Based Simulator. *Proceedings of the 2nd International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design: Driving Assessment 2003*.

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

Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement*, *38*(6), 432-449.

Atchison, L. (2016). Reducing casualties involving young drivers and riders in Europe. *European Transport Safety Council*, Brussels. Retrieved from: https://etsc.eu/reducing-casualties-involving-young-drivers-and-riders-in-europe/

Barnard, Y., Utesch, F., Van Nes, N., Eenink, R., & Baumann, M. (2016). The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. *European Transport Research Review, 8*(2), 8-14.

Baron-Epel, O., Obid, S., Fertig, S., & Gitelman, V. (2015). Perceived descriptive safety-related driving norms within and outside Arab towns and villages in Israel. *Traffic Injury Prevention*, *17*(1), 51-57.

Bem, D.J. (1970). *Beliefs, attitudes, and human affairs*. Belmont, CA: Brooks/Cole.

Benda, H. V., & Hoyos, C. G. (1983). Estimating hazards in traffic situations. *Accident Analysis and Prevention, 15*(1), 1-9.

Blockey, P. N., & Hartley, L. R. (1995). Aberrant driving behaviour: errors and violations. *Ergonomics, 38*(9), 1759–1771.

Boholm, Å. (1998). Comparative studies of risk perception: A review of twenty years of research. *Journal of Risk Research*, *1*, 135–163.

Borowsky, A., & Oron-Gilad, T. (2013). Exploring the effects of driving experience on hazard awareness and risk perception via real-time hazard identification, hazard classification, and rating tasks. *Accident Analysis and Prevention, 59*, 548-565.

Borowsky, A., Oron-Gilad, T., & Parmet, Y. (2009). Age and skill differences in classifying hazardous traffic scenes. *Transportation Research Part F: Traffic Psychology and Behaviour, 12*(4), 277-287.

Borowsky, A., Oron-Gilad, T., & Parmet, Y. (2010). The role of driving experience in hazard perception and categorization: A traffic-scene paradigm. *World Academy of Science, Engineering and Technology 66*, 305-309.

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

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

Bränström, R., Kristjansson, S., & Ullen, H. (2006). Risk perception, optimistic bias, and readiness to change sun related behaviour. *European Journal of Public Health 16*, 492–497.

Brown, I. D. (1967). Decrement in skill observed after seven hours of car driving. *Psychonomic Science, 7*(4), 131-132.

Brown, I. D., & Groeger, J. A. (1988). Risk perception and decision taking during the transition between novice and experienced driver status. *Ergonomics, 31,* 587-597.

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

Castro, C.,Ventsislavova, P., Peña-Suárez, E.,Gugliotta, A.,García-Fernández, P., Eisman, E., & Crundall, D. (2016). Proactive listening to a training commentary improves hazard prediction. *Safety Science, 82*, 144–154.

Catchpole, J., & Leadbeatter, C. (2000). Redevelopment of Victoria's Hazard Perception Test. *Proceedings of Road Safety Research: Policing and Education Conference, 2000, Brisbane, Queensland*.

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

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.*

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

Chen, J. Y., Oden, R. N., Kenny, C., & Merritt, J. O. (2010). Stereoscopic displays for robot teleoperation and simulated driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 4*(19), 1488-1492.

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

Cheng, Y. K., Wong, K. H., Tao, C. H., Tam, C. N., Tam, Y. Y., & Tsang, C. N. (2016). Front blind spot crashes in Hong Kong. *Forensic Science International, 266*, 102–108.

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

Cockerton, C. G. & Isler, R. B. (2003). Developing real-life driving simulations for novice driver education. In Lassner & C. McNaught (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* 2003 (pp. 316-317). Chesapeake, VA: AACE.

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

Congdon, P. (1999). VicRoads Hazard Perception Test. Can it Predict Accidents? Australian Council for Educational Research, Victoria, Australia.

Congdon, P., & Cavallo, A. (1999). Validation of the Victorian Hazard Perception Test. In: *Research Policing Education Road Safety Conference*. (ATSB: Canberra, ACT).

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

Crundall, D. (2018, June). Understanding hazard perception. Presented at the International Congress of Applied Psychology, Montreal, CA.

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

Crundall, D., Chapman, P., Phelps, N., & Underwood, G. (2003). Eye movements and hazard perception in police pursuit and emergency response driving, *Journal of Experiment 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 and Prevention, 45*, 600–609.

Crundall, D., & Kroll, V. (2018). Prediction and perception of hazards in professional drivers: Does hazard perception skill differ between safe and less-safe fire-appliance drivers? *Accident Analysis and Prevention*, *121*, 335-346.

Crundall, D. E., & Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics*, *41*(4), 448-458.

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

Crundall, D., Underwood, G., & Chapman, P. (2002). Attending to the peripheral world while driving. *Applied Cognitive Psychology, 16*(4), 459–475.

Darby, P., Murray, W., & Raeside, R. (2009). Applying online fleet driver assessment to help identify, target and reduce occupational road safety risks. *Safety Science, 47*(3), 436-442.

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

De Craen, S., Twisk, D. A. M., Hagenzieker, M. P., Elffers, H., & Brookhuis, K.A. (2011). Do young novice drivers overestimate their driving skills more than experienced drivers? Different methods lead to different conclusions? *Accident Analysis Prevention*, *43*(5), 1660–1665.

DeJoy, D.M. (1989). The optimism bias and traffic accident risk perception. *Accident Analysis & Prevention 21*, 333–340.

Department for Transport (2017). Annual Reports and Accounts. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/741685/dft-annual-report-18-tagged.pdf

Department for Transport (2011). Local Transport Note 1/11. Shared Space. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/732739/ltn-1-11.pdf

Department for transport (2018). Reported Road Casualties in Great Britain: 2017. Annual Report. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/744077/reported-road-casualties-annual-report-2017.pdf

Department for Transport (2015). Report Road Casualties in Great Britain. Main Results 2015. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/533293/rrcgb-main-results-2015.pdf

Department for Transport (2016). Road Use Statistics Great Britain 2016. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/514912/road-use-statistics.pdf

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., … Knipling, R. R. (2006). The 100-Car Naturalistic Driving Study: Phase II -

Results of the 100-Car Field Experiment. (DOT HS 810 593) Virginia Tech Transportation Institute.

Dobkins, K. R., & Bosworth, R. G. (2001). Effects of set-size and selective spatial attention on motion processing. *Vision Research, 41*(12), 1501-1517.

Dodge, R. (1923). The human factor in highway regulation and safety. *Highway Research Board Proceedings, 2*(32), 73-78.

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.

Durso, F. T., & Gronlund, S. D. (1999). Situation awareness. In F. T. Durso (Ed.), Handbook of applied cognition (pp. 283-314). New York, NY, US: John Wiley & Sons Ltd.

Egea-Capparos, D. A. (2012). Estudiando la percepción de peligros en tráfico (hazard perception) en el laboratorio: una revisión. *Anales de Psicología, 28*(1), 240-265.

Egea-Caparrós, D., García-Sevilla, J., Pedraja, M., Romero-Medina, A., Marco-Cramer, M., & Pineda-Egea, L. (2016). Late detection of hazards in traffic: A matter of response bias? *Accident Analysis and Prevention*, *94*, 188-197.

Elias, W. (2018). The role of fines and rewards in the self-regulation of young drivers. *European Transport Research Review, 10*(6), 1-8.

Elliott, M. A. (2006). Can we use the Cohort II data to explore the effects of taking Pass Plus? In Behavioural Research in Road Safety: Sixteenth Seminar. London: Department for Transport, 225-236.

Endsley, M. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*.

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

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*(2), 97-120.

European Transport Safety Council (2016). Ranking EU Progress on Road Safety. 10th Road Safety Performance Index Report. Brussels, Belgium: Adminaite, D., Jost, G., Stipdonk, H., & Ward, H. Retrieved from: https://etsc.eu/wp-content/uploads/AR-2016-FINAL.pdf

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

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

Fisher, D., & Garay-Vega, L. (2012). Advance yield markings and drivers' performance in response to multiple-threat scenarios at mid-block crosswalks. *Accident Analysis and Prevention, 44*(1), 35-41.

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*, 25-29.

Flaherty, J., Gaviria, F., Pathak, D., Mitchell, T., Wintrob, R., Richman, J., & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous Mental Disease, 176*(5), 257–263.

Gao, L., Yu, X.T., & Hou, L.T. (2015). Research for the novice driver's Capacity of hazard perception and response. Y.H.Kim, *Proceedings of the2015 international conference on engineering management, engineering education and information technology 36,* 355–360.

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.

Geisinger, K. F. (1992). The metamorphosis to test validation. *Educational Psychologist, 27*(2), 197-222.

Gitelman, V., Dain, Y., Levi, S., & Eizenman, A. (2003). Causes and Characteristics of Road Accidents in the Non-Jewish Population. *The National Road Safety Authority-The Ministry of Transport, Jerusalem.*

Glendon, A. I., Dorn, L., Matthews, G., Gulian, E., Davies, D. R., & Debney, L. M. (1993). Reliability of the Driving Behaviour Inventory. *Ergonomics, 36*(6), 719-726.

Grayson, G. B., Maycock, G., Groeger, J. A., Hammond, S. M., & Field, D. T. (2003). Risk, hazard perception and perceived control. (TRL560). Retrieved from: https://trl.co.uk/sites/default/files/TRL560.pdf

Groeger, J. A. (2000). *Understanding Driving.* Taylor and Francis: London.

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

Groeger, J. A., Field, D., & Hammond, S. (1998). Processes involved in responding to risk. Behavioural research in road safety VIII, TRL, Crowthorne, UK.

Guggenheim, N., & Taubman – Ben-Ari, O. (2015). Ultraorthodox young drivers in Israel – Driving through cultural lenses. *Transportation Research Part F: Traffic Psychology and Behaviour*, *33*, 87-96.

Gugliotta, A., Ventsislavova, P., Garcia-Fernandez, P., Peña-Suarez, E., Eisman, E., Crundall, D., & Castro, C. (2017). Are Situation Awareness and Decision-Making in driving totally conscious processes? Results of a Hazard Prediction task. *Transportation research Part F: Traffic Psychology and Behaviour, 44*, 168-179.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309-333.

Hartos, J. L., Eitel, P., Haynie, D. L., & Simons-Morton, B. G. (2000). Can I take the car? Relations among parenting practices and adolescent problem-driving practices. *Journal of Adolescent Research, 15*(3), 352-367.

Harzing, A.-W. (2005). Book Review: Crossing Cultures: Insights from Master Teachers. *International Journal of Cross Cultural Management, 5*(2), 225–228.

Helman, S., Grayson, G. B., & Parkes, A. M. (2010). How can we produce safer new drivers? A review of the effects of experience, training and limiting exposure on the collision risk of new drivers. (INS005). Crowthorme, UK. Retrieved from: https://trl.co.uk/sites/default/files/INS005.pdf

Helman, S., Palmer, M., Delmonte, E. & Buttress, S. (2012). Development of a video measure of hazard perception skill and a group-discussion-based hazard perception training package for motorcyclists. Published Project Report (PPR615). Crowthorne: Transport Research Laboratory.

Hergovich, A., Arendasy, M. E., Sommer, M., & Bognar, B. (2007). The Vienna Risk-Taking Test - Traffic. *Journal of Individual Differences*, *28*(4), 198-204.

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

Hoffman, Y., & Rosenbloom, T. (2016). Driving experience moderates the effect of implicit versus explicit threat priming on hazard perception test. *Accident Analysis and Prevention*, *92*, 82-88.

Horrey, W. J., Lesch, M. F., Mitsopoulos-Rubens, E., & Lee, J. D. (2015). Calibration of skill and judgement in driving: Development of a conceptual framework and the implications for road safety. Accident Analysis and Prevention, 76, 25-33.

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), 939-44.

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

Horswill, M. S., & Helman, S. (2003). A behavioral comparison between motorcyclists and a matched group of non-motorcycling car drivers: factors influencing accident risk. *Accident Analysis and Prevention, 35*(4), 589-597.

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

Horswill, M. S., Marrington, S. 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*, 212–218.

Horswill, M. S., & McKenna, F. P. (1999). The development, validation, and application of a video-based technique for measuring an everyday risk-taking behaviour: drivers' speed choice. *Journal of Applied Psychology*, *84*(6), 977-985.

Horswill, M. S., & McKenna, F. P. (2004). Drivers' hazard perception ability: situation awareness on the road. In: Banbury, S., Tremblay, S. (Eds.), A Cognitive Approach to Situation Awareness: Theory and Application. Ashgate Publishing, Aldershot, Hampshire, pp. 155–175.

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

International Test Commission (ITC), 2010. International Test Commission Guidelines for Translating and Adapting Tests. Download From. Retrieved from: http:// www.intestcom.org&gt.

International Transport Forum (2017), Road Safety Annual Report 2017, OECD Publishing, Paris. http://dx.doi.org/10.1787/irtad-2017-en

International Transport Forum-OECD (2015). ITF Transport Outlook 2015. Retrieved from: https://read.oecd-ilibrary.org/transport/itf-transport-outlook-2015_9789282107782-en#page1

Isler, R., & Cockerton, C. (2003). A Computer-Based Interactive Multimedia Training CD-ROM for Novice Drivers in New Zealand. In D. Lassner & C. McNaught (Eds.), *Proceedings of ED-MEDIA 2003--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 799-800). Honolulu, Hawaii, USA: Association for the Advancement of Computing in Education (AACE).

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 and Prevention, 41*, 445–452.

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

Johnston, K. & Scialfa, C. (2016). Hazard perception in emergency medical service respondents. *Accident Analysis and Prevention, 95*, 91-96.

Kagitcibasi, C., Ataca, B., & Diri, A. (2010). Intergenerational Relationships in the Family: Ethnic, Socioeconomic, and Country Variations in Germany, Israel, Palestine, and Turkey. *Journal of Cross-Cultural Psychology, 41*(5-6), 652-670.

Kaparias, I., Bell, M. G. H., Singh, A., Dong, W., Sastrawinata, A., Wang, X., & Mount, B. (2013). Analysing the perceptions and behaviour of cyclists in street environments with elements of shared space. Paper presented at the 45th Annual

Conference of the Universities' Transport Study Group, January 2013, Oxford, UK.

Karndacharuk, A., Wilson, D. J., & Dunn, R. (2014). A Review of the Evolution of Shared (Street) Space Concepts in Urban Environments. *Transport Reviews, 34*(2), 190-220.

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 and Prevention*, *42*(3), 827–834.

Konstantopoulos, P., & Crundall, D. (2008). The Driver Prioritisation Questionnaire: Exploring drivers' self-report visual priorities in a range of driving scenarios. *Accident Analysis and Prevention, 40*(6), 1925-1936.

Korn, L., Weiss, Y., & Rosenbloom, T. (2017). Driving violations and health promotion behaviors among undergraduate students: Self-report of on-road behavior. *Traffic Injury Prevention*, *18*(8), 813-819.

Kruysse, H. W. (1991). The subjective evaluation of traffic conflicts based on an internal concept of dangerousness. *Accident Analysis and Prevention, 23*(1), 53-65.

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

Lavallière, M., Reimer, B., Mehler, B., D'Ambrosio, L., Wang, Y., Teasdale, N., & Coughlin, J. (2011). The Effect of Age and Gender on Visual Search During Lane Changing. *Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2011*.

Lee, J.D., Young, K.L., & Regan, M.A. (2008). Defining driver distraction. In: Regan, M.A., Lee, J.D., Young, K.L. (Eds.), Driver Distraction: Theory, Effects, and

Mitigation. CRC Press Taylor & Francis Group, Boca Raton, FL, USA, pp. 31–40.

Lehtonen, E., Lappi, O., Koirikivi, I., & Summala, H. (2014). Effect of driving experience on anticipatory look-ahead fixations in real curve driving. *Accident Analysis and Prevention,70*, 195-208.

Lehtonen, E., Sahlberg, H., Rovamo, E., & Summala, H. (2017). Learning game for training child bicyclists' situation awareness. *Accident Analysis and Prevention, 105*, 72-83

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.

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

Lindemann, P., & Rigoll, G. (2017). A diminished reality simulation for driver-car interaction with transparent cockpits. *2017 IEEE Virtual Reality (VR)*.

Lund, I., & Rundmo, T. (2009). Cross-cultural comparisons of traffic safety, risk perception, attitudes and behaviour. *Safety Science 47*, 547–53.

Mackenzie, A., & Harris, J. (2014). Visual attention and driving: How to measure it and how to train it. *i-Perception, 5*(5), 477-477.

Malone, S. & Brünken, R. (2013). Assessment of driving expertise using multiple choice questions including static vs. animated presentation of driving scenarios. *Accident Analysis and Prevention, 51*, 112-119.

Malone, S., & Brünken, R. (2015). Hazard perception assessment: how much ecological validity is necessary? *Journal of Manufacturing Processes, 3*, 2769–2776.

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

Mattsson, M., O'Brien, F., Lajunen, T., Gormley, M., & Summala, H. (2015). Measurement invariance of the Driver Behavior Questionnaire across samples of young drivers from Finland and Ireland. *Accident Analysis & Prevention, 78*, 185–200.

Mayhew, D. R., Simpson, H. M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. *Accident Analysis and Prevention, 35*(5), 683–691.

McCartt, A., Mayhew, D., Braitman, K., Ferguson, S., & Simpson, H. (2009). Effects of Age and Experience on Young Driver Crashes: Review of Recent Literature. *Traffic Injury prevention*, 10, 209-19.

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

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* (Supplement), S15–23 1.

McGowan, A., & Banbury, S. (2004). Evaluating interruption-based techniques using embedded measures of driver anticipation. In: Banbury, S.P., Tremblay, S. (Eds.), A Cognitive Approach to Situation Awareness. Ashgate Publishing, Aldershot, Hampshire, pp. 176–192.

McKenna, F.P., & Crick, J.L. (1991). Hazard Perception in Drivers: A Methodology for Testing and Training (Final Report). Transport and Road Research Laboratory, Crowthorne, UK.

McKenna, F.P., & Crick, J. (1997). Hazard Perception in Drivers: A Methodology for Testing and Training. Transport Research Laboratory, Crowthorne, UK.

McKenna, F.P. & Horswill, M.S. (1998). *The effect of skill acquisition on risk-taking*. Department of Psychology, University of Reading, Reading.

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. (2006). Does anticipation training affect drivers' risk taking? *Journal of Experimental Psychology: Applied*, 12(1), 1–10.

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

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.

Mills, K. L., Hall, R. D., McDonald, M., & Rolls, G. W. P. (1998). The effects of hazard perception training on the development of novice driver skills, Research Report. London: Department for Transport, UK. Available at: http://www.dft.gov.uk.

Mourant, R. R., & Rockwell, T. H. (1970). Visual Information Seeking of Novice Drivers. *SAE Technical Paper Series.*

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

Niewoehner, W., & Berg, A. (2005). Endangerment of Pedestrians and Bicyclists at Intersections. Retrieved from http://www-nrd.nhtsa.dot.gov

Nordfjærn, T., Jørgensen, S., & Rundmo, T. (2011). A cross-cultural comparison of road traffic risk perceptions, attitudes towards traffic safety and driver behaviour. *Journal of Risk Research, 14*(6), 657-684.

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*, 91–98.

Nordfjærna, T., Şimşekoğlu, Ö., & Rundmo, T. (2014). Culture related to road traffic safety: A comparison of eight countries using two conceptualizations of culture. *Accident Analysis & Prevention 62*, 319–328.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education, 15*(5), 625-632.

NZ Transport Agency (2012). Annual report. Retrieved from: https://www.nzta.govt.nz/assets/resources/annual-report-nzta/docs/nzta-annual-report-2012.pdf

Olsen, E. C., Lee, S. E., & Simons-Morton, B. G. (2007). Eye Movement Patterns for Novice Teen Drivers. *Transportation Research Record: Journal of the Transportation Research Board, 2009*(1), 8-14.

Olsson-Collentine et al (2019). The Prevalence of Marginally Significant Results in Psychology Over Time. *Psychological Science, 30*, 576-586.

Owens, R. E., Hanna, G. S., & Coppedge, F. L. (1970). Comparison of multiple-choice tests using different types of distractor selection techniques. *Journal of Educational Measurement, 7*(2), 87-90.

Özkan, T., Lajunen, T., Chliaoutakis, J., Parker, D., & Summala, H. (2006). Cross-cultural differences in drivinbehaviorsrs: A comparison of six countries. *Transportation Research Part F: Traffic Psychology and Behaviour, 9*, 227–242.

Özkan, T., Lajunen, T., Parker, D., Sümer, N., & Summala, H. (2011). Aggressive driving among British, Dutch, Finnish and Turkish drivers. *International Journal of Crashworthiness, 16*(3), 233–238.

Özkan, T., Lajunen, T., & Summala, H. (2006). Driver Behaviour Questionnaire: A follow-up study. *Accident Analysis & Prevention, 38*(2), 386–395.

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

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

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

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

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

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

Petzoldt, T., Weiss, T., Krems, J. F., & Bannert, M. (2013). The Development of a Cognitive Skills Training to Support Driver Education – Comparing Performance of Experienced and Trained Learner Drivers. *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2013*.

Pollatsek, A., Fisher, D. L., & Pradhan, A. (2006). Identifying and Remedying Failures of Selective Attention in Younger Drivers. *Current Directions in Psychological Science*, *15*(5), 255-259.

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, 48*(3), 447-464.

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

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

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

Pritschet, L., Powell, D., Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, 27, 1036–1042.

Rameau, F., Ha, H., Joo, K., Choi, J., Park, K., & Kweon, I. S. (2016). A Real-Time Augmented Reality System to See-Through Cars. *IEEE Transactions on Visualization and Computer Graphics*, *22*(11), 2395-2404.

RAC, Report on Motoring (2017). Feeling the Squeeze. Retrieved from: https://www.rac.co.uk/pdfs/report-on-motoring/rac_rom_2017.pdf

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

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied, 9*(2), 119-137.

Report on the Development of Road Traffic Safety 2017 (2017). Beijing. Retrieved from http://auto.china.com.cn/news/20171219/685572.shtml.

Rosenbloom, T., Nemrodov, D., & Barkan, H. (2004). For heaven's sake follow the rules: pedestrians' behavior in an ultra-orthodox and a non-orthodox city. *Transportation Research Part F: Traffic Psychology and Behaviour*, *7*(6), 395-404.

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

Rosenbloom, T., Shahar, A., & Perlman, A. (2008). Compliance of Ultra-Orthodox and secular pedestrians with traffic lights in Ultra-Orthodox and secular locations. *Accident Analysis & Prevention*, *40*(6), 1919-1924.

Rowe, R., Roman, G., McKenna, F, Barker, E., & Poulter, D. (2015). Measuring errors and violations on the road: A bifactor modeling approach to the Driver Behavior Questionnaire. *Accident Analysis & Prevention*, *74*, 118–125.

Rudloff, C., Schoenauernauer, R., & Fellendorf, M. (2013). Comparing a calibrated shared space simulation model to real life data. *Transportation Research Record: Journal of the Transportation Research Board*, *2390*, 44-52.

Rundmo, T., Granskaya, J., & Klempe, H. (2012). Traffic culture as symbol exchange – A cross-country comparison of Russia and Norway. *Safety Science, 50*(5), 1261–1267.

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

Savage, S. W., Potter, D. D., & Tatler, B. W. (2013). Does preoccupation impair hazard perception? A simultaneous EEG and Eye Tracking study. *Transportation Research Part F: Traffic Psychology and Behaviour, 17*, 52-62.

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 and Prevention, 43*(1), 204–208.

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

See, J.E., Warm, J.S., Dember, W.N., & Howe, S.R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors, 39*, 14-29.

Sexton, B. (2000). Development of hazard perception testing. *Proceedings of the DETR Novice Drivers Conference*, Bristol. Available at: http://www.dft.gov.uk

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

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

Shimazaki, K., Ito, T., Fujii, A., & Ishida, T. (2017). Improving drivers' eye fixation using accident scenes of the Hazard Touch driver-training tool. *Transportation Research Part F: Traffic Psychology and Behaviour, 51*, 81–87.

Shinar, D. (1984). Actual versus estimated night-time pedestrian visibility. *Ergonomics, 27*(8), 863-871.

Sjöberg, L. (2000). Factors in Risk Perception. *Risk Analysis 20*(1), 1–12.

Smith S., Horswill M., Wetton M., & Chambers, B. (2009). Sleepiness and hazard perception while driving (Road Safety Grant Report No. 2009-001). Canberra, Australia: the Department of Infrastructure, Transport, Regional Development and Local Government. Retrieved from http://www.infrastructure.gov.au/roads/safety/publications/2009/RSRG 200 9001.aspx on 7th April 2009.

Spicer, (1964) cited in Pelz, D. C. & Krupat, E. (1974). Caution profile and driving record of undergraduate males. *Accident Analysis and Prevention, 6,* 45-58.

Standards for Highways online resources. (2018). Design Manual for Roads and Bridges 2018. Retrieved from: https://www.gov.uk/guidance/standards-for-highways-online-resources#the-design-manual-for-roads-and-bridges

Stanojević, P., Lajunen, T., Jovanović, D., Sârbescu, P., & Kostadinov, S. (2018). The driver behaviour questionnaire in South-East Europe countries: Bulgaria, Romania and Serbia. *Transportation Research Part F: Traffic Psychology and Behaviour, 53*, 24–33.

Stecklov, G., & Goldstein, J. R. (2004). Terror attacks influence driving behavior in Israel. *Proceedings of the National Academy of Sciences*, *101*(40), 14551-14556.

Stecklov, G., & Goldstein, J. R. (2010). Societal Responses to Endemic Terror: Evidence from Driving Behavior in Israel. *Social Forces*, *88*(4), 1859-1884.

Summerskill, SJ, Marshall, R, Cook, S, Lenard, J, & Richardson, J. (2015) The use of volumetric projections in Digital Human Modelling software for the identification of large goods vehicle blind spots. *Applied Ergonomics: human factors in technology and society, 53*(Part A), 267-280.

Taubman-Ben-Ari, O., & Katz-Ben-Ami, L. (2012). The contribution of family climate for road safety and social environment to the reported driving behavior of young drivers. *Accident Analysis and Prevention, 47,* 1-10.

Taubman-Ben-Ari, O., Mikulincer, M., & Gillath, O. (2004). The multidimensional driving style inventory-scale construct and validation. *Accident analysis and prevention 36*(3), 323-32.

Taubman-Ben-Ari, O. & Skvirsky, V. (2016). The multidimensional driving style inventory a decade later: Review of the literature and re-evaluation of the scale. Accident Analysis and Prevention, 93, 179–188.

Taubman – Ben-Ari, O., Skvirsky, V., Greenbury, T. J., & Prato, C. G. (2018). Explaining risks behind the wheel – Comparing Israeli and Queensland young drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 184-192.

The Central Bureau of Statistics (Israel)-Government Authorities, 2012. Available online: http://www.cbs.gov.il/webpub/pub/text_page.html?publ=78&CYear=2012&CMonth=1 (Accessed February 3, 2017)

The Central Bureau of Statistics (Israel)-Government and Local Authorities-Local Authorities. Available online at: http://www.cbs.gov.il/reader/?MIval=cw_usr_view_SHTML&ID=446 (Accessed February 3, 2017)

The Central Bureau of Statistics. *Press Release*. Hebrew. CDC (2016), 1–6. Available online at: http://www.cbs.gov.il/reader/newhodaot/hodaa_template.html?hodaa=201627020

Toroyan, T. (2009). Global status report on road safety. *Injury Prevention, 15*(4), 286–286.

Tūskė, V., Šeibokaitė, L., Endriulaitienė, A., & Lehtonen, E. (in Press). Hazard perception test development for Lithuanian drivers. *IATSS Research*.

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

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

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

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

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

UK Government (2015). Hazard perception clips get a modern makeover. Retrieved form: https://www.gov.uk/government/news/hazard-perception-clips-get-a-modern-makeover

Unverricht, J., Sanuel, S., & Yamani, Y. (2018). Latent Hazard Anticipation in Young Drivers: Review and meta-analysis of training studies. *Transportation Research Record, 2672*(33) 11–19.

Van Schagen, I., & Sagberg, F. (2012). The potential benefits of naturalistic driving for road safety research: Theoretical and empirical considerations and challenges for the future. *Procedia - Social and Behavioral Sciences, 48*, 692-701.

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

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

Vlakveld, W.P. (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.

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

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

Wang, W., Cheng, Q., Li, C., André, D., & Jiang, X. (2019). A cross-cultural analysis of driving behavior under critical situations: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour, 62*, 483-493.

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

Warner, H. W., Özkan, T., Lajunen, T., & Tzamalouka, G. (2011). Cross-cultural comparison of drivers' tendency to commit different aberrant driving behaviours. *Transportation Research Part F: Traffic Psychology and Behaviour, 14*(5), 390–399.

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

Wegman, F. (2017). The future of road safety: A worldwide perspective. *IATSS Research, 40*(2), 66-71.

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.

West, R. J., Wilding, J. M., French, D. J., Kemp, R., & Irving, A. (1993). Effect of low and moderate doses of alcohol on driving hazard perception latency and driving speed. *Addiction, 88*(4), 527-532.

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

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 and Prevention*, *54*, 57-66.

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

White, M., Cunningham, L. C., & Titchener, K. (2011). Young drivers' optimism bias for accident risk and driving skill: Accountability and insight experience manipulations. *Accident Analysis and Prevention, 43*(4), 1309-1315.

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

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1*(2), 202-238.

World Health Organisation (2013). Global Status Report on Road Safety 2013. Supporting a decade of action. Retrieved from: file:///C:/Users/User/Downloads/9789241564564_eng.pdf

World Health Organisation (2013). Global Status Report on Road Safety 2013.Time for Action. Retrieved from: http://apps.who.int/iris/bitstream/handle/10665/44122/9789241563840_eng.pdf?sequence=1

World Health Organisation (2015). Global Status Report on Road Safety 2015. Retrieved from: https://www.who.int/violence_injury_prevention/road_safety_status/2015/GSRRS2015_Summary_EN_final2.pdf

World Health Organisation (2018). Global Status Report on Road Safety 2018. Retrieved from: file:///C:/Users/User/Downloads/9789241565684-eng.pdf

World Life Expectancy (2018). Road Traffic Accidents. Age-Adjusted Death Rate. Retrieved from: http://www.worldlifeexpectancy.com/

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools*. New York: Longman**.**

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

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

Yanagi, T., Fernando, C. L., Saraiji, Y., Minamizawa, K., Tachi, S., & Kishi, N. (2015). Transparent cockpit using telexistence. *Proceedings IEEE Virtual Reality Conference, VR 2015*, Alres, France.

Yasuda, H., & Ohama, Y. (2012). Toward a practical wall see-through system for drivers: How simple can it be? *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*.

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

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

Zhang,Y., & Ioannou, P.A. (2016).Combined variable speed limit and lane change control for high way traffic. *IEEE Transactions on Intelligent Transportation System, 99*,1–12.

Appendix A

Health and Safety Guidelines

HSE

Health and Safety Executive

Safe filming and recording involving vehicles

HSE information sheet

Entertainment Information Sheet No 22 (Revision 1)

Introduction

In this information sheet, '**must**' denotes a legal obligation. Words such as 'do', 'should' etc are used to give advice on good practice and are not compulsory.

This information sheet is one of a series produced by HSE in consultation with the Joint Advisory Committee for Entertainment (JACE). It gives general guidance on the risks associated with filming and recording of and from vehicles and is relevant to film, television and radio. The general principles can be applied to all road vehicles.

This information sheet only deals with matters relating to health and safety. For issues relating to the Road Traffic Acts, the police should be consulted.

Activities this guidance relates to:

- Road testing or other car or motorcycle shooting on public roads.
- Filming of motorcycles from cars at normal speeds on public roads.
- Road testing and driving shots of cars and motorcycles at high(er) speeds on closed tracks.
- Recording interviews, dialogue or commentary by a driver or passenger of a moving vehicle.
- Filming or recording commentary of motor racing, from trackside or using mounted cameras.
- Recording commentary as a passenger in a motor racing car.
- Use of A-frames, low-loaders or specialist modified vehicles.
- Off-road use, including 4 x 4 vehicles.

Legislation

The main legal requirements covering recording from vehicles are the Health and Safety at Work etc Act 1974, the Management of Health and Safety at Work Regulations 1999 as amended, (the Management Regulations) and the Road Traffic Acts of 1988 and 1991 and related legislation.

The Management Regulations require a suitable and sufficient risk assessment to be conducted. This must be carried out by employers and by the self-employed to assess the risk to themselves, their employees and others who may be affected by their activities. It should be used to identify control measures that can be implemented to control the risks identified. The opportunity should be taken during risk assessment to consider the application of any other relevant health and safety legislation, including the requirement to consider fire precautions and emergency procedures.

If a vehicle is to be used on the public highway all road traffic legislation must be complied with, including the Road Traffic Acts 1988 and 1991. The Highway Code and local by-laws, as well as health and safety legislation, will also be relevant.

Where there are concerns over compliance with road traffic legislation, the police or local authority should be consulted as appropriate. Advice should be sought from the local authority responsible for the location where the action will be filmed. Where road traffic legislation does not apply, health and safety legislation will still be relevant.

The activities of the production must not put any other road users at risk. This includes making sure warning signs for all vehicles are fully operational and visible to other road users, eg braking lights.

There are certain incidents that occur on the public highway that may be reportable to HSE as well as the police. The most likely situation in filming would be if someone who is unloading by the roadside is injured. Any incident resulting in major injury or over 3-day injury on land where the road traffic legislation does not apply is reportable to HSE. An over 3-day injury is one that is not 'major' but which results in the injured person being away from work or unable to do the full range of their normal duties for more than three days.

For the latest requirements go to www.hse.gov.uk/riddor.

Competence

All drivers need to be competent for the activity they are to be involved in; anything that they are asked to do should be within their ability. This is of particular concern when the driver is an artiste or a member of the public. Adequate time should be allowed for rehearsal, especially if the driver is inexperienced in delivering pieces to camera etc. Competence includes being aware of the vehicle's limitations and the limitations of their own driving ability. People should not be encouraged to exceed them.

A person at the wheel should not be distracted from driving, for example by being in sole charge of recording equipment. Where specialist driving is required, for example vehicle stunts or high-speed work, the driver should have appropriate experience and, where applicable, qualifications. The vehicle should be suitable for the activity.

Unless the camera operator and bike rider are experienced in the activity, manned cameras on motorcycles should not be used.

Risk assessment

Operations on, in or near vehicles involve risk and therefore a risk assessment must be carried out before any production activity. This must be reviewed as necessary if more information becomes available or if circumstances change. The risk assessment will determine what control measures are needed.

The risk assessment may require an exchange of information with other parties, eg event organisers or police. These other parties should also have carried out a risk assessment of their own activities. The producer/event organiser should liaise with these bodies to make sure that all hazards arising from the event have been identified and that everyone is aware of the risks arising from each other's activities and of the control measures that are in place.

What can go wrong?

For example:

- injury to crew, presenters or other workers;
- injury to members of the public/onlookers;
- hearing damage - exposure to high levels of noise could cause hearing damage;
- inability to communicate between presenters and crews;
- accident at high speed (race speeds on motor racing circuits and maximum speed limits on public roads);
- in-car camera/recording equipment could injure passengers in vehicles;
- distracting other drivers.

Participation

If the impression that a performer is participating in a race is required, you should consider whether participation is really necessary to achieve the recording or pictures that are needed. Recording the piece without pressures of participation is a good control measure. All speeds and manoeuvres must be kept to safe limits. Sufficient rehearsal time should be allowed.

Position of equipment/personnel

The equipment and any operator must not obstruct the driver's view, distract their attention while driving or put other persons at risk. The position of personnel must be safe in all circumstances, including all foreseeable emergencies.

When working in motor racing pits, production teams should have a 'minder' with them to look out for dangerous movement of vehicles or equipment in potentially dangerous positions. These are high-risk areas and the risk assessment must clearly demonstrate the need to be present, the nature of the activities and the proposed control measures with specific reference to communications, positioning of personnel and provision of appropriate personal protective equipment.

Trackside positions should take account of the possibility of vehicles leaving the track at speed, eg beyond bends or at wet/muddy stretches. Where this risk exists, advice should be taken from competent persons, eg a racing driver, about filming positions and an assessment made which considers the use of protective barriers or locked off or remote cameras.

When filming from the passenger seats of vehicles, people should be properly seated with seat belts or in a safe and secure position. People should not crouch in a footwell. The camera can be fixed in the required position, and the picture viewed via a remote monitor.

When shooting through a side window, operators should not lean out or let the camera lens protrude beyond the plane of the vehicle. Similarly, a side-mounted camera may be a hazard to vehicles overtaking and/or pedestrians if it protrudes.

When filming 'up and passes' or 'run bys' on a public highway, the camera and equipment should be a safe distance from the edge of the road and on the inside of any bends ('up and passes' describes the way the camera is used to film the vehicle.)

Filming by a pillion passenger facing backwards on a bike is only acceptable if the bike is properly adapted, the rider and camera operator are competent specialists and the activity complies with any road traffic legislation/local by-laws. For all work on or close to the public highway, the need for appropriate high-visibility clothing must be considered in the production risk assessment.

Restraints on equipment and people

While a vehicle is being driven on a public highway drivers and passengers must be effectively restrained in accordance with road traffic legislation. Wherever possible, lightweight cameras should be used. Camera operators must not stand up and shoot through the sunroof while the vehicle is moving.

There must be effective measures to prevent equipment striking any occupant in the event of an emergency stop or collision. Steps should be taken to prevent small items of equipment from causing a hazard to the normal operation of the vehicle or striking somebody in the event of a collision.

The likelihood that convertible-type or open-backed vehicles may overturn must be considered and appropriate controls put in place. The controls should be identified as a result of the risk assessment and may include the reduction of speed, use of vehicles fitted with roll bars etc. Any film or recording equipment should either be securely mounted or independently secured by safety lanyards. An exception may be for a small, hand-held camera used in the front seat, but even then the risk to a driver in a side-impact must be considered. The effect of the deployment of airbags must also be considered as any equipment in the vicinity may be pushed violently towards occupants of the vehicle.

Communication

Drivers and crews must be briefed beforehand of the shots to be attempted. There must be effective communication at all times.

When using a tracking vehicle the driver of any overtaking vehicle should decide when it is safe to pass. Only competent and authorised personnel who have appropriate training or expertise and compliant clothing may direct traffic.

Selection of equipment

When it is not possible to use a full-size camera safely, for example to achieve close-up, car-to-car shots of wheels, a remote mini-camera should be used. Mini-camera equipment and mounts must be fit for purpose and should only be fitted by competent persons. Unmanned, fixed mini-cameras can be used for pieces to camera on motor vehicles. Equipment must be placed so as not to increase the likelihood of an injury in the event of an accident. When mounted on motorcycles, consideration should be given to the location so as not to adversely affect the balance of the bike.

Always use an appropriate vehicle for the planned activity. Vehicles must be fully maintained and in a roadworthy condition, complying with the road traffic and vehicle legislation that applies. They must also be used in the way they were designed to be used. If there is any modification to a vehicle, this must be subject to a full risk assessment.

Pieces to camera while driving

Performers need to keep their eyes on the road and look at the camera for no longer than they would normally look away to inspect their instruments or look in the rear-view mirror. They should also try to keep both hands on the wheel when recording pieces to camera, except for the occasional gesture or for operation of controls.

Tracking vehicles and the use of low-loaders and 'A-frames'

There should be a nominated competent person who will be responsible for the control of the tracking vehicle and who will work in conjunction with the driver. Tracking on public highways should only be done with the agreement of the local authority/ highways agency and the local police authority. Tracking shots (car-to-car) should only be done from a specially modified tracking vehicle or one in which the structural integrity is not compromised by having doors/tailgates open. The route should be checked prior to filming to identify potential hazards such as speed bumps etc.

All the crew should be suitably restrained when the vehicle is in motion. Care should be taken to ensure no loose objects will fall out of the moving car - the camera and any other kit should be secured appropriately. The vehicle and all technical equipment should be fully secured by a competent person. Also, the lights should not create glare that may distract other motorists.

Care should be taken when shooting cars mounted on low-loaders or A-frames. If a passenger is to be carried on the back, you must consult with the local police force. Anyone filming or recording should be securely harnessed to appropriate anchor points and no one should be permitted to sit on the handrails.

Safety rails on low-loaders and A-frames must not be altered for any filming purpose unless full approval is given by the equipment supplier and the work is carried out by experienced and qualified vehicle technicians.

The maximum weight and limit on numbers of persons for low-loaders and A-frames must not be exceeded. Information regarding the maximum weights permitted can be found on a plate located either in the cab of the vehicle or on the nearside of the trailer. Further information may be obtained from the 'plating certificate' issued to the vehicle operator/owner.

If you are intending to use a low-loader with an extension platform fitted and the vehicle exceeds the legal road width, you should consult the local authority/highways agency.

All unit crew personnel working on or around a low-loader or A-frame must be fully briefed in the safety measures.

Consideration should be given to extra front/rear 'buffer' vehicles to give advance warning of unforeseen hazards ahead and to create a safe area behind the moving scene.

Noise

The Control of Noise at Work Regulations 2005 requires employers to take specific action at certain action values. These relate to levels of exposure of noise to employees calculated over a working day or week and the maximum noise (peak sound pressure) to which employees are exposed in a working day.

Employers are required to carry out a noise assessment if their employees are likely to have a daily or weekly exposure of 80 dB (A weighted) or to be exposed to a peak sound pressure of 135 dB (C weighted). Employees are not to be exposed to noise above the exposure limit (137 dB (C weighted). The peak sound pressure level of 135 dB could be exceeded by acts such as starting up performance vehicles and motorbikes. There are further requirements placed on employers by the Regulations to:

- Take action to reduce the exposure to the noise producing risk.
- Provide employees with suitable hearing protection if you cannot reduce the noise exposure enough by using other methods.
- Make sure the noise exposure limits are not exceeded.
- Provide employees with information, instruction and training.
- Carry out health surveillance where your risk assessment shows there is a risk to health.

Vehicle noise levels are a particular problem at motoring events, as talkback may have to be at high levels for operators to hear it clearly. The consequent risk of hearing damage can be reduced by wearing sound-excluding headphones to reduce the effect of vehicle noise or allowing loud talkback through ordinary headphones, but wearing earplugs to reduce the noise levels reaching the ear.

Further reading

Health and safety in audio-visual production: Your legal duties Leaflet INDG360 HSE Books 2002 (priced packs of 5 ISBN 978 0 7176 2551 2) www.hse.gov.uk/pubns/indg360.pdf

Management of health and safety at work. Management of Health and Safety at Work Regulations 1999. Approved Code of Practice and guidance L21 (Second edition) HSE Books 2000 ISBN 978 0 7176 2488 1 www.hse.gov.uk/pubns/books/l21.htm

Controlling noise at work. The Control of Noise at Work Regulations 2005. Guidance on Regulations L108 (Second edition) HSE Books 2005 ISBN 978 0 7176 6164 0 www.hse.gov.uk/pubns/books/l108.htm

Health and safety at motor sport events: A guide for employers and organisers HSG112 HSE Books 1999 ISBN 978 0 7176 0705 1 www.hse.gov.uk/pubns/books/hsg112.htm

Road Traffic Acts 1988 (c.52), 1991 (c.40) The Stationery Office

The Official Highway Code: Revised 2007 edition The Stationery Office 2007 ISBN 978 0 11 552814 9

www.hse.gov.uk/risk

Further information

For information about health and safety, or to report inconsistencies or inaccuracies in this guidance, visit www.hse.gov.uk/. You can view HSE guidance online and order priced publications from the website. HSE priced publications are also available from bookshops.

The Stationery Office publications are available from The Stationery Office, PO Box 29, Norwich NR3 1GN Tel: 0870 600 5522 Fax: 0870 600 5533 email: customer.services@tso.co.uk Website: www.tsoshop.co.uk/ (They are also available from bookshops.) Statutory Instruments can be viewed free of charge at www.legislation.gov.uk/.

This document contains notes on good practice which are not compulsory but which you may find helpful in considering what you need to do.

Appendix B

Risk Assessment

School/Division: **Psychology**

Date Of Last Assessment: **06/02/15**

Responsible Person Assessment: Name: **Prof. D. Crundall**

Section: Name of Assessor: **Prof. D. Crundall**

Date Of Next Assessment: n/a

Signature: **D. E. Crundall**

This is the risk assessment for filming naturally occurring hazard perception clips on the real roads using two internally mounted cameras and two externally mounted cameras.

Activity	Hazards	Existing Precautions	OK/ NOK	Deficiencies	Hazard Severity	Risk	Extent	Risk Severity
RF1	Hazard of the cameras attached to the car falling off.	Attaching the cameras will adhere to health and safety regulations (see Appendix A). The external cameras will be attached to the car via secure suction mounts and will be tethered to the vehicle. Therefore, in the event that a suction mount fails, the camera will still remain connected to the vehicle and will not fly off.			1	1	2	**2**

RF2	Hazard of attached camera obscuring the view of the driver or other road users.	The researchers will ensure that the external cameras do not extend further than the wing mirrors of the car and do not obscure the mirrors or vision of the drivers. Please refer to Health and Safety Entertainment Information Sheet No 22 (Appendix A).			1	1	1	**1**
RF3	Hazard of The film car driver changing their driving style so as to increase crash risk above the normal level	The driver (Crundall) has extensive experience in driving while filming. He is aware of the factors that might change a driver's behaviour while filming hazard perception clips, and can therefore avoid them. Only drives that are necessary for reasons other than filming will be used and therefore no additional risk or exposure should be incurred.			1	1	2	**2**

APPENDIX C: PUBLISHED ARTICLE: Ventsislavova, Crundall, Baguley, Castro, Gugliotta, Garcia-Frnandez, Zhang, Ba, & Li (2019)

Accident Analysis and Prevention 122 (2019) 268–286

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

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

Petya Ventsislavova[a,*], David Crundall[a,*], Thom Baguley[a], Candida Castro[b], Andrés Gugliotta[b], Pedro Garcia-Fernandez[b], Wei Zhang[c], Yutao Ba[c,1], Qiucheng Li[c]

[a] Nottingham Trent University, Department of Psychology, School of Social Sciences, UK
[b] University of Granada, CIMCYC, Mind, Brain and Behaviour Research Centre, Spain
[c] Tsinghua University, Department of Industrial Engineering, Beijing, China

ARTICLE INFO

Keywords:
Hazard perception
Hazard prediction
Driving safety

ABSTRACT

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

1. Introduction

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

1.1. Traffic death and injury as a global problem

Injuries and fatalities arising from traffic collisions are a global problem. The World Health Organization (2015) estimates the number of global fatalities due to traffic collisions to be 1.25 million, with up to 90% of these occurring in low to middle-income countries. Currently road traffic collisions are the 8th leading cause of death in the world, but are predicted to rise to the 5th leading cause by 2030 unless a concerted effort is made to avert this disaster. In response to this growing problem, the United Nations declared a Decade of Action for Road Safety which began in 2011, with the aim of first stabilising and then reducing road traffic fatalities and injuries by 2020. With over 100 countries pledged to assist, the Decade of Action is focused upon '5 pillars' of road safety: road safety management, developing safer roads, developing safer vehicles, developing safer road users, and improving emergency responses to incidents. One key aspect of this is the technological promise of automated vehicles, but the Institute of Electrical and Electronics Engineers (http://www.ieee.org/about/news/2012/5september_2_

* Corresponding authors.
E-mail addresses: petya.petrova@ntu.ac.uk (P. Ventsislavova), david.crundall@ntu.ac.uk (D. Crundall).
[1] Yutao Ba is now based at IBM China, Zhongguancun Software Park 19, Haidian, Beijing, 100094, China.

https://doi.org/10.1016/j.aap.2018.10.010
Received 23 August 2017; Received in revised form 17 October 2018; Accepted 17 October 2018
Available online 02 November 2018

2012.html) estimates that we will have to wait until 2040 before 75% of global driving stock is automated. Even if this ambitious target is met, millions will die before we reach this point without a wider range of safety initiatives, and it will be the low and middle-income countries that will continue to bear the brunt of this automotive pandemic.

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

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

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

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

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

1.2. An international perspective on hazard perception

Although the UK was the first country to include an HP test as part of the licensing procedure, both Australia and the Netherlands have since developed their own HP tests. In addition, there are research groups around the world who have developed HP tests in their own countries, including Australia Spain, Germany, The Netherlands, Israel, Singapore, Malaysia, Canada, Hong Kong, China, Japan, and New Zealand (e.g. Borowsky et al., 2010; Cheng et al., 2011; Cocron et al., 2014; Gao et al., 2015; Horswill et al., 2010, 2015; Malone and Brünken, 2016; Isler et al., 2008; Lim et al., 2013; Rosenbloom et al., 2011; Scialfa et al., 2014; Shimazaki et al., 2017; Ventsislavova et al., 2016; Vlakveld, 2011, 2014; Wang et al., 2007; Wetton et al., 2011, 2010; Yeung and Wong, 2015). Unfortunately, the results of many studies from around the world are mixed, with some researchers finding safe to perform better than les-safe drivers (e.g.Wallis and Horswill, 2007; Horswill et al., 2015), while others fail to find this basic effect (e.g.Sagberg and Bjørnskay, 2006; Lim et al., 2013; Yeung and Wong, 2015).

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

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

To our knowledge, only Lim et al. (2013, 2014) have measured performance on the exact same test across two different countries[2]. In 2013 they compared Malaysian and UK drivers' hazard perception performance on clips filmed in both countries. They found that the UK drivers responded to many more hazards than the Malaysian drivers, especially when they were presented with Malaysian clips. A difference between novice and experienced drivers did not materialise however (in both Malaysian and UK participants). The authors suggested that cultural differences in hazard criterion (the internal threshold at which one considers an event to be a 'hazard') impacted more on test performance than experience. As Malaysian drivers typically encounter more hazards on the road than UK drivers, these events become normalised to the extent that a scenario must be extremely dangerous before they consider it to be a 'hazard', rather than just an everyday event. However, without finding a difference between UK novice and experienced drivers, they could not firmly conclude that the hazard perception test could not transfer between countries (as they could not establish the effect in the UK in the first place with their clips).

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

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

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

1.3. The current experiments

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

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

2. Experiment 1

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

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

Key to this study was the requirement that the country-specific tests were as similar as possible in all other ways. Thus, all clips from each country were filmed and edited for this specific study, rather than co-opting previously captured video footage for inclusion. From the experience of filming in Malaysia, thought was given as to how best capture hazards that might not be suitable for the single-camera forward view favoured in the official UK and Australian tests. In order to accommodate the potential increase in overtaking hazards that may occur outside the UK, we used additional cameras attached to the film vehicle to record the views that one would see in the rear view mirror, and the two side mirrors. These video streams were then synchronised with the forward view and edited into mirror placeholders created in a graphic overlay of a car interior. Mirror information has been used

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

previously in hazard perception clips. For instance, Borowsky et al. (2012) included an inset rear-view mirror in their clips, though their clips did not require attention to the mirror information. Crundall et al. (2012a, b), and Shahar et al. (2010; 2012) included both side mirror and rear-view mirror information, inset into the forward view, in their hazard perception clips, though only the latter study required participants to use the mirror information to decide when it was safe to change lanes. The current study however combines mirror information with a graphic overlay to create a more immersive environment, providing precursors for hazards that appear from behind the film-car. We predicted that this test format would differentiate between experienced and inexperienced drivers in each country (using experience as a surrogate for crash likelihood). It was also considered likely that experience might interact with clip origin and participant nationality, such that UK experienced drivers may only out-perform UK inexperienced drivers on UK clips, etc. Such findings would at least demonstrate that the test format is culturally agnostic, if not the actual stimuli filmed in the three countries.

3. Method

3.1. Participants

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

Demographic details for each group can be found in Table 1. Over all three countries the average experienced driver was 29.5 years old, passed the driving test in 2005 (with 10 years of experience), and drove 11,804 miles per year. The average inexperienced driver was 21.1 years old, passed the driving test in 2014 and had an annual mileage of only 63 miles.

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

3.2. Materials and apparatus

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

Once the clips had been selected, the mirror footage was synchronised with the forward view and edited into mirror placeholders that were contained in a graphic overlay of the interior of a car. The graphic overlay was generated from internal photographs of a Ford Focus. The A-pillars and roof were designed to be semi-transparent, allowing the forward view to be seen through these sections of the overlay, although at a reduced fidelity. This was done to simulate the fact that real-world

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

Demographics	Chinese Participants		Spanish Participants		UK Participants	
	Novice	Exp'd	Novice	Exp'd	Novice	Exp'd
Study 1: Hazard Perception						
Total N (female N)	31 (13)	19 (2)	25 (17)	26 (8)	24 (13)	23 (21)
Age	22	28.6	19.2	35.9	22.1	24.1
Post-licence Experience (years)	1	6	1	17	1	7
Annual Mileage	82.8	4274.3	28	22041.9	78.2	9095.7
Study 2: Hazard Prediction						
Total N (female N)	24 (9)	26 (4)	25 (21)	27 (4)	27 (18)	23 (23)
Age	22.7	25.3	20.2	40.9	19.4	24.4
Post-licence Experience (years)	1	5	1	21	1	7
Annual Mileage	33.7	5474	28.9	20,183	266.7	5587

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

Clip Number	Hazards (*with occlusion points for experiment 2 italicised*)	Duration of the clip (ms)
	CHINESE CLIPS	
1	A pedestrian is visible at the right edge of the road, looking to cross. The pedestrian is obscured by a turning car at the point of stepping into the road. By the time the pedestrian is visible again, he has already stepped into the road becoming a hazard. *For experiment 2, the clip occludes just as the pedestrian starts to become visible as the obscuring car moves past.*	55000
2	There are parked cars on both sides of a narrow street that might occlude pedestrians. A pedestrian steps into the road in front of you, from between two parked vehicles. *The clip occludes as the pedestrian first becomes visible stepping out from between the parked cars.*	57000
3	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A cyclist emerges from the side road, obscured by the parked vehicles, and makes a wide turn in front of your vehicle, before cycling towards you. *The clip occludes as the front wheel of the bicycle enters the view.*	26000
4	Your car slows on approach to a junction. A cyclist approaches from the left and is partially obscured by the A-frame of the semi-transparent graphic overlay. The cyclist cuts directly across your path. *The clip occludes as the cyclist makes a change in direction to cut across your path.*	63000
5	Your car is driving slowly and there are parked cars on the right. A parked car on the right side of your lane indicates late before attempting to pull out in front of you. *The clip occludes following one flash of the indicator from the manoeuvring car.*	57000
6	A car immediately behind you, visible in the rear-view mirror and left side mirror, decides to overtake by entering a slip road to your left. It is forced to immediately pull back into your lane, in-front of you, as the slip road ends. *The clip occludes when the car is no longer visible in the left mirror, but flash of it is visible in the left window.*	31000
7	A car behind you is visible in the rear-view mirror and left side mirror. The car undertakes you on the left by entering a bus lane. Once past you, it cuts into your lane and is forced to brake suddenly due to slowing traffic ahead. *The clip occludes when the car is no longer visible in the right mirror, but it quickly appears next to the right window of your car.*	50000
8	A car behind you is visible in the rear-view mirror and left side mirror. While attempting to exit a multilane road, the car from behind accelerates to undertake your vehicle, forcing you to hold off moving into the desired lane. *The clip occludes when the car is no longer visible in the right mirror.*	62000
9	A lorry approaches fast from the right-hand side. The lorry enters the main road from a side road on the right, cutting into your lane. *The clip occludes at the moment in which the lorry is about to enter into the main road.*	38000
10	A car behind you is visible in the rear-view mirror and left side mirror. The car indicates with the front lights. Then undertakes via a bus lane at speed, immediately cutting in front of your vehicle and braking. *The clip occludes when the car is no longer visible in the right mirror.*	55000
	SPANISH CLIPS	
1	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A motorcyclist emerges from the side road, partially obscured by the parked vehicles, and enters the main carriageway immediately in front of your vehicle. *The clip occludes when the front part of the motor is visible.*	41000
2	A pedestrian is stood in the road next to a parked car waiting for her friend to exit the car. As your car approaches the driver's door of the parked vehicle opens. *The clip occludes when the pedestrians stops next to the car.*	42000
3	On entering a side road, a rider on a scooter is checking over her shoulder in order to pull out around a vehicle blocking her lane. She then pulls out in front of your vehicle, as you finish turning into the side road. *The clip occludes following one flash of the indicator of the scooter.*	34000
4	While travelling on a dual carriageway, in the distance a pedestrian enters from the left side of the road. The pedestrian continues to cross the street, forcing you to slow and stop. *The clip occludes at the moment when the pedestrian enters the road.*	37000
5	A car emerges from a side road and stops in front of you before indicating that it is going to reverse into a parking space at the road edge. *The clip occludes following one flash of the indicator from the car.*	53000
6	A van is approaching fast from a side road on the right. The van tries to pull out and it halts abruptly when already partially out of the road, but nevertheless forces you to brake suddenly. *The clip occludes when the van is approaching from the right and almost enters your lane.*	48000
7	A car ahead stops on a zebra crossing due to congestion ahead. A pedestrian, unable to cross on the actually crossing, steps into the road slightly in advance of the zebra crossing. As she steps out she is partially obscured by parked vehicles on the right. *The clip occludes just as the pedestrian first become visible.*	42000
8	A double-length ('bendy') bus in the right lane indicates and pulls off from a bus stop immediately in front of you, after you have just exited from a roundabout. *The clip occludes when the bus turns to enter the road and following a flash of the indicator of the bus.*	50000
9	While driving on a dual carriageway, a motorcycle undertakes in the right lane and is forced to pull in-front of your car as traffic in the right lane slows due to congestion. *The clip occludes when the motorcycle is no longer visible in the right mirror, but part of it is visible at the right window.*	26000
10	A pedestrian is approaching from the left, partially obscured by a pillar. The pedestrian crosses the road ahead from the left. *The clip occludes when the pedestrian is approaching the zebra crossing.*	47000
	UK CLIPS	
1	A car ahead overshoots a red traffic signal. As the cross traffic begins to enter the junction, the reversing light of the car turns on and the car ahead reverses towards you. *The clip occludes following initial illumination of the reversing light.*	64000
2	You are driving in a street with shops and parked vehicles on the left side. There is a pedestrian coming out of one of the shops, approaching the street. The pedestrian steps out from between two parked cars on the left just as you accelerate after waiting in standing traffic. *The clip occludes when the pedestrian turn his head to look at your car.*	49000
3	A distracted pedestrian is walking towards the street. The pedestrian crosses the road from the right without looking. *The clip occludes when the pedestrian approaches the road in order to cross.*	54000
4	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind undertakes you on the right on a multilane road. Once past you, it cuts into your lane and is forced to brake suddenly due to a red traffic light. *The clip occludes when the car is no longer visible at the left mirror, but a flash of it is visible on the left window.*	43000
5	There are pedestrians on both side of a narrow street. A pedestrian with a child's push chair enters the road from the right without looking. Her entrance is partially obscured by pedestrians standing on the right. *The clip occludes when the push chair is partially visible among the pedestrians.*	32000
6	A bus in a bus lane signals to pull away from a bus stop. Due to parked vehicles ahead in the bus lane, it pulls out into your lane forcing you to stop. *The clip occludes following one flash of the indicator from the bus.*	31000
7	A car ahead emerges from a side road on the left and crosses your lane (too far away to be considered a hazard). It then indicates again and immediately cuts across your lane once more to park in a layby. This second lane crossing is close enough to your vehicle to constitute a hazard. *The clip occludes just before the car starts to cross your lane and following one flash of the indicator of the car.*	42000

(*continued on next page*)

Table 2 (*continued*)

Clip Number	Hazards (*with occlusion points for experiment 2 italicised*)	Duration of the clip (ms)
8	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind overtakes your vehicle on a blind rural bend. The appearance of an oncoming vehicle in the opposite lane forces the overtaking vehicle to pull back into your lane immediately in front of you. *The clip occludes when the car is no longer visible at the rear-view mirror and the right mirror.*	34000
9	While travelling at speed along a country road, a blind bend ahead reveals a queue of standing traffic, forcing you to slow and stop. *The clip occludes immediately after passing the blind bend, when the brake lights of the cars ahead are partially visible.*	70000
10	While your car is slowing due to congestion ahead, a pedestrian looks over his shoulder before deciding to run in front of your vehicle forcing to slow more abruptly than otherwise required. *The clip occludes at the moment when the pedestrian is visible at the left pillar and is looking at your car.*	35000

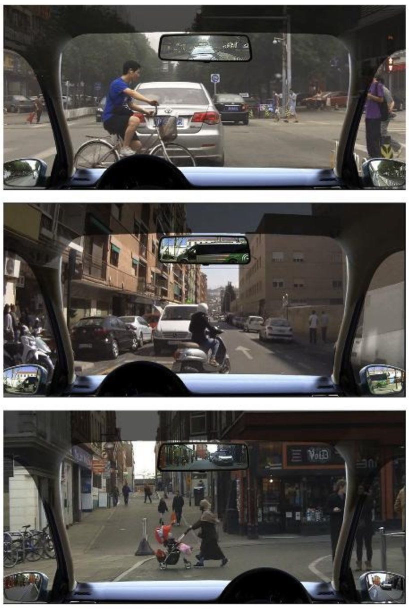


Fig. 1. Three screen shots taken from hazard perception clips filmed in China (top panel), Spain (middle panel) and the UK (bottom panel).

obscuration by A-pillars is offset somewhat by stereopsis and small head movements. Equally, head movements can often bring objects back into the visual field that are obscured by the roof (e.g. If one is first in a queue of vehicles at a red traffic signal, the nearest set of signals can easily be hidden by the roof, necessitating the driver to lean forward slightly to be able to look up at the red light). The dashboard, and mirror placeholders were however fully opaque. The final edited video clips created a seamless driving experience. When passing a car travelling in the opposite direction the vehicle would disappear briefly (into the driver's blind spot) before reappearing in the mirrors. Screen shots from each country can be viewed in Fig. 1.

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

Clips were displayed on a Lenovo (ThinkPad) computer with resolution of 1920×1080 and screen size of $34.5\,\text{cm} \times 19.5\,\text{cm}$ in all three countries and the programme used was E-Prime 2.0 Software (Psychology Software Tools, Inc. (2012)). Participants responded with a mouse connected to the laptop.

3.3. Design

A $2 \times 3 \times 3$ mixed factorial design was used. The between-group factors were the driving experience of participants (experienced vs. inexperienced) and their nationality (Chinese vs. Spanish vs. UK). The within-group factor was the clip origin (China vs. Spain vs. UK). The dependent variables included the percentage of hazards that participants correctly identified and their response times to these hazards.

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

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

3.4. Procedure

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

Participants were told that they would see 30 video clips from the driver's perspective, recorded in three different countries, and that each contained at least one hazardous situation. They were asked to view these clips as if they were the driver, and to press the mouse button as soon as they saw a hazard occurring. A hazard was defined as an object

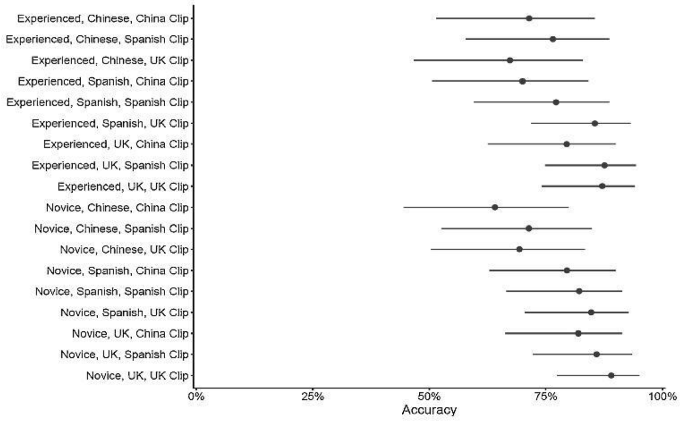


Fig. 2. The percentage of hazards correctly responded to by all participant groups across the three countries of origin for the clips (with error bars).

or event in the road environment that could increase the risk of a collision if an evasive manoeuvre such as braking or steering was not performed (following Crundall, 2016). It was made clear that participants did not need to locate the hazard on the screen using the mouse, but merely had to press the button to record their response. After each clip they were asked to rate how hazardous they thought that particular situation was on a scale from 1 to 7 by pressing the corresponding button on the keyboard (1 = not at all hazardous; 7 = extremely hazardous). Before the start of the experiment each participant saw a practice clip from their own country in order to familiarise themselves with the task. If the participant failed to perform on the practice task as expected, the experimenter explained the instructions again. In total the experiment took an average of 35 min.

4. Results

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

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

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

4.1. Response accuracy to hazards

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

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

However there was a main effect of participant nationality, ΔG^2 (2) = 25.5, p < .001. On average accuracy was 70.1%, 95% CI [59.3, 79.1], for Chinese drivers, 80.3%, 95% CI [71.6, 86.7], for Spanish drivers and 85.4% 95% CI [78.3, 90.5] for UK drivers. Follow-up tests indicate that differences between all three nationalities were statistically significant.[3] The odds ratio (*OR*) of the difference between Chinese and Spanish drivers was 0.58, p < .005, 95% CI [0.41, 0.81], whilst between Chinese and UK drivers it was, 0.40, p < .005, 95% CI [0.28, 0.57], narrowing to 0.69, p < .005, 95% CI [0.49, 0.99], between Spanish and UK drivers. Thus the data suggest a clear pattern of differences between drivers of different nationalities – one that reflects the rank order of these countries in road safety according to the World Health Organisation; WHO, 2015).

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

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

4.2. Response times to hazards

In order to calculate response times (RTs) for the hazards, hazard onsets and offsets were defined for each clip. Hazard onset times for each clip were subtracted from button-press times to give the RTs. Where participants failed to make a response during a particular clip, they were assigned a maximum response time plus 1 ms. The resulting data were therefore right-censored and, as is common with response times, positively skewed. To address these features of the data a multilevel regression model treating clips and participants as random factors and modelling response times as a right-censored lognormal distribution was employed. This is superior to standard approaches to such data that treat censored data as missing or treat censored responses as known with certainty. A 2 × 3 × 3 factorial model was fitted as a Bayesian model using the probabilistic programming language Stan (Carpenter et al., 2017) via the R package brms (Bürkner, 2017) using the weakly informative default priors. As brms does not provide frequentist likelihood ratio tests, models were compared using the information criteria WAIC (Vehtari et al., 2018) and followed up using 95% Bayesian central posterior probability intervals.

For an intercept only model (with only random effects) WAIC was 56559.7 decreasing to 56553.5 (ΔWAIC = 6.2) for a model with all main effects – indicating a substantial improvement in fit.[4] Fig. 4 shows the predicted mean response times by condition. The two way model (WAIC = 56551.6) provided only a modest additional improvement in fit (ΔWAIC = 1.9). Accordingly only main effects were followed up further.

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

4.3. Extra hazard responses

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

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

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

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

Fig. 6 helps clarify the patterns of extra hazard responses. Panel (a) indicates that the lower rate of hazard responses for Chinese participants relative to UK and Spanish participants is obtained across all clip types but is particularly pronounced for clips of Chinese origin. In contrast, panel (b) reveals that the lower hazard response rate for UK clips is observed regardless of experience, but that the difference between the Chinese and UK clips is larger for the experienced than inexperienced drivers. Simple main effects for the Chinese versus Spanish

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

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

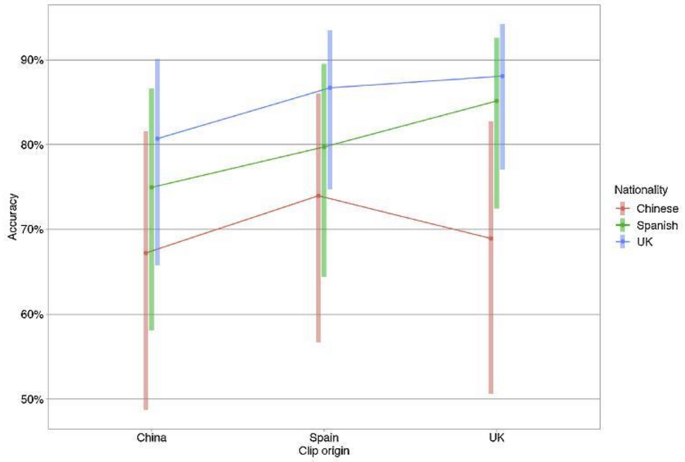


Fig. 3. The percentage of hazards correctly responded for each block of clips across the three groups of participants (with error bars).

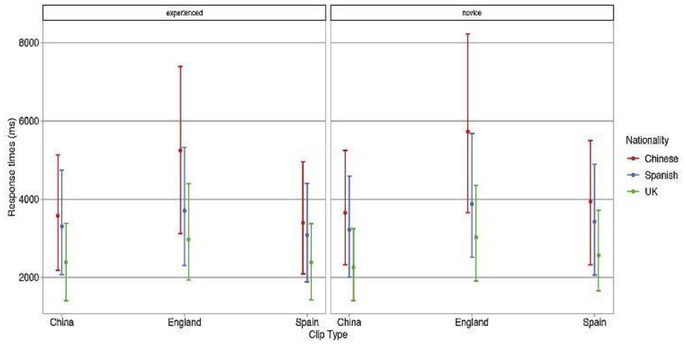


Fig. 4. Response times across all participants nationality, experience groups and 3 clips sets recorded in each country (with error bars).

differences are not significant for either experienced or novice drivers (both $p > .05$).

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

4.4. Hazardousness ratings for all hazard clips

Following each clip, participants were asked to provide a hazardousness rating on a scale of 1–7 (where higher numbers reflect greater levels of perceived hazardousness). These ratings were analysed with a three-way factorial design (participant experience vs. nationality vs. clip origin) using multilevel ordinal logistic regression via the ordinal package (Christensen, 2018). The intercept only model (G^2 (8) = 14,226) was a worse fit than a model with all main effects (ΔG^2 (5) = 40.0, $p < .0001$), which in turn was a worse fit than a model with all two-way interactions (ΔG^2 (8) = 40.8, $p < .0001$). Adding the three-way interaction did not further improve the model (ΔG^2 (4) = 1.8, $p = .78$).

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

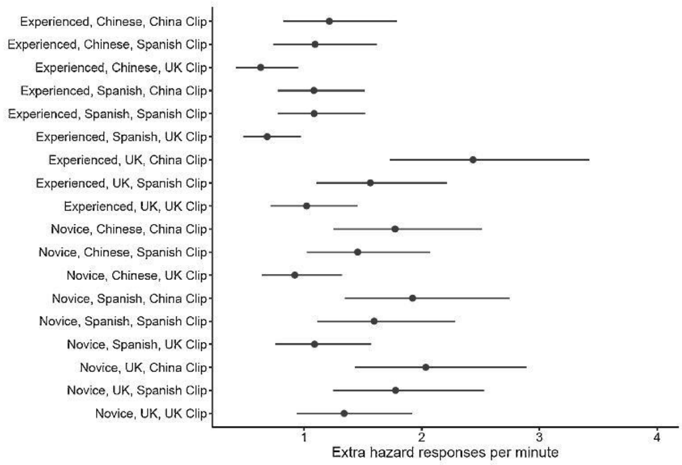


Fig. 5. Extra hazard responses across all participant groups and 3 sets of clips (with error bars).

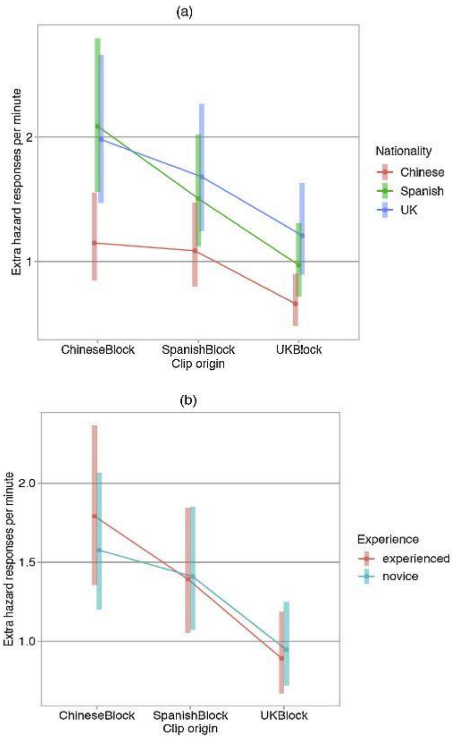


Fig. 6. Extra hazard responses across clip origin vs nationality (with error bars).

final factor of driver experience did not reveal a difference between the two groups (4.2 vs. 4.1 for experienced and inexperienced drivers, respectively), G^2 (1) = 1.4, p = 0.23.

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

5. Discussion

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

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

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

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

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

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

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

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

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

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

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

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

6. Experiment 2

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

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

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

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

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

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

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

7. Method

7.1. Participants

A hundred and fifty-three participants took part in Experiment 2. The sample was composed of 50 Chinese, 52 Spanish and 51 UK drivers. One participant was later excluded (from the UK sample) due to difficulties categorising the individual as experienced or inexperienced. All of the participants held a full or a learner-driver licence from their country. Participants were split into two sub-groups of experienced and inexperienced drivers following the method used in Experiment 1. In China we recruited 26 experienced drivers (mean age of 25.3, an average of 5 years of post-licensure experience, and a mean annual mileage of 5474 miles) and 24 inexperienced drivers (mean age of 22.7, an average of 1 year of post-licensure experience, and a mean mileage of 33.7 miles). In Spain we recruited 27 experienced drivers (mean age of 40.9, an average of 21 years of post-licensure experience, and a mean mileage of 20,183 miles) and 25 inexperienced Spanish drivers (mean age of 20.2, an average of 1 year of post-license mean experience and mean mileage of 28.9 miles). In the UK 23 experienced UK drivers were recruited (mean age of 24.4, an average of 7 years of post-license mean experience and mean annual mileage of 5587 miles), along with 27 inexperienced UK drivers (mean age of 19.4, an average of 1 year of post-license experience, and a mean annual mileage of 266.7). Across all countries, the mean age of experienced drivers was 30.2 years, with an average of 11 years of post-licensure experience, and they had driven an average of 10,415 miles in the previous year, while inexperienced drivers had a mean age of 20.8, with an average of 1 year of post-licensure experience, and had driven an average of 109.6 miles.

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

7.2. Materials and apparatus

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

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

7.3. Design and procedure

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

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

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

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

8. Results

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

To test whether there were differences in the accuracy of hazard prediction performance across the factors, the $2 \times 2 \times 3$ factorial design was analysed as in Experiment 1 using a multilevel logistic regression with participant and clip as random factors. The between-groups factors were the experience level of drivers (experience vs. inexperienced drivers) and their nationality (Chinese vs. Spanish vs. UK). The within group factor was the origin of the clip (China vs. Spain vs. UK). An intercept only model (with no predictors) estimated the *SD* of the participant random effect as 0.804 and the *SD* of the clip random effect as 1.278 indicating that only 28% variation at level 2 of the model is attributable to participants – with variability in clips accounting for the majority (72%) of level 2 variance. This indicates that a traditional ANOVA analysis – that treats variation between clips as zero – would substantially underestimate standard errors. The deviance (likelihood ratio Chi Square, G^2) for the intercept only model was 5023.8 and decreased to 5012.9 for a model including main effects of nationality, experience and clip type. This improvement in model fit was not statistically significant, ΔG^2 (5) = 9.9, $p = .077$. However, G^2 decreased dramatically for a model with all two-way interactions, ΔG^2 (8) = 24.2, $p < .002$, with a negligible improvement with the addition of the three-way interaction, ΔG^2 (4) = 2.8, $p > .05$. The two-way interaction model therefore appears to be the most informative.

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

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

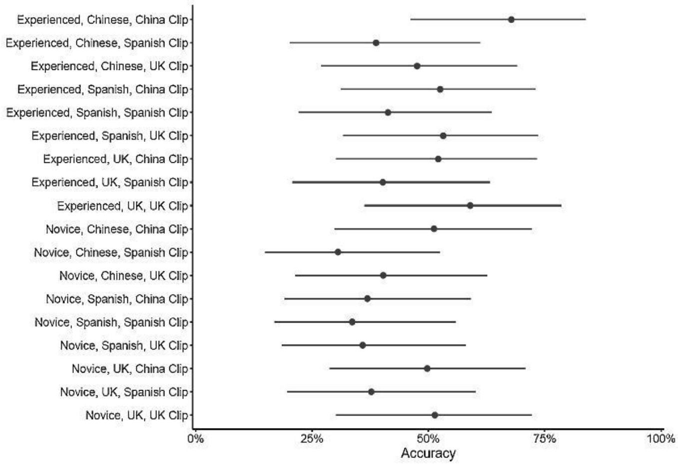


Fig. 7. Percenatges of accuracy in prediction across driving groups, participant's nationality and clips (with error bars).

8.1. *Hazardousness ratings for hazard prediction*

As with the hazard perception study, participants were asked to provide a hazardousness rating on a scale of 1–7 (where higher numbers reflect greater levels of perceived hazardousness). Again a multilevel ordinal logistic regression was used to analyse the ratings using a factorial design (with experience, nationality and clip origin as factors). The intercept only model (G^2 (8) = 15,252) was a worse fit than a model with all main effects (ΔG^2 (5) = 29.3, $p <$ .0001), which in turn was a worse fit than a model with all two-way interactions (ΔG^2 (8) = 35.4, $p <$ .0001). Adding the three-way interaction only marginally further improved the model (ΔG^2 (4) = 5.8, $p =$ .21).

There was a main effect of clip origin, G^2 (2) = 8.8, $p =$ .012. Chinese clips were rated as more hazardous than Spanish Clips (4.26 vs. 3.73) with a smaller difference between the Spanish clips and the UK clips (3.73 vs. 3.63). Hochberg corrected post hoc tests revealed significant differences between the Chinese and Spanish (adjusted $p =$ .025) and UK clips (adjusted $p =$ .008) but not between Spanish and UK (adjusted $p =$. 61). Nationality also produced a significant main effect, G^2 (2) = 20.2, $p <$ .0001, with Spanish drivers giving the highest ratings (4.24) followed by UK drivers (4.02), with Chinese drivers giving the lowest hazard ratings (3.33). Hochberg corrected pairwise tests revealed Chinese drivers to give lower ratings than both Spanish and UK drivers (adjusted $p <$ .0001 and $p <$ .002 respectively), with no difference between Spanish and UK drivers ($p =$ .26).

In regard to driver experience, though experienced drivers were not found to give significantly different ratings to those provided by novices overall (3.86 vs. 3.88; G^2 (1) = 0.2, $p =$ .64), driver experience did

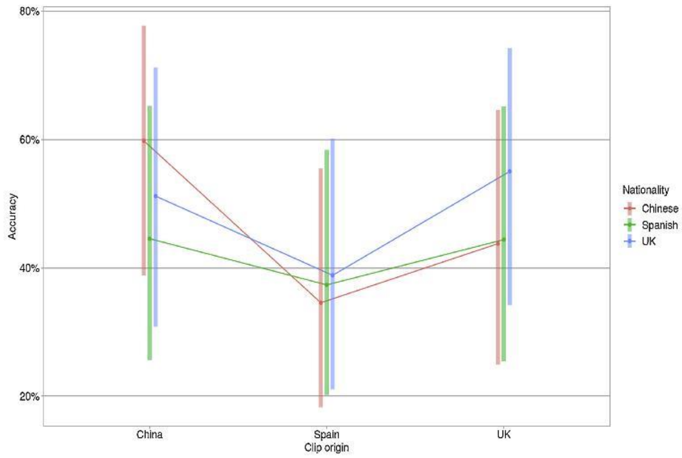


Fig. 8. Accuracy in prediction across different nationality vs clip origin (with error bars).

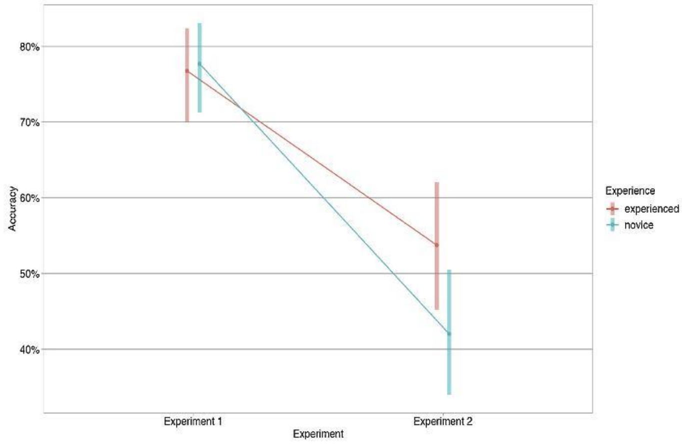


Fig. 9. Percentages of prediction accuracy for test type across experience (with error bars).

interact with clip origin (G^2 (2) = 12.8, p = .002). Both driver groups rate Chinese clips as more dangerous than Spanish clips, but this effect is more pronounced in the experienced driver group. There was also an interaction between clip origin and nationality, G^2 (5) = 15.1, $p <$ .005. This largely followed the pattern obtained for the main effects except that the UK participants rated Spanish clips more hazardous than UK clips (adjusted $p <$ .05).

9. Discussion

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

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

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

10. Comparison of the two tests

It is possible to directly compare the performance on the two tests, using the prediction accuracy from experiment 2 and the percentage of hazards that received a timed button response in experiment 1 (though note that we cannot claim that all responses that fell in the scoring window in the hazard perception test were referencing the actual hazard – this is one of the problems with the traditional HP methodology). In the analysis reported below we only focus on the main effect of *test type* (whether accuracy scores differ across the hazard *perception* and hazard *prediction* tests), and any emerging interactions with test type.

Accuracy rates for the two tests were compared with a $2 \times 2 \times 3 \times 3$ factorial design using a multilevel logistic regression model with the factors of test type (Perception vs. Prediction), participant experience (inexperienced vs. experienced), participant nationality (Chinese, Spanish or from the UK), and clip origin (China, Spain, UK). For these models the main effects only model was more informative than the intercept only model, ΔG^2 (6) = 198.1, $p <$ .0001, with the two-way interaction model improving model fit further still, ΔG^2 (13) = 76.7, $p <$ .0001. The three-way interaction model also offered additional improvement in fit, but was not significantly better than the two-way model, ΔG^2 (12) = 20.8, $p >$ .05, with almost no change in fit after adding the four-way interaction, ΔG^2 (4) = 1.4, $p >$ .05. As the focus here is on differences between the tests we report only tests of effects comparing *hazard perception* to *hazard prediction.*

The results showed that there was a main effect for the type of test, G^2 (1) = 168.0, $p <$ .0001. Participants scored higher on average for the hazard perception test compared to the hazard prediction test

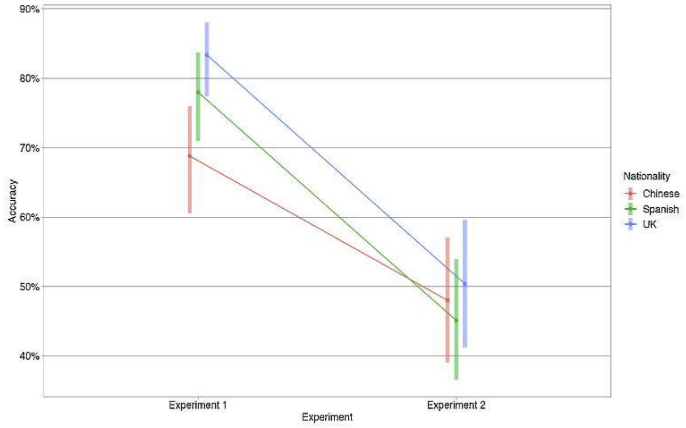


Fig. 10. Percentage of prediction accuracy for test type across nationality (with error bars).

(77.2% vs. 47.8%). A significant interaction was also found for test-type and experience, G^2 (1) = 168.0, $p < .0001$. Despite the prediction test appearing more difficult than the perception test, it is clear that the benefit of experience only holds for hazard prediction rather than hazard perception (see Fig. 9).

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

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

11. General discussion

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

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

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

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

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

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

Crucially, the clips edited for the hazard prediction test were the same as those used in the hazard perception test, allowing a direct comparison of the two tests. This is the first time that the hazard perception and hazard prediction tests have been directly compared in a single analysis[5], though this was complicated by the fact that the two tests record very different primary measures: response times and percentage accuracy, respectively. However, as the hazard perception test required response times to fall within a temporal scoring window around the appearance of the hazard, the presence or absence of a response allowed the calculation of an accuracy score that could be compared to the hazard prediction test.

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

11.1. Do different countries produce different hazards?

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

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

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

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

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

11.2. The limitations of hazard prediction tests

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

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

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

If drivers' individual thresholds for reporting a hazard are considered important enough to warrant assessment (and we believe they

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

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

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

12. Conclusions

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

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

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

References

Allen, J., Walsh, J.A., 2000. A construct-based approach to equivalence: methodologies for cross-cultural/ multicultural personality assessment research. In: Dana, R.H. (Ed.), Handbook of Cross-Cultural and Multicultural Personality Assessment. Personality in Clinical Psychology Series. Lawrence Erlbaum Associates, Mahway, NJ, pp. 63–85.

Baguley, T., 2012. Serious Stats: a Guide to Advanced Statistics for the Behavioural Sciences. Palgrave Macmillan, London.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48.

Barnard, Y., Utesch, F., van Nes, N., Eenink, R., Baumann, M., 2016. The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. Eur. Transp. Res. Rev. 8 (14), 1–10.

Borowsky, A., Oron-Gilad, T., Meir, A., Parmet, Y., 2012. Drivers' perception of vulnerable road users: a hazard perception approach. Accid. Anal. Prev. 44, 160–166.

Borowsky, A., Shinar, D., Oron-Gilad, T., 2010. Age, skill and hazard perception in driving. Accid. Anal. Prev. 42 (4), 1240–1249.

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

Bürkner, P.-C., 2017. brms: an R Package for Bayesian multilevel models using Stan. J. Stat. Softw. 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. J. Stat. Softw. 76 (1). https://doi.org/10.18637/jss.v076.i01.

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

Castro, C., Ventsislavova, P., Peña-Suárez, E., Gugliotta, A., García-Fernández, P., Eisman, E., Crundall, D., 2016. Proactive listening to a training commentary improves hazard prediction. Saf. Sci. 82, 144–154. https://doi.org/10.1016/j.ssci.2015.09.018.

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

Chapman, P., Underwood, G., Roberts, K., 2002. Visual search patterns in trained and untrained novice drivers. Transp. Res. Part F Traffic Psychol. Behav. 5, 157–167.

Cheng, A.S.K., Ng, T.C.K., Lee, H.C., 2011. A comparison of the hazard perception ability of accident-involved and accident-free motorcycle riders. Accid. Anal. Prev. 43 (4), 1464–1471.

Christensen, R.H.B., 2018. Ordinal - regression models for ordinal data. R Package Version 2018. pp. 4–19.

Cocron, P., Bachl, V., Früh, L., Koch, I., Krems, J.F., 2014. Hazard detection in noise-related incidents - the role of driving experience with battery electric vehicles. Accid. Anal. Prev. 73, 380–391.

Crundall, D., 2016. Hazard prediction discriminates between novice and experienced drivers. Accid. Anal. Prev. 86, 47–58. https://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. Accid. Anal. Prev. 42 (6), 2117–2124.

Crundall, D., Chapman, P., Phelps, N., Underwood, G., 2003. Eye movements and hazard perception in police pursuit and emergency response driving. J. Exp. Psychol. Appl. 9 (3), 163–174. https://doi.org/10.1037/1076-898X.9.3.163.

Crundall, D., Crundall, E., Clarke, D., Shahar, A., 2012a. Why do car drivers fail to give way to motorcycles at t-junctions? Accid. Anal. Prev. 44 (1), 88–96.

Crundall, D., Chapman, P., Trawley, S., Collins, L., van Loon, E., Andrews, B., Underwood, G., 2012b. Some hazards are more attractive than others. Accid. Anal. Prev. 45, 600–609.

Crundall, D., Kroll, V., 2018. Prediction and perception of hazards in professional drivers: Does hazard perception skill differ between safe and less-safe fire-appliance drivers? Accid. Anal Prev. 121, 335–346.

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

Deery, H.A., 1999. Hazard and risk perception among young novice drivers. J. Safety Res. 30 (4), 225–236.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Peres, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knipling, R.R., 2006. NHTSA report DOT HS 809 593. The 100-car Naturalistic Driving Study, Phase II: Results of the 100-car Field Experiment. NHTSA, Washington, US.

Drummond, A.E., 2000. Paradigm lost! Paradigm gained? An Australian's perspective on the novice driver problem. Proceedings of the Novice Driver Conference (retrieved 18.01.09). http://www.dft.gov.uk/pgr/roadsafety/drs/novicedrivers/conference/theaustralianperspectiveonth4667.

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

Endsley, M.R., 1988b. Situation awareness global assessment technique (SAGAT). Proceedings of the National Aerospace and Electronics Conference. IEEE., New York, pp. 789–795.

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

Fitts, P.M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 47, 381–391.

Foss, R.D., Martell, C.A., Goodwin, A.H., O'Brien, N.P., Center, U.H.S.R., 2011. Measuring Changes in Teenage Driver Crash Characteristics During the Early Months of Driving. AAA Foundation for Traffic Safety, Washington, DC.

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

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

Groeger, J.A., 2000. Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task. Psychology Press, New York, NY.

Hertzum, M., Hornbaek, K., 2010. How age affects point with mouse and touchpad: a comparison of young, adult and elderly users. Int. J. Hum. Interact. 26 (7), 703–734.

Horswill, M.S., Anstey, K.J., Hatherley, C.G., Wood, J.M., 2010. The crash involvement of odler drivers is associated with their hazard perception latencies. J. Int. Neuropsychol. Soc. 16 (5), 939–944.

Horswill, M.S., Hill, A., Wetton, M., 2015. Can a video-based hazard perception test used for driver licensing predict crash involvement? Accid. Anal. Prev. 82 (23), 213–219.

Horswill, M.S., Marrington, S., McCullough, C.M., Wood, J., Pachana, N.A., McWilliam, J., 2008. The hazard perception ability of older drivers. The journals of gerontology. Series B, Psychol. Sci. Soc. Sci. 63 (4), 212–218.

Horswill, M.S., Taylor, K., Newnam, S., Wetton, M., Hill, A., 2013. Even highly experienced drivers benefit from a brief hazard perception training intervention. Accid. Anal. Prev. 52, 100–110. https://doi.org/10.1016/j.aap.2012.12.014.

International Test Commission (ITC), 2010. International Test Commission Guidelines for Translating and Adapting Tests. Download From. [retrieved, July 26, 2012]. http://www.intestcom.org&gt.

Isler, R.B., Starkey, N.J., Williamson, A.R., 2008. Video-based road commentary training improves hazard perception of young drivers in a dual task. Accid. Anal. Prev. 41, 445–452.

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

Judd, C.M., Westfall, J., Kenny, D.A., 2012. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. J. Pers. Soc. Psychol. 103 (1), 54–69.

Lim, P.C., Sheppard, E., Crundall, D., 2013. Cross-cultural effects on drivers' hazard perception. Transp. Res. Part F Traffic Psychol. Behav. 21, 194–206.

Lim, P.C., Sheppard, E., Crundall, D., 2014. A predictive hazard perception paradigm differentiates driving experience cross-culturally. Transp. Res. Part F Traffic Psychol. Behav. 26, 210–217.

Malone, S., Brünken, R., 2016. The role of ecological validity in hazard perception assessment. Transp. Res. Part F Traffic Psychol. Behav. 40, 91–103.

McCartt, A.T., Shabanova, V.I., Leaf, W.A., 2003. Driving experience, crashes and traffic citations of teenage beginning drivers. Accid. Anal. Prev. 35 (3), 311–320.

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

McDonald, C.C., Goodwin, A.H., Pradhan, A.K., Romoser, M.R.E., Williams, A.F., 2015. A review of hazard anticipation training programs for young drivers. J. Adolesc. Health 57 (Supplement), S15–23 1.

McKenna, F.P., Crick, J.L., 1991. Hazard Perception in Drivers: a Methodology for Testing and Training. Final Report. Transport Research Laboratory, Crowthorne, UK.

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

Padilla, A.M., Medina, A., 1996. Cross-cultural sensitivity in assessment: using tests in culturally appropriate ways. In: Suzuki, L.A., Meller, P.J., Ponterotto, J.G. (Eds.), Handbook of Multicultural Assessment: Clinical, Psychological, and Educational Applications. Jossey-Bass Publishers., San Francisco, pp. 3–28.

Parmet, Y., Meir, A., Borowsky, A., 2014. Survival analysis: a fresh approach for analyzing response times in driving-related hazard perception paradigms. Transp. Res. Part F Traffic Psychol. Behav. 25, 98–107.

Pelz, D.C., Krupat, E., 1974. Caution profile y driving record of undergraduate males. Accid. Anal. Prev. 6, 45–58.

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

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

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

Core Team, R., 2018. R: a Language and Environment for Statistical Computing (Version 3.2.5) [computer Program]. R Foundation for Statistical Computing, Vienna, Austria.

Renge, K., 1998. Drivers' hazard and risk perception, confidence in safe driving, and choice of speed. Int. Assoc. Traffic Saf. Sci. Res. 22, 103–110.

Rosenbloom, T., Perlman, A., Pereg, A., 2011. Hazard perception of motorcyclists and car drivers. Accid. Anal. Prev. 43 (3), 601–604.

Sagberg, F., Bjørnskay, T., 2006. Hazard perception and driving experience among novice drivers. Accid. Anal. Prev. 28 (2), 407–414.

Scialfa, C.T., Pereverseff, R.S., Borkenhagen, D., 2014. Short-term reliability of a brief hazard perception test. Accid. Anal. Prev. 73, 41–46.

Shaffer, J., 1986. Modified sequentially rejective multiple test procedures. J. Am. Stat. Assoc. 81 (395), 826–831.

Shahar, A., Alberti, C.F., Clarke, D., Crundall, D., 2010. Hazard perception as a function of target location and the field of view. Accid. Anal. Prev. 42 (6), 1577–1584.

Shahar, A., Van Loon, E., Clarke, D., Crundall, D., 2012. Attending overtaking cars and motorcycles through the mirrors before changing lanes. Accid. Anal. Prev. 44 (1), 104–110.

Shimazaki, K., Ito, T., Fujii, A., Ishida, T., 2017. Improving drivers' eye fixation using accident scenes of the HazardTouch driver-training tool. Transp. Res. Part F Traffic Psychol. Behav. 51, 81–87.

Thomas, F.D., Rilea, S.L., Blomberg, R.D., Peck, R.C., Korbelak, K.T., 2016. Evaluation of the Safety Benefits of the Risk Awareness and Perception Training Program for Novice Teen Drivers (Report No. DOT HS 812 235). National Highway Traffic Safety Administration., Washington, DC.

Underwood, G., Chapman, P., Bowden, K., Crundall, D., 2002. Visual search while driving: skill and awareness during inspection of the scene. Transp. Res. Part F Traffic Psychol. Behav. 5 (2), 87–97.

Underwood, G., Ngai, A., Underwood, J., 2013. Driving experience and situation awareness in HP. Saf. Sci. 56, 29–35.

Vehtari, A., Gabry, J., Yao, Y., Gelman, A., 2018. Loo: efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.0.0.

Ventsislavova, P., Crundall, D., 2018. The hazard prediction test: a comparison of free-response and multiple-choice format. Saf. Sci. 109, 246–255.

Ventsislavova, P., Gugliotta, A., Peña-Suarez, E., Garcia-Fernandez, P., Eisman, E., Crundall, D., Castro, C., 2016. What happens when drivers face hazards on the road? Accid. Anal. Prev. 91, 43–54.

Vlakveld, W.P., 2011. Hazard Anticipation of Young Novice Drivers. Proefschrift Rijksuniversiteit Groningen. SWOV-Dissertatiereeks. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV, Leidschendam.

Vlakveld, W.P., 2014. A comparative study of two desktop hazard perception tasks suitable for mass testing in which scores are not based on response latencies. Transp. Res. Part F Traffic Psychol. Behav. 22, 218–231. https://doi.org/10.1016/j.trf.2013.12.013.

Wahlström, J., Svensson, J., Hagberg, M., Johnson, P.W., 2000. Differences between work methods and gender in computer mouse use. Scand. J. Work Environ. Health 26 (5), 390–397.

Wallis, T.S.A., Horswill, M.S., 2007. Using fuzzy signal detection theory to determine why experienced and trained drivers respond faster than novices in a hazard perception test. Accid. Anal. Prev. 39, 1177–1185.

Wang, Y., Peng, P., Liang, L.J., Zhang, W., Wu, S., 2007. Road hazard reaction testing using driving simulation: The novice vs. The experienced drivers. Helander, M., Xie, M., Jaio, M., Tan, K.C. (Eds.), 2007 IEEE International Conference on Industrial Engineering and Engineering Management.

Watts, G.R., Quimby, A.R., 1979. Design and Validation of a Driving Simulator for Use in Perceptual Studies. Report 907. Crowthorne. Transport Research Laboratory, UK.

Wells, P., Tong, S., Sexton, B., Grayson, G., Jones, E., 2008. Cohort II: A Study of Learner and New Drivers. Department of Transport, DOT, London, UK.

Wetton, M.A., Hill, A., Horswill, M.S., 2011. The development and validation of a hazard perception test for use in driver licensing. Accid. Anal. Prev. 43, 1759–1770.

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

Williams, A.F., Tefft, B.C., 2014. Characteristics of teens-with-teens fatal crashes in the United States, 2005–2010. J. Safety Res. 48, 37–42.

World Health Organisation, 2015. Global Status Report on Road Safety, 2015. (last accessed on 19/12/16). http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/.

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

Yamauchi, T., Seo, J.H., Jett, N., Parks, G., Bowman, C., 2015. Gender differences in mouse and cursor movements. Int. J. Hum. Interact. 31 (12), 911–921.

Yeung, J.S., Wong, Y.D., 2015. Effects of driver age and experience in abrupt-onset hazards. Accid. Anal. Prev. 78, 110–117.

Zhai, S.M., Kong, J., Ren, X.S., 2004. Speed-accuracy tradeoff in Fitts' law tasks – on the equivalency of actual and nominal pointing precision. Int. J. Hum. Stud. 61 (6), 823–856.

APPENDIX D: PUBLISHED ARTICLE: Ventsislavova & Crundall (2018)

Safety Science 109 (2018) 246–255

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

The hazard prediction test: A comparison of free-response and multiple-choice formats

Petya Ventsislavova*, David Crundall

Department of Psychology, School of Social Sciences, Nottingham Trent University, UK

ARTICLE INFO

Keywords:
Hazard perception
Hazard prediction
Multiple-choice format
Driving

ABSTRACT

Hazard perception skill is often related to lower crash risk, and the hazard perception test has been widely employed to measure this ability in drivers. An increasingly popular test-variant is the hazard prediction test: driving videos are occluded immediately prior to a hazard and participants are asked to predict how the situation will develop. Early versions of this test asked participants to provide a free-response answer which was subsequently coded. Later versions, however, have used a multiple-choice format where participants are provided with four options presented on screen. While the benefits of a multiple-choice format are obvious in terms of providing immediate feedback without relying on subjective coding, it is unclear whether this change in format affects the discriminative validity of the test. For the current study, a free-response test and a multiple-choice test were created using the same video clips. The free-response test (experiment 1) was found to successfully discriminate between novice and experienced drivers, with the latter predicting more hazards correctly. The answers provided by participants in Experiment 1 were then used to generate the options for a multiple-choice test (experiment 2). This second test was also found to discriminate between novice and experienced drivers, and a comparison between the two tests failed to reveal an advantage for one over the other. Despite this, correlations between prediction accuracy and both years of post-license driving, and annual mileage, were only significant for the multiple-choice test. The results suggest that the multiple-choice format is not only time- and cost-efficient, but is ostensibly as good as the free-response test in discriminating between driver groups.

1. General introduction

Hazard perception refers to the skill of detecting on-road dangers in sufficient time to avoid a collision. Following 50 years of research in this field, there is now a general consensus that the hazard perception (HP) skill is related to crash risk (Horswill, 2016). The hazard perception test (HPT) has been widely employed to investigate this skill within the field of traffic and transport psychology, and traditionally involves the presentation of video clips from a driver's perspective, filmed from a moving vehicle. Hazards appear during the clips (e.g. a pedestrian may step into the road, a car may emerge from a side street, etc.), and participants are required to press a button as soon as they spot the danger (McKenna and Crick, 1991; Scialfa et al., 2011; Wetton et al., 2011). It has been documented over many studies that novice drivers and crash-involved drivers are slower, and less likely, to detect hazards in these clips than safer or more experienced drivers (e.g. Cheng et al., 2011, Horswill et al., 2010, McKenna and Horswill, 1999, Rosenbloom et al., 2011). Since the first known hazard perception test (Spicer, 1964, cited in Pelz and Krupat, 1974) many different versions have been developed for research purposes. In addition to the traditional speeded, push-button responses required by the typical hazard perception test, some tests have used levers or sliding scales for participants to indicate a level of hazardousness (Watts and Quimby, 1979; Crundall et al., 2003). Other test variants have required drivers to locate the hazard via a mouse click or a touch screen response (Wetton et al., 2010, 2011). An increasingly popular version requires drivers to predict imminent hazards following sudden occlusion of the video clips (often termed the 'What Happens Next?' test, or the hazard prediction test), measuring HP skill in terms of prediction accuracy (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009).

Tests also differ in the medium chosen to present hazards to participants. The majority of research-based tests employ video clips of driving situations, filmed from a camera mounted on a moving vehicle to record the driver's view of the road. Other variants range from the use of computer-generated imagery to create clips (as introduced in the official UK hazard perception test in 2015), to the use of static road images (as used in the official hazard test in the Netherlands). Even those tests that adopt the more traditional video-based approach can

* Corresponding author.
E-mail addresses: petya.petrova@ntu.ac.uk (P. Ventsislavova), david.crundall@ntu.ac.uk (D. Crundall).

https://doi.org/10.1016/j.ssci.2018.06.004
Received 19 January 2018; Received in revised form 5 June 2018; Accepted 11 June 2018
Available online 18 June 2018
0925-7535/

differ in the way they present these clips: some are presented across multiple screens (e.g. Shahar et al., 2010), and have mirror information available to participants (e.g. Ventsislavova et al., submitted), while others present a single forward view (Horswill et al., 2015). Other differences across tests include variations in instructions given to participants (e.g. Farrand and McKenna, 2001), the method of analysis (for example, the contentious issue of dealing with missing values; Parmet and Borowsky, 2014), the nature of the hazards (Crundall, 2016) and even the definition of what constitutes a hazard (e.g. Pradhan and Crundall, 2017; Crundall et al., 2012).

Given that the primary aim of any HP test is to discriminate between safe and unsafe drivers (often on the basis of surrogate measures such as previous crash history, or driving experience), then policy makers might be unconcerned about the particular design of a specific test, providing it successfully separates these driver groups. Certainly, at its best, hazard perception research has demonstrated both retrospective and prospective sensitivity to crash-likelihood, and even offers great hope that HP training may produce on-road improvements in driver safety (Chapman et al., 2002; Pradhan et al., 2009, Horswill, 2016; Thomas et al., 2016). However, the research field is littered with examples of failed attempts to discriminate between crash-involved, inexperienced drivers and their safer, more-experienced counterparts (e.g. Crundall et al., 1999; Lim et al., 2013; Sagberg, and Bjørnskau, 2006; Underwood et al., 2013; Yeung and Wong, 2015). Given the huge variety of test designs across research groups it has proved difficult to understand why some tests are successful and others are not.

For this reason, we argue that test design should be developed through empirical research, with each facet compared and analysed to assess whether it contributes to the validity of the test. This process has already begun in some research groups. For instance, Scialfa et al. (2013) correlated response time performance to hazards presented in static images and dynamic clips (though their results were inconclusive in identifying which was the better test).

Other researchers have noticed that their stimuli can produce different effects in participants depending on the nature of the hazards. Zimasa et al. (2017) found that their clips containing vulnerable road users elicited quicker response times than those involving cars. This distinction appears particularly relevant when comparing typical drivers to individuals on the autistic spectrum, with the latter showing reduced sensitivity for such 'social' hazards (Bishop-Johnson et al., 2017; Sheppard et al., 2010).

The underlying structure of hazards has also been explored. Crundall (2016) compared 'environmental prediction' and 'behavioural prediction' hazards, and found the former to better discriminate between novice and experienced drivers. While 'behavioural prediction' hazards can be predicted by the actions of the soon-to-be hazard (e.g. the erratic driving of the car ahead), 'environment prediction' refers to hazards that appear out of obscuration (e.g. an oncoming car from around a blind bend, a pedestrian from behind a parked truck). In these scenarios, the environment is the only clue to the possible upcoming hazard (i.e. the blind bend, the parked truck). These types of hazards have been a particular focus of the Risk Awareness and Perceptual Training programme designed by Fisher et al. (Fisher et al., 2010; Pradhan et al., 2009), where drivers are trained to spot occluded objects that may hide hazards. Similar levels of detailed analysis are also being given to the design of HP training programmes in Australia (Horswill et al., 2017; Wetton et al., 2013).

Despite these studies, more research is needed to provide the basic blueprint for a valid hazard perception test. While much of a test's validity is likely to lie with the content (i.e. the particular hazards that form the stimuli), there are many finer points of test design that may provide significant discriminative gains. These may include the method of presentation, the instructions given, and the required responses, to name but a few.

2. The hazard prediction test

The hazard prediction test differs from the traditional hazard perception test in that it shuns response times in favour of accuracy for predicting what happens next following an occlusion that occurs just as the hazard begins to develop. It is argued that this test format removes many of the potential problems associated with recording response times to hazards (Crundall, 2016). For instance, response-time measures require a scoring window to be defined. If a response is made between the onset and offset of a hazard, then the response is considered to be correct. However, there are no clear guidelines on how to define onsets and offsets, and there is always the possibility that excellent drivers will spot very subtle cues to upcoming hazards, and respond just before the scoring window (which would be counted as a miss). Even if drivers do press within the scoring window, we do not know if they are responding to the actual hazard, or to some other less hazardous aspect of the scene (see Crundall, 2016, for an argument as to why localised hazard responses are not a suitable solution for a lack of accuracy in the traditional test). Finally, 'hazard perception' is confounded by post-perceptual processes, such as criterion bias: expert drivers may delay or refrain from responding to hazards because they believe the unfolding event falls within the boundaries of their driving skill (Pradhan and Crundall, 2017). The hazard prediction test (or 'What happens next?' test) mitigates these confounds by removing reliance on response times, replacing them with the accuracy of drivers to predict what happens next following occlusion of the developing hazard.

A number of studies have demonstrated the ability of the hazard prediction test to successfully discriminate between safer, experienced drivers, and less-safe, inexperienced drivers (Jackson et al., 2009; Crundall, 2016; Castro et al., 2014, 2016, Ventsislavova et al., 2016, Gugliotta et al., 2017; Lim et al., 2014). Several of these studies have also developed this occlusion-based methodology through a number of targeted experiments focusing on design elements. For instance, Jackson et al. (2009) demonstrated that an occlusion is necessary to discriminate driver groups, rather than just pausing on the final frame. A freeze-frame provides an unrealistic amount of time for novice drivers to identify clues to the impending hazard, whereas an occlusion ensures that the driver must be looking in the right place at the right time. As safer drivers are more likely to prioritise those areas of the scene that may develop into hazards, the occlusion is therefore more likely to identify the safest drivers (Crundall and Kroll, 2018).

Crundall (2016) addressed a number of methodological questions, including the impact of clip length on predictive accuracy. He found that longer clips resulted in lower prediction accuracy, especially for novice drivers, suggesting that novices suffer a greater vigilance decrement over time. In a separate experiment, Crundall manipulated the occlusion point. The results demonstrated a decline in prediction accuracy as the occlusion point became more temporally distant from the hazard. The novice/experienced driver distinction remained however and did not interact with the occlusion point. Thus it seems that hazards can be extrapolated from relatively early information (in this case, over a second prior to hazard onset), though at a reduced level of accuracy. Participants' confidence ratings in their predictions fall to baseline levels however at the most distal occlusion points.

These initial studies suggest that the hazard prediction test can provide a robust and simpler alternative to the more traditional hazard perception test. However, one of the problems with the version of the test used by many researchers (Castro et al., 2014, 2016; Crundall, 2016; Jackson et al., 2009) is that participants give free-response answers which must be hand coded. This introduces the potential for rater error, and renders the test impractical for use on a wide scale. An alternative is to provide the participants with multiple options to choose from following occlusion, instead of inviting a verbal or typed response. This approach simplifies the test further and allows for automatic and unambiguous coding. This variant of the hazard prediction test was first employed by Lim et al. (2014), and was followed by Ventsislavova et al. (2016), but there has never been a direct comparison between a multiple-choice test variant and hand-coded free responses. This paper aims to make that comparison, in order to assess the impact of using the (more practical) multiple-choice method, compared to free responses.

To this end, a new hazard prediction test was developed. This new test uses full HD video from four cameras attached to a moving vehicle. These four video streams are then synchronised and combined with a graphic overlay of a car interior, to provide the viewer with access to mirror information. We believe that providing mirror information may evoke more realistic responses to hazards, even when the hazard never appears in the mirrors (Alberti et al., 2014). Novice and experienced drivers were then tested on this new prediction task in order to assess whether it could discriminate between these groups. This first study (Experiment 1) required free responses to the question 'What happens next?', which is more typical of studies in this area. The second study (Experiment 2) presented the same videos, but participants were provided with 4 options to choose from at the end of each clip. These options were developed from the free-response answers provided in the first study. Finally, a matched sub-group of participants from study 1 and 2 were compared in terms of their prediction accuracy with the only difference being the response mode: free-response or selection of an answer from four options. We hypothesised that both test variants would discriminate between drivers to a certain degree, but we made no prediction about which test would be the better discriminator in the final analysis.

3. Experiment 1: A free-response hazard prediction test

3.1. Introduction

With both the improvement of camera technology and our evolving approach to the presentation of hazards to participants, it is inevitable that new tests will be developed and will require validating. The current stimuli differed from those used in Crundall (2016) both in terms of HD quality and the amount of information provided to participants. For instance, the current test includes mirror information that provides a more realistic level of perceptual demand on the driver, and also allows for hazards to approach from behind (see Fig. 1).

To validate this new version of the test we recruited experienced and inexperienced drivers to take part. In addition to the basic validation of this test, we took the opportunity to revisit one of the questions raised by Crundall (2016): when should one occlude the clip to maximise discriminability? As noted above, Crundall found that occlusions that were proximal to the hazard resulted in the most accurate predictions. While the most distal occlusions (an average of 1200 ms prior to hazard onset) significantly reduced accuracy, participants were still able to predict the hazard on 57% of trials (with a free response). Surprisingly, the change in occlusion point did not affect the discriminability of the driver groups. One might suppose that experienced drivers would be relatively better than novices at more distal occlusions, as they may make better use of weaker hazard evidence (Pradhan and Crundall, 2017). However, as occlusions become extremely separated from the hazard we might further predict that all drivers would reach the same nadir, in the absence of even subtle cues to the nature of the upcoming hazard. Neither of these eventualities occurred in the study of Crundall (2016), possibly due to the particular occlusion points that were chosen and the nature of the individual hazards in those clips. Given the need to validate the new free-response version of our prediction test on experienced and inexperienced drivers, we also had the opportunity to investigate whether a novel set of stimuli produced different findings when the occlusion points were varied across three levels: proximal (temporally closest to the hazard), intermediate, and distal occlusion points.

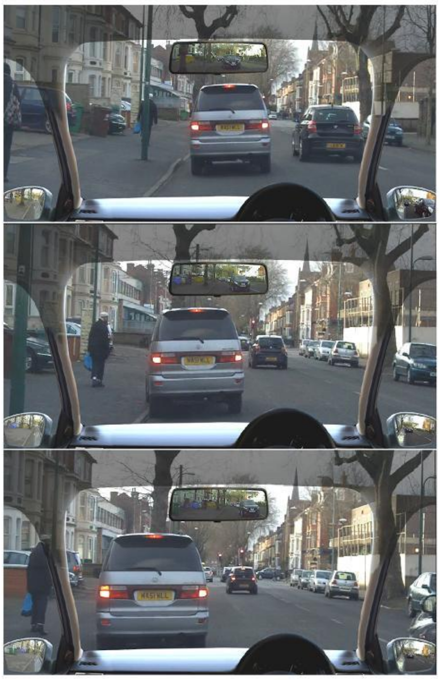


Fig. 1. Three frames illustrating the three occlusion points for one clip. While driving along a road, the car in front pulls over. This requires the film car to overtake the stopping vehicle. However, a safe driver should be aware that there is a vehicle visible in the right side mirror that might pose a problem (top panel; distal occlusion). As the clip progresses, the car in the side mirror disappears, as it moves into the blind spot ready to overtake. Thus, the evidence that this vehicle is going to pose a problem increases considerably (middle panel; intermediate occlusion). Finally, the overtaking car becomes visible in the forward view confirming that the car from behind is indeed overtaking and poses a hazard for the film car's intended action (bottom panel; proximal occlusion).

3.2. Method

3.2.1. Participants

Sixty-one participants took part in this experiment. The sample was divided into experienced and inexperienced drivers. Experienced drivers (N = 30, mean age = 21.5 years) had a minimum of 1 year of driving experience since passing their driving test and inexperienced drivers (N = 31; mean age = 19.5 years) had less than 1 year of post-test driving experience, and included 14 learner drivers. Mean mileages for experienced and inexperienced drivers were 4553.3 miles and 290.9 miles in the year prior to the study, respectively. All participants had normal or corrected-to-normal vision and were University students and staff.

3.2.2. Design

A 2 × 3 between-group factorial design was employed, where the independent variables were the driving experience of participants (experienced vs. inexperienced) and the occlusion points (proximal vs. intermediate vs. distal). Five cells contained 10 participants, with the distal/inexperienced group containing 11 participants. The proximal occlusion point was the closest to the onset of the hazard, and would typically show several frames of the hazard following onset. This meant that participants who were looking in the right place at the right time, would have their predictions confirmed. For instance, if the participant believed the pedestrian ahead may step into the road, a fixation on this pedestrian immediately prior to occlusion (proximal condition only) would confirm this, as the pedestrian would be seen to move a leg in the direction of the road. Intermediate occlusion points were an average of 618 ms earlier than the proximal occlusion points, and would contain evidence of precursors, but without any confirmation of the hazard triggering. In this case, the pedestrian might be seen walking towards the edge of the pavement, perhaps turning her head in the direction of the film car. Finally, the distal occlusion points were an average of 1222 ms earlier than the proximal points. Typically, these would still contain precursors to the hazard, but the evidence of hazardousness would be weak. For instance, the pedestrian may be visible on the pavement, but still be walking parallel to the road with no obvious intention to cross. See Fig. 1 for an alternative example of these three occlusion points.

The dependent variable was the percentage of the accuracy with which participants correctly identified the hazards across a total of 15 clips. Following occlusion, participants were presented with a free-text entry box to provide their answer. These typed responses were subsequently coded with one point awarded for a correct answer. Participants were instructed to consider the identity of the hazardous object, it's location within the scene, and how the event would unfold (e.g. "the pedestrian on the left is about to step into the road in front of me"). Where participants failed to report the three suggested items in their answer, but it was still unambiguously correct, they were still awarded the point. For instance, the response "The pedestrian is about to step into the road" could still be awarded the point if this was the only visible pedestrian, even though the response failed to locate the hazardous object as being on the left-side pavement. Two independent evaluators scored each answer. Cohen's kappa was 0.88 for the distal occlusion point, 0.88 for the intermediate occlusion point, and 0.80 for the proximal occlusion point. Where raters disagreed in their initial rating, a final score was agreed via discussion.

3.2.3. Materials

In order to create the hazard prediction test, new video stimuli were recorded on roads in and around Nottingham (UK). Four HD mini cameras were attached to the car. The first camera was attached on the inside of the front window, with the second camera attached on the inside of the rear window to obtain the footage for the rear-view mirror. The two remaining cameras were attached externally to the driver and passenger side windows to capture side-mirror footage. All clips were recorded in the daytime during fine weather conditions. We extracted a selection of hazard perception clips from the footage, each containing a naturally occurring hazard as judged by two traffic and transport psychologists (see Table 1 for a description of the hazards). Only fully developed hazards were selected (i.e. we did not select a clip with a pedestrian who *looked* like she might step into the road; we only selected the clip if she *did* step into the road, and at a close enough distance to be hazardous, as assessed by our judges). The duration of each clip varied between 31 s and 64 s prior to editing the occlusions.

Following selection, the four video streams were synchronised, with the three rear-facing video streams edited into mirror placeholders contained in a graphic overlay of the inside of a car. The A-pillars and roof were designed to be semi-transparent, allowing the forward view to be seen through these sections of the overlay and the dashboard and mirror placeholders were fully opaque. The reason behind the semi-transparent design is that real-world obscuration by A-pillars is offset somewhat by stereopsis and small head movements. A fully opaque overlay could therefore over-exaggerate the obscuration that a car frame causes.

The clips were edited to occlude (i.e. cut to an immediate black screen) at three separate points: proximal, intermediate and distal from the fully developed hazard. A proximal cut would allow several frames of the actual hazard to be seen (providing the participant was already looking at the location of the imminent hazard), confirming the participant's prediction. Intermediate occlusions and distal occlusions occurred an average of 618 and 1222 ms earlier than the proximal occlusion, respectively.

Clips were displayed on a Lenovo (ThinkPad) computer running E-Prime 2.0 Software (Psychology Software Tools, 2012), with a resolution of 1920 × 1080 and screen size of 34.5 cm × 19.5 cm. Participants responded using the keyboards and the mouse.

3.2.4. Procedure

Once participants gave consent, they were seated 60 cm from the screen and asked to answer demographic questions, including age, sex, year of obtaining driving license, driving collisions in the past 24 months, and miles driven in the past 12 months. Participants were then given both verbal and on-screen instructions for the test. They were told that they would watch 15 short video clips from a driver's perspective. They were asked to watch each clip carefully, as at some point the clip would stop and be occluded by a black screen. Following this, they were asked to describe "What happens next?" (i.e. how the driving situation was going to develop), by typing their answer in a free-response box on the screen (with a 150 character limit). They were instructed to include in their answer any source of potential hazard, the location of that source, and how the situation was going to develop. Following each answer, participants were asked to rate how hazardous they thought the predicted event would be for them, using a 7-point Likert scale (where 1 is 'not hazardous at all', and 7 is 'extremely hazardous').

Participants then watched a practice trial where they had the opportunity to familiarise themselves with the task. This practice trial also included feedback: following occlusion and their answer, the clip was replayed without an occlusion allowing participants to see what the hazard actually was. This practice trial was the only time when they received feedback about their performance. Once participants felt comfortable with the instructions and the practice, they began the experiment. In total, the experiment took an average of 20 min.

3.3. Results

All participants were given a score out of 15 clips for the accuracy of their predictions. The scores were based on the interpretation of the typed free responses by two expert raters. Cohen's Kappa for each condition was acceptable (= > .8), with disagreements reconciled through discussion.

Participants' scores were converted to percentages and subjected to a 2 × 3 Analysis of Variance (ANOVA; experienced vs. inexperienced drivers × 3 levels of occlusion point). A main effect was found for driving experience with experienced drivers being more accurate when predicting hazards than inexperienced (54.9% vs. 45.6%; $F(1,55) = 7.13$, $MSe = 161.2$, $p < .01$, $\eta_p^2 = .11$). A main effect was also found across all levels of occlusion point, with the distal, intermediate and proximal conditions producing average accuracy rates of 28.6%, 50.7%, and.72.3%, respectively ($F(2,55) = 60.6$, $MSe = 161.2$, $p < .001$, $\eta_p^2 = .69$). Planned repeated contrasts were conducted comparing distal to intermediate, and intermediate to proximal occlusion points. Both were significant (*ps* < .001).

Despite a visible trend to suggest that the distal condition is the weakest discriminator of the driver groups (see Fig. 2), no significant interaction was found between experienced group and occlusion point ($F(2,55) = .94$, $MSe = 151.2$, $p = .39$, $\eta_p^2 = .03$).

Ratings of hazardousness were also subjected to a 2 × 3 ANOVA but no differences were found across experience or occlusion points. The average hazard rating of experienced and inexperienced drivers for the

Table 1
A description of the *a priori* hazards selected within each clip.

N°	Video clips	Duration(ms)	Last sketch prior the clip occlusion	Hazard
1	A car ahead overshoots a red traffic signal. As the cross-traffic begins to enter the junction, the reversing light of the car ahead illuminates, and this car then reverses towards you to get out of the junction. The clip occludes following initial illumination of the reversing light	64,000		Reversing car ahead
2	As you are driving, the car indicates and pulls over blocking your path. As you try to overtake the car and continue your path, a car from behind overtakes you (passing from the right mirror, into the blind spot, before emerging into your forward view). The clip occludes at the point where the overtaking car enters the blindspot	34,000		A car from behind overtakes you
3	You are driving along a narrow urban street and the head of a pedestrian is visible above the parked cars on the left. The clip occludes as the pedestrian moves between the parked cars to step into the road	34,000		The pedestrian from the left crosses the street
4	You come to a standstill due to congestion in an urban street with shops and parked vehicles. A pedestrian exits a shop and then steps out from between two parked cars on the left, just as the traffic moves and you begin to accelerate. The clip occludes when the pedestrian turns his head to look at you, while stepping forward	49,000		A pedestrian from the left crosses the street
5	A distracted pedestrian is walking towards the street. The pedestrian crosses the road from the right without looking. The clip occludes as the pedestrian begins to cross	54,000		A distracted pedestrian crosses the street from the right
6	A car behind you is visible in the rear-view mirror and left side mirror. As you approach traffic lights, this car undertakes you and pulls in front of you, forcing you to brake. The clip occludes when the car is just emerging from the blind spot	43,000		The car from the left undertakes the camera car
7	There are pedestrians on both side of a narrow street. A pedestrian with a child's pushchair enters the road from the right without looking. Her entrance is partially obscured by pedestrians standing on the right. The clip occludes when the pushchair is partially visible among the pedestrians	32,000		A pedestrian with a child's pushchair enters the road from the right without looking
8	You are driving along a narrow urban road. A pedestrian approaches from the right with shopping bags. He is temporally obscured by a large plant pot, before he steps out into the road. The clips occludes just as he begins to step out from behind the plant pot	46,000		A pedestrian from the right crosses the street without looking for upcoming cars
9	A bus in a bus lane signals to pull away from a bus stop. Due to parked vehicles ahead in the bus lane, it pulls out into your lane forcing you to stop. The clip occludes following one flash of the indicator from the bus	31,000		The bus pulls out into your lane
10	An oncoming car cuts across your lane to park in a layby on your left. The clip occludes just before the car starts to cross your lane, following one flash of its indicator	42,000		The car in front cuts across your lane
11	A car behind you is visible in the rear-view mirror and right-side mirror. The car from behind overtakes your vehicle on a blind bend. The appearance of an oncoming car in the opposite lane forces the overtaking vehicle to pull back into your lane immediately in front of you. The clip occludes when the overtaking car begins to emerge from your blind spot, and the oncoming car becomes obvious	34,000		The car behind overtakes on a blind bend
12	While travelling at speed along a country road, a blind bend ahead reveals a queue of standing traffic, forcing you to slow and stop. The clip occludes immediately after passing the blind bend, when the brake lights of the cars ahead are partially visible	70,000		Queue of standing traffic
13	While your car is slowing due to congestion ahead, a pedestrian looks over his shoulder before deciding to run in front of your vehicle forcing you to slow more abruptly than otherwise required. The clip occludes at the moment when the pedestrian looks directly at you, as he steps forward	35,000		A pedestrian crosses the street from the left
14	Driving through a school area, two pedestrians are visible on the right. You can see that the woman is looking at the car. The clip occludes at the moment she gesticulates with her arm, thinking that you have given way for her and her child to cross	42,000		Pedestrians from the right cross the street after indicating to the camera car
15	You are driving along a narrow road with parked cars on both sides, which are obscuring the view ahead. Suddenly an oncoming van appears ahead. The clip occludes at the moment the white van is partially visible	46,000		A van appears suddenly behind parked cars

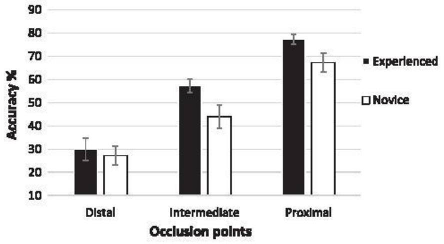


Fig. 2. Percentage of correct responses across all 3 occlusion points and experienced groups (with standard error bars added).

distal clips was 3.46 and 3.40, respectively. For the intermediate clips, experienced drivers gave a rating of 3.76 while inexperienced drivers rated them at 3.72. Finally, for the proximal clips, the ratings for experienced and inexperienced drivers were 3.25 and 4.04.

3.4. Discussion

The primary finding of the current experiment is that this new version of the hazard prediction test, using a new set of clips from those used by Crundall (2016), has successfully discriminated between experienced and inexperienced drivers. With appropriate occlusion points, the clips successfully elicit more accurate responses from experienced drivers compared to inexperienced drivers.

The occlusion-point factor ostensibly produced identical findings to those of Crundall (2016), with declining accuracy as the temporal separation between occlusion and hazard increases. Crundall (2016) failed to find an interaction, and the current interaction between experience and occlusion point also failed to reach significance. However, a closer look at the data is suggestive that the ability of clips to discriminate between the driver groups at the most distal level of occlusion is noticeably degraded. While the data do not support any strong conclusions, there is a clear suggestion that cutting the clip too early reduces the link between precursors and hazards to such an extent that all drivers approach a floor. Note however that the overall accuracy to the distal condition in the current study is far below the 57% score in the comparable condition of Crundall (2016). This suggests that the distal occlusions in the current study are cut more severely in relation to the actual hazards. This cannot be due to absolute differences in distal occlusion points across the Crundall study (2016) and the current one (1200 ms vs. 1222 ms), so must therefore be due to differences in unfolding nature of the hazard precursors. It is likely that the lag between first evidence of a hazard, and the actual hazard beginning to develop, was greater in the Crundall (2016) clips, than in the current selection.

In regards to hazardousness ratings, no differences were noted across either factors, mirroring the results of Jackson et al. (2009), who also failed to find significant differences regarding the hazardousness of the future situation. A possible explanation for this is that drivers are basing their ratings more on the nature of the scene that they *do* see (for more than 30 s prior to occlusion), rather than the predicted hazard itself which predominantly remains unseen.

In conclusion, the current study has validated the new set of prediction clips by demonstrating an experiential difference in predictive accuracy. Furthermore, we have demonstrated that the choice of occlusion point has limited influence upon the discriminability of the test, within reasonable boundaries. If set too early however, none of the participants are likely to score highly due to the removal of all precursors to the upcoming hazard.

An additional advantage of the current study however is that the free responses provided by participants offer a range of incorrect predictions that can be developed into multiple-choice distracter answers for Experiment 2. The reduced accuracy of the intermediate and distal conditions was very useful in this regard, as the greater number of incorrect responses was concomitant with a greater variety of response, providing us with sufficient plausible answers to provide three distracter options for each clip in the following study.

4. Experiment 2: A multiple-choice hazard prediction test

4.1. Introduction

There have been several previous attempts to create a multiple-choice version of the hazard prediction test. Lim et al. (2014) used a multiple-choice hazard prediction test with Malaysian and UK drivers. Their results showed that experienced drivers were able to predict more hazards in comparison to the novices (though only for the Malaysian clips). While this result was in contrast to a previously unsuccessful study using response-time hazard perception clips (Lim et al., 2013), they did not directly compare the two datasets across the studies. More recently, Ventsislavova et al. (2016) created a multiple-choice Spanish prediction test in order to test the differences between experienced and novice drivers, asking both “What was the hazard” and “What happens next?”. They found that experienced drivers were more accurate in identifying and predicting the hazard than novices.

Two recent studies have also compared multiple-choice prediction tests directly to response-time tests (Crundall and Kroll, 2018; Malone and Brünken, 2016). Malone and Brünken (2016) described their two test variants as differing on ecological validity, with the response-time test representing high ecological validity, while the multiple-choice version represented low ecological validity. Even though they expected that the high ecological validity test would discriminate better between experienced groups, the multiple-choice version of the hazard perception test also yielded a significant difference between experienced and novice drivers. It should be noted however that Malone et al. used hazards that fully materialised. As participants saw the whole hazard, this test did not isolate the act of prediction, and is therefore less applicable to the current discussion. Crundall and Kroll (2018) compared a response-time HP test and a multiple-choice prediction test across groups of fire appliance drivers (with clips filmed from fire appliances on blue-light training runs). They found the multiple-choice test to be more sensitive to group differences.

Multiple-choice formats have also been employed in the training of hazard perception (Cockerton and Isler, 2003; Isler and Cockerton, 2003; Petzoldt et al., 2013). Petzoldt et al. (2013) compared a paper-based intervention with computer-based training using a multiple-choice format, with the latter being more beneficial for participants to spot earlier critical cues and scan relevant areas in the visual field.

One concern that has been raised regarding the development of multiple-choice tests relates to the plausibility of the distracters (Andrich and Marais, 2014). Poor or implausible distracter options can be easily rejected allowing the participant to choose the correct response by default. For this reason, we followed the guidelines of Worthen et al. (1999), creating the distracter items using the incorrect answers that were provided by the participants in Experiment 1. In order to test this new format, we divided our sample into four groups: learners, novices, moderately experienced drivers and highly experienced drivers. We predicted that there would be clear differences in prediction accuracy across our driving groups.

4.2. Method

4.2.1. Participants

Fifty-one participants were split into four different experienced groups: 12 learner drivers (L), 10 novice drivers (N), 15 drivers that have driven for less than 6 years (E < 6y) and 14 drivers that have driven more than 6 years (E > 6y) (see Table 2 for demographics). All participants had normal or corrected-to-normal vision and were

Table 2
Mean age, mileage and driving experience for the different driving groups.

Participants	Mean age	Mean driving experience since test (in years)	Mean mileage in the previous 12 months
Learner drivers	18.3	0	9
Novice drivers	18.4	11 months	555
E < 6 years	21.4	3 years	4933.3
E > 6 years	35.9	16 years	7857.1

identified as University students and staff.

4.2.2. Design, materials and procedure

The proximal occlusion clips from Experiment 1 were edited to include a multiple-choice question, with the most frequent incorrect answers from the previous experiment chosen as the three distracters for each clip. Both the correct answer and all distracter options were written in short, simple sentences, following guidelines provided by Haladyna (1999). All apparatus remained the same as that used in Experiment 1.

A 1 × 4 between-subjects design was employed, where the independent variable was the level of experience of the participants and the dependent variable was the percentage accuracy of hazards predicted. The procedure was identical to that used in Experiment 1 with the only difference being that this time participants were asked to predict the driving situation by choosing one of 4 options after each of 15 clips (see Fig. 3). Participants were aware that only one option would be correct and, if they had been looking in the correct place at the time of occlusion, that they would have seen very brief confirmation that their prediction was correct. Following their selection of an option, participants were asked to rate how hazardous the predicted situation would have been for them, by selecting a point on a 1–7 Likert Scale (where 1 is 'not at all hazardous' and 7 is 'extremely hazardous'). Prior to the actual experiment, all participants saw a practice trial. The experiment took an average of 20 min to complete.

4.3. Results

Accuracy was analysed using a 1 × 4 ANOVA, where the between-group factor was driver experience. A main effect was found (see

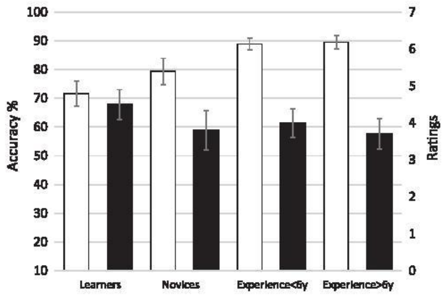


Fig. 4. Percentage accuracy (white bars) and ratings (from 1 to 7; black bars) across all experienced groups: learners, novices, experienced drivers with less than 6 years of driving experience, and experienced drivers with more than 6 years of driving experience. Standard error bars are added.

Fig. 4), with performance improving in line with increasing experience across the four groups (71.7%, 79.3%, 88.9% and 89.5% for learners, novices, moderately experienced and highly experienced drivers, respectively; $F(3,47) = 6.9$, $MSe = 134.2$, $p < .001$, $\eta_p^2 = .31$). Planned contrasts were undertaken to unpack this main effect. The first contrast pooled the learners and novices and compared them to the pooled experienced driver groups (mirroring the analysis in experiment 1 which just compared inexperienced drivers to experienced drivers). The difference was found to be significant with the experienced drivers outperforming the learner/novice group ($t(47) = 4.2$, $p < 0.001$). Repeated contrasts (learners vs. novices, novices vs. moderately-experienced drivers, and moderately-experienced vs. highly-experienced drivers) were also conducted to identify whether the test was sensitive to more subtle changes in experience, but the only significant comparison was found between novices and moderately-experienced drivers ($t(47) = 2.0$, $p < 0.05$). This suggests a clear experiential boundary in the ability of the test to discriminate between groups at the transition between novice and moderately-experienced drivers.

Ratings were also analysed and a main effect was found for driving experience across the four groups ($F(3,47) = 2.9$, $MSe = .85$, $p < .05$,

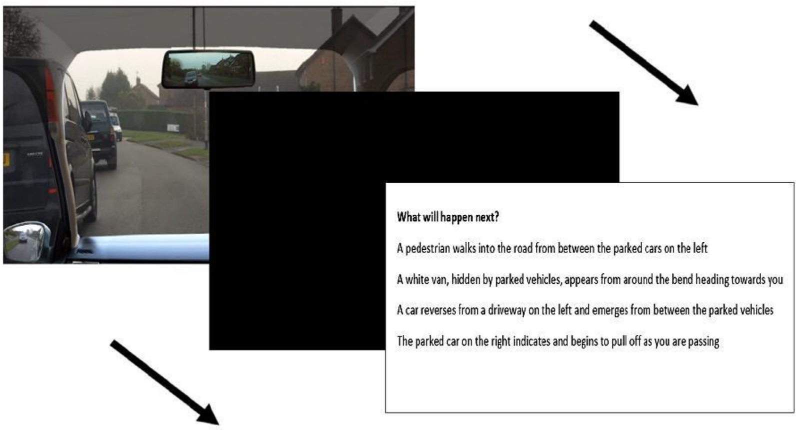


Fig. 3. Three panels represent the end of a clip: the first panel is the final frame (where the impending hazard is an oncoming white van, hidden by parked vehicles). This is followed by an occlusion panel, before four multiple-choice options are displayed on screen.

$\eta_p^2 = .16$). While learners gave the highest risk ratings for the clips (4.5) and the highly-experienced group gave the lowest ratings (3.7), the experiential effect was not sufficient for any of the planned contrasts (L/N vs. all experienced drivers; L vs. N; N vs. E < 6 yr; E < 6 yr vs. E > 6yrs) to reach significance (see Fig. 4).

4.3.1. Comparison of the two tests

A third additional analysis was conducted in order to directly compare the two tests and investigate whether there were differences between the test types in terms of accuracy and discrimination. For that purpose, all the participants that took part in the proximal occlusion condition of Experiment 1 (N = 20; 10 novices and 10 experienced) were matched to participants with the most similar driving experience (in terms of passing the driving test and annual mileage) from Experiment 2 (N = 20; 10 novices and 10 experienced). The average difference between matched pairs in terms of mileage was 76 miles for the novices and 400 miles for the experienced drivers, while the average difference in years' experience since passing the driving test was of 0.4 years for the novice drivers and 0.2 for the experienced.

A 2 × 2 between-groups ANOVA was conducted on the two data sets. A marginal difference was found between the tests ($F(1,36) = 3.5$, MSe = 152.8, $p = .07$, $\eta_p^2 = .09$), with performance on the multiple-choice test being ostensibly higher than that on the free-response version. The main effect of driving experience remained with experienced drivers outperforming novices (82.3 vs. 69.7; $F(1,36) = 10.5$, MSe = 152.8, $p < .01$, $\eta_p^2 = .23$), although there was no interaction between the test type and the experienced groups (see Fig. 5). No differences were found for the ratings. The hazards were not rated as more hazardous in terms of test type. These results suggest that the tests do not differ regarding their ability to discriminate on the basis of driving experience, and that the type of question does not influence perceived hazardousness.

Pearson correlations were conducted to assess the relationship between the accuracy of responses, annual mileage and years of driving (post-license) for both types of test. Significant correlations were found for accuracy and years of driving ($r(18) = 0.51$, $p < .05$), and accuracy and miles driven ($r(18) = 0.56$, $p < .05$), but only with the data from the multiple-choice test. Neither of these correlations was significant for the free-response data from Experiment 1 (see Table 3).

5. Discussion

The multiple-choice hazard prediction test was able to discriminate between the driving groups on the basis of experience. Experienced drivers (E < 6y and E > 6y) outperformed the learner and novice drivers in predicting hazards. Planned contrasts revealed that the threshold for performance differences lies between the novice and moderately-experienced group, suggestive of a step change in the ability to predict hazards following 1–2 years of post-license driving experience.

This is in keeping with the literature that consistently reports novice

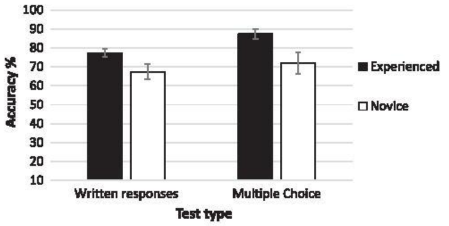


Fig. 5. Percentage accuracy of correct responses across experience groups and test type (with standard error bars added).

Table 3
Pearson correlations for accuracy vs. years of driving and accuracy vs. miles for both types of tests (multiple-choice and free-response type).

	Multiple-choice test			Free-response test		
	Accuracy	Years	Miles	Accuracy	Years	Miles
Accuracy	–			–		
Years	.512 (p < .05)	–		.382 (p = .096)	–	
Miles	.562 (p < .01)	.697 (p < .01)	–	.395 (p = .085)	.531 (p < .05)	–

drivers, especially within the first 12 months of post-license driving, to be overrepresented in crashes. This suggests that safety-relevant skills are still developing in this first year of independent driving (Foss et al., 2011; McCartt et al., 2003; Williams and Tefft, 2014; Pradham and Crundall, 2017).

When overall accuracy rates for the two tests are examined, the multiple-choice test is ostensibly easier (though this effect did not reach conventional levels of statistical significance; $p = 0.07$). Such an effect can be understood in terms of the opportunity for participants to guess, or infer, the correct answer out of the multiple-choice options. Arguably, an act of inference may still tap into hazard prediction skill (as participants may choose the most likely response on the basis of the evidence they had gathered up to the point of occlusion). Pure guesses are potentially more problematic, but as the ostensible increase in accuracy did not negatively impact on the ability of the test to discriminate between our driver groups, we argue that it is of little consequence.

The ability of the multiple-choice test to discriminate between the driver groups suggests that it might have similar validity to that of the free-response test. Considering the correlation analyses, one might be tempted to argue that the multiple-choice test is even more sensitive to driving experience than the free response test, with years of driving (post-license) and mileage only correlating with the former but not the latter.

On the basis of the current results, it appears that the multiple-choice test is no worse than the free response test at discriminating between our driver groups. Given the positive correlations with measures of experience, and the pragmatic advantages that a multiple-choice format offers (objective marking, immediate feedback, quick administration, etc.), we argue that providing participants with four options to choose between should be the preferred format.

5.1. General discussion

The current study compared two different versions of the hazard prediction test: free response and multiple-choice formats. The results suggest that both versions are able to discriminate between experienced and novice drivers, however significant relationships between both driving experience and mileage with prediction accuracy were only found for the multiple-choice format.

While several studies have reported finding significant differences in accuracy between novice and experienced drivers when using multiple-choice formats (Lim et al., 2014, Ventsislavova et al., 2016), others have failed to find these differences (Malone and Brünken, 2013). However, when Malone and Brunken (2016) compared the traditional reaction time paradigm with the multiple-choice format, they found that experienced drivers outperformed novices in both test types. Following these results, they concluded that the ability of a test to discriminate between driver groups is more likely to be due to the scenario type (i.e. the particular videos used) rather than the assessment methodology employed.

On the basis of the current results, and those of other researchers reported in the literature, it appears that the multiple-choice format is no less discriminative than other test formats, yet pragmatically it offers a range of other advantages. For instance, the multiple-choice format allows quick administration and objective scoring of the answers. In contrast, a free response prediction test can take much longer to administer, and, without the development of natural language processing algorithms, must be scored offline by hand. Not only is this time consuming and risks errors during subjective coding, but it also means that participants cannot receive an immediate score for their performance. For research purposes, participants do not need to be given a score at the end of the test (though they typically prefer this). However, if the prediction format was to be used in driver training or licensing procedures, then the immediate feedback that can be provided with a multiple-choice format is a necessity.

There are other potential concerns however regarding the use of multiple-choice formats. Malone and Brunken (2016) argued that providing multiple options for drivers to choose between might simply tap into memory processes rather than driving skill. This was particularly pertinent for their study which only presented the options after a hazard clip had been viewed in its entirety. In this case, drivers might have to choose an option that refers to a hazard that they had seen tens of seconds ago. While such a short temporal gap might not be considered to place a huge strain on memory, the point remains valid: any gap between the target event and the probe question might produce results confounded by memory processes. However, in the current study, the clip occludes just as the hazard begins to develop, and the options are immediately provided. The safest drivers will be aware of this hazard at the point of occlusion, and thus this answer should be readily available in working memory.

One further concern with the multiple-choice format is the generation of incorrect options that provide plausible alternatives to the correct answer. If the distracter options are too unlikely, then the viewer may be able to guess the correct answer by rejecting the implausible options. For this reason, our distracters were developed using the responses given by participants in Experiment 1. We recommend this as a method for generating plausible and realistic distracters. While this process is time consuming, it helps ensure that the correct answer does not stand out.

In addition to supporting the development of a multiple-choice format, the current studies have provided additional insight into the act of hazard prediction. First, the fact that we have two further experiments that demonstrate the effectiveness of the hazard prediction test serves to bolster the standing of this methodology, adding to the growing literature that has successfully employed the occlusion format (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009; Lehtonen et al., 2017; Lim et al., 2014; Ventsislavova et al., 2016). This is especially important given the more equivocal nature of the findings in the field when using more traditional response-time hazard perception tests (Wetton et al., 2010; Horswill et al., 2015).

Specifically, the results from Experiment 1 again demonstrate the robustness of the discriminative validity across varying occlusion points (see also Crundall, 2016). Admittedly, the gap between experienced and novice driver performance appears at its weakest with the distal occlusion, though this is understandable: if the clip is cut before there is any evidence to point towards the impending hazard, then neither group should be able to provide a correct answer. Conversely, if the clip is cut *after* the hazard has fully appeared, one might expect all driver groups to get the answer correct (though, interestingly, Malone and Brünken, 2016, still found experiential differences using probes that were presented after the hazard had fully developed).

Experiment 2 also provided new insight via the comparison of four experiential groups. The experiential differences lay between novices and the moderately-experienced driver group. The fact that novices were not significantly better than learners, could suggest that hazard prediction is a skill that is still developing during the first year of independent driving. This lack of improvement in novice drivers over learners is surprising given that all of our novices trained for, and passed, the official UK hazard perception test within (at most) 12 months prior to taking part in the study. This raises the possibility that the introduction of the UK hazard perception test in 2002 into the licensing procedure has not reached its full potential for improving the hazard detection skills of our new drivers. Certainly, the young driver problem has not been eradicated in the UK, with one recent report from the Department of Transport noting that young car drivers are involved in 18% of road collisions, despite only accounting for 5% of the total miles driven (DfT, 2015). It is possible that the focus on *responding* to hazards in the UK test, rather than *predicting* them, may have missed an opportunity to ensure learners are trained in the most important aspects of hazard avoidance (Pradhan and Crundall, 2017).

It should be noted however that any test is only as good as the stimuli that comprise it, and vagaries in stimuli may always prevent wider generalisation of effects. In this particular study, for example, 7 out of 15 clips contained pedestrian hazards. This may be considered to place too much emphasis on this particular type of hazard. In defence of the clip selection, we draw attention to the fact that all stimuli were collected naturalistically, across a variety of roadways (rural, dual carriageway, urban and suburban). Furthermore, other studies have demonstrated the discriminative validity of the hazard prediction paradigm using different clips and different hazards (e.g. Lim et al., 2014, Castro et al., 2014), and the current results are consistent with the emerging pattern in the literature. Nonetheless, the need to assess the hazard prediction test across a wider range of hazards should be addressed in future research, and ideally with larger samples.

In conclusion, the results suggest that a multiple-choice format of the hazard prediction test can provide an efficient and effective tool for discriminating between driver groups on the basis of driving experience. It is particularly sensitive to the boundary between novice and moderately-experienced drivers, which accords with the literature that identifies the first 12 months post-license as being particularly problematic for young drivers. With the growing evidence-base for hazard prediction tests as a valid method of discriminating driver groups, this study has demonstrated how such a test could be designed for widespread automated deployment, thus providing a valid alternative to the current UK hazard perception test that still relies on confounded response-time measures.

References

Alberti, C.F., Shahar, A., Crundall, D., 2014. Are experienced drivers more likely than novice drivers to benefit from driving simulations with a wide field of view? Transportation Res. Part F: Traffic Psychol. Behav. 27, 124–132.

Andrich, D., Marais, I., 2014. Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. Appl. Psychol. Meas. 36, 432–449.

Bishop, H.J., Biasini, F.J., Stavrinos, D., 2017. Social and non-social hazard response in drivers with autism spectrum disorder. J. Autism Dev. Disord. 47 (4), 905–917.

Castro, C., Padilla, J.L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B., López-Ramón, M.F., Crundall, D., 2014. Development and validation of the Spanish HP Test. Traffic Inj. Prev. 15 (8), 817–822.

Castro, C., Ventsislavova, P., Peña-Suárez, E., Gugliotta, A., García-Fernández, P., Eisman, E., Crundall, D., 2016. Proactive listening to a training commentary improves hazard prediction. Saf. Sci. 82, 144–154.

Chapman, P., Underwood, G., Roberts, K., 2002. Visual search patterns in trained and untrained novice drivers. Transp. Res. Part F 5, 157–167.

Cheng, A.S.K., Ng, T.C.K., Lee, H.C., 2011. A comparison of the hazard perception ability of accident-involved and accident-free motorcycle riders. Accid. Anal. Prev. 43 (4), 1464–1471.

Cockerton, C., Isler, R.B., 2003. Developing real-life driving simulations for novice driver education. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake, VA.

Crundall, D., 2016. Hazard prediction discriminates between novice and experienced drivers. Accid. Anal. Prev. 86, 47–58.

Crundall, D., Chapman, P., Phelps, N., Underwood, G., 2003. Eye movements and HP in police pursuit and emergency response driving. J. Exp. Psychol.: Appl. 9 (3), 163–174.

Crundall, D., Crundall, E., Clarke, D., Shahar, A., 2012. Why do car drivers fail to give way to motorcycles at T-junctions? Accid. Anal. Prev. 44, 88–96.

Crundall, D., Kroll, V.R., 2018Crundall and Kroll, in press. Prediction and perception of hazards in professional drivers: Does hazard perception skill differ between safe and less-safe fire-appliance drivers? Accid. Anal. Prev (In press).

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

DfT, 2015. Facts on young car drivers. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/448039/young-car-drivers-2013-data.pdf (last accessed, 21/12/17).

Farrand, P., McKenna, F.P., 2001. Risk perception in novice drivers: the relationship between questionnaire measures and response latency. Transportation Res. Part F: Traffic Psychol. Behav. 4 (3), 201–212.

Fisher, D.L., Thomas, F.D., Pradhan, A.K., Pollatsek, A., Blomberg, R.D., Reagan, I., 2010. Development and Evaluation of a PC-based Attention Maintenance Training Program (Report No. DOT HS 811 252). National Highway Traffic Safety Administration, Washington, DC.

Foss, R.D., Martell, C.A., Goodwin, A.H., O'Brien, N.P., 2011. Measuring Changes in Teenage Driver Crash Characteristics During the Early Months of Driving. AAA Foundation for Traffic Safety.

Gugliotta, A., Ventsislavova, P., Garcia-Fernandez, P., Peña-Suarez, E., Eisman, E., Crundall, D., Castro, C., 2017. Are situation awareness and decision-making in driving totally conscious processes? Results of a hazard prediction task. Transportation Res. Part F: Traffic Psychol. Behav. 44, 168–179.

Haladyna, T.M., 1999. Developing and Validating Multiple Choice Test Items, second ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Horswill, M.S., 2016. Hazard perception in driving. Curr. Directions Psychol. Sci. 25 (6), 425–430.

Horswill, M., Anstey, K.J., Wood, J., Hatherly, C., 2010. The crash involvement of older drivers is associated with their hazard perception latencies. J. Int. Neuropsychol. Soc. 16 (5), 939–944.

Horswill, M.S., Gartha, M., Hill, A., Watson, M.O., 2017. The effect of performance feedback on drivers' hazard perception ability and self-ratings. Accid. Anal. Prev. 101, 135–142.

Horswill, M.S., Hill, A., Wetton, M., 2015. Can a video-based hazard perception test used for driver licensing predict crash involvement? Accid. Anal. Prev. 82 (23), 213–219.

Isler, R.B., Cockerton, C., 2003. A computer-based interactive multimedia training CD-ROM for novice drivers in New Zealand. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake, VA.

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

Lehtonen, E., Sahlberg, H., Rovamo, E., Summala, H., 2017. Learning game for training child bicyclists' situation awareness. Accid. Anal. Prev. 105, 72–83.

Lim, P.C., Sheppard, E., Crundall, D., 2013. Cross-cultural effects on drivers' hazard perception. Transp. Res. Part F 21, 194–206.

Lim, P.C., Sheppard, E., Crundall, D., 2014. A predictive HP paradigm differentiates driving experience cross-culturally. Transp. Res. Part F 26, 210–217.

Malone, S., Brünken, R., 2013. Assessment of driving expertise using multiple choice questions including static vs. animated presentation of driving scenarios. Accid. Anal. Prev. 51, 112–119.

Malone, S., Brünken, R., 2016. The role of ecological validity in hazard perception assessment. Transportation Res. Part F: Traffic Psychol. Behav. 40, 91–103.

McCartt, A.T., Shabanova, V.I., Leaf, W.A., 2003. Driving experience, crashes and traffic citations of teenage beginning drivers. Accid. Anal. Prev. 35 (3), 311–320.

McKenna, F.P., Crick, J.L., 1991. Hazard Perception in Drivers: A Methodology for Testing and Training. Transport and Road Research Laboratory, Crowthorne, UK.

McKenna, F.P., Horswill, M.S., 1999. Hazard perception and its relevance for driver licensing. J. Int. Assoc. Traffic Saf. Sci. 23 (1), 36–41.

Parmet, Y., Borowsky, A., 2014. Survival analysis: A fresh approach for analyzing response times in driving-related hazard perception paradigms. Transp. Res. Part F 25, 98–107.

Pelz, D.C., Krupat, E., 1974. Caution profile y driving record of undergraduate males. Accid. Anal. Prev. 6, 45–58.

Petzoldt, T., Weiß, T., Franke, T., Krems, J.F., Bannert, M., 2013. Can driver education be improved by computer based training of cognitive skills? Accid. Anal. Prev. 50, 1185–1192.

Pradhan, A.K., Crundall, D., 2017. Hazard avoidance in young novice drivers: Definitions and a framework. In: Fisher, D.L. (Ed.), Handbook of Teen and Novice Drivers: Research, Practice, Policy and Directions. CRC Press, Boca Raton, FL.

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

Rosenbloom, T., Perlman, A., Pereg, A., 2011. Hazard perception of motorcyclists and car drivers. Accid. Anal. Prev. 43 (3), 601–604.

Sagberg, F., Bjørnskau, T., 2006. Hazard perception and driving experience among novice drivers. Accid. Anal. Prev. 38 (2), 407–414.

Scialfa, C.T., Borkenhagen, D., Lyon, J., Deschênes, M., 2013. A comparison of static and dynamic hazard perception tests. Accid. Anal. Prev. 51, 268–273.

Scialfa, C.T., Deschênes, M.C., Ference, J., Boone, J., Horswill, M.S., Wetton, M.A., 2011. A hazard perception test for novice drivers. Accid. Anal. Prev. 43 (1), 204–208.

Shahar, A., Alberti, C.F., Clarke, D., Crundall, D., 2010. Hazard perception as a function of target location and the field of view. Accid. Anal. Prev. 42 (6), 1577–1584.

Sheppard, E., Ropar, D., Underwood, G., van Loon, E., 2010. Brief report: Driving hazard perception in autism. J. Autism Dev. Disord. 40, 504–508.

Spicer, 1964. citado en Pelz, D.C., Krupat, E., 1974. Caution profile and driving record of undergraduate males. Accid., Anal. Prev., 6, 45–58.

Thomas, F.D., Rilea, S.L., Blomberg, R.D., Peck, R.C., Korbelak, K.T., 2016. Evaluation of the Safety Benefits of the Risk Awareness and Perception Training Program for Novice Teen Drivers (Report No. DOT HS 812 235). National Highway Traffic Safety Administration, Washington, DC.

Underwood, G., Ngai, A., Underwood, J., 2013. Driving experience and situation awareness in HP. Saf. Sci. 56, 29–35.

Ventsislavova, P., Gugliotta, A., Peña-Suarez, E., García-Fernandez, P., Eisman, E., Crundall, D., Castro, C., 2016. What happens when drivers face hazards on the road? Accid. Anal. Prev. 91, 43–54.

Watts, G.R., Quimby, A.R., 1979. Design and Validation of a Driving Simulator for Use in Perceptual Studies. Report 907. Transport Research Laboratory, Crowthorne, UK.

Wetton, M.A., Hill, A., Horswill, M.S., 2011. The development and validation of a hazard perception test for use in driver licensing. Accid. Anal. Prev. 43 (5), 1759–1770.

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? Accid. Anal. Prev. 54, 57–66.

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. Accid. Anal. Prev. 42 (4), 1232–1239.

Williams, A.F., Tefft, B.C., 2014. Characteristics of teens-with-teens fatal crashes in the United States, 2005–2010. J. Safety Res. 48, 37–42.

Worthen, B.R., White, K.R., Fan, X., Sudweeks, R.R., 1999. Measurement and Assessment in Schools. Longman, New York.

Yeung, J.S., Wong, Y.D., 2015. Effects of driver age and experience in abrupt-onset hazards. Accid. Anal. Prev. 78, 110–117.

Zimasa, T., Jamson, S., Henson, B., 2017. Are happy drivers safer drivers? Evidence from hazard response times and eye tracking data. Transp. Res. Part F: Traffic Psychol. Behav. 46, 14–23.

APPENDIX E: Data output, Hazard frames of clips, Informed Consent and Debrief forms

Appendix E contains a link to a folder that contains the SPSS outputs for each experiment, hazard frames of the videos featured for each experiment, and consent forms and debriefs.

All the materials can be accessed via the following link: tiny.cc/PetyaPhD

-The SPSS outputs are in separate folders named after each experiment. Each output is named after the respective analysis for each experiment.

-The folders containing the hazard frames are placed in the following sub-folders:

- Chinese clips
- Israeli clips: Hazardous clips and Quasi-hazardous clips
- Spanish clips
- UK clips

The number of each clip is identical to those representing each hazard in Table 2.1 (UK Clips), Table 2.2 (Chinese clips), Table 2.3 (Spanish clips), Table 2.4 (Israeli clips).

-Informed consent and Debrief forms for each experiment have been placed in the respective folder containing sub-folders for each experiment.