Bachelor Thesis Psychology:

The role of task features in task difficulty and discriminatory power within a video-based hazard prediction test for driving.

Author: Chris Roelfsema (s2526832)

Supervisor: Erik Roelofs

Second Supervisor: Simone Borsci

18th of August, 2024

Abstract:

This research aims to investigate how task features within a video-based hazard prediction test affect item difficulty and the ability to discriminate between novice and experienced drivers. A video-based hazard prediction test was taken online by 77 participants. Along with the test, a small questionnaire and several demographic questions was administered to the participants.

All scenarios within the hazard perception test were assessed based on several key task features, that lead up to a total task complexity score. The scored task features were related to performance on each scenario. Along with that, the degree to which each scenario was able to discriminate between experienced and novice drivers was also related to the task features of each scenario.

The analysis showed no significant correlations between any of the task features and the performance on the test (test difficulty). A trend for the available time to think and the performance was found, this effect neared significance ($r = 0.34$, $p = 0.06$).

The full test was unable to discriminate successfully between novice and experienced drivers. No significant differences were found for any of the scenarios in the intended direction. In scenarios 10 and 20, novice drivers significantly outperformed experienced drivers, directly contradicting the goal of the test. The correlational analysis between the discriminatory power of the scenarios and the task features resulted in no significant correlations. A strong significant correlation was however found between the discriminatory power and the difficulty of the test ($r = 0.59$, $p = 0.04$).

In conclusion, the study did not confirm a clear relationship between task features and either the discriminatory value or test difficulty.

Introduction

In 2022, the Netherlands saw an increase of 20% in traffic-related fatalities and injuries compared to the previous year (SWOV, 2023). This translates to 745 fatalities due to traffic and another 134.000 victims of traffic incidents that were treated in the emergency care of hospitals. Traffic-related fatalities are the second leading cause of death for people between the ages of 10-30 (Centraal Bureau Statistiek (CBS), 2023).

Up to 95% of traffic incidents are believed to be a result of human errors (Habibzadeh omran et al., 2023; Stanton & Salmon, 2009) According to Stanton & Salmon (2009), these human errors can be categorized into action errors, cognitive and decision-making errors, observation errors, information retrieval errors, and violations. Notably, cognitive and decision-making errors play a significant role, as highlighted by Treat et al. (1979), who found that these type of errors caused or played a role in 56% of traffic crashes. Understanding the cognitive processes behind driving is thus of significant importance in understanding traffic-related incidents.

Out of all the cognitive processes involved in driving, the ability to perceive and predict hazardous situations stands out. Among all driving-related skills, only hazard perception and hazard prediction skills have been shown to significantly correlate to traffic crash involvement (M. Horswill & Mckenna, 2004; M. S. Horswill et al., 2020)

Currently, hazard perception is used in several countries as part of the drivers' licensing system. It has been researched widely for years. A commonly used definition of hazard perception is "The ability to identify dangerous situations on the road" (Crundall, 2016; M. Horswill & Mckenna, 2004). It can be seen as a multifaceted skill, consisting of several processes, including at least:

1. The detection of the potential hazard.

2. The judgement of the situation and whether it could potentially cause a conflict.
3. The classification of the event, and whether it requires a response (Wetton et al., 2010).

Hazard perception is seen as a complex cognitive process, and a significant body of research is present, investigating ways to measure and assess hazard perception reliably (Borowsky et al., 2009).

In the Netherlands, hazard perception assessment has been part of the requirements that are set for licensing since 2009 (SWOV, 2014). It is part of a theory test, which is mandatory to complete to be able to get a driver's license. Hazard perception is assessed through still image tests, in which the participant sees an image of a situation, and subsequently has to decide whether to brake, let go of the gas or do nothing (CBR, n.d.).

Worldwide, hazard perception tests prevailed in the United Kingdom and Australia. Later, several countries in the EU also implemented it as a mandatory part of the driver tests (European road safety observatory, n.d.). Even more globally there have been moves toward hazard perception tests, with several Asian countries also looking into the matter.

However, contrary to the still-image tests used in the Netherlands, assessment of hazard perception is commonly done through video-based tests. In these video-based tests, the participant must react promptly upon perceiving a potential hazard (Wetton et al., 2010). The participants' hazard perception skill is often judged based on the reaction time required to respond to the presented hazards. Variants in which the participant must choose between several answers are also commonly used (hazard perception questionnaires). Whether video-based or still-image assessment of hazard perception is objectively superior over the other is not clear-cut, with sources reporting contradicting findings (Cao et al., 2022; Habibzadeh omran et al., 2023). Several sources report still-image tests to correlate only weakly with

driving experience, while other sources report the correlation to be found in both still-image and video-based hazard perception tests.

Both still image and video-based tests have thus been shown to be able to discriminate between novice and experienced drivers in most cases (operationalized as drivers with ten or more years of driving experience) (Habibzadeh omran et al., 2023; Wetton et al., 2010). Additionally, the results of the video-based tests and the still-image tests were shown to correlate weakly between each other. According to Habibzadeh omran et al. (2023), this could indicate that both methods could be measuring different dimensions of hazard perception. Although still image tests are typically cheaper to create, there is no concrete proof that it could replace a video-based test due to this discrepancy in results and possible difference in measurement dimensions, and implications in the correlations between driving experience and performance on the test.

Within the types of hazard perception tests, several performance measures are used, each presenting its own advantages and drawbacks. Although reaction time is the most widely used form of measure for hazard perception, it is not without limitations (Vlakveld, 2014). For example, it leaves room for measurement errors, where participants may react to stimuli other than the presented threat. Other commonly used metrics, in the form of computer-based hazard perception tests, which can be done in both still-image and video-based forms, also presents its own drawbacks, with . Other, less used, metrics are hazard hit rate and several types of measures of eye fixation including fixation duration, variance, probability, and several others (Cao et al., 2022).

Although hazard perception tests of several sorts are commonly used in practical contexts as part of the driver's licensing system, more recently, hazard prediction tests have been used to measure a drivers' hazard perception skill as well (Horswill et al., 2020)

Hazard prediction is defined the prediction of hazards before it is present, based on the hazard evidence present in precursors, it happens before hazard perception takes place (Pradhan & Crundall, 2016). This means that drivers would notice evidence of potentially dangerous situations before they occur, thus predicting the danger, rather than merely responding to it.

A hazard prediction test usually consists of video clips from real or simulated traffic. Drivers have to predict what happens next when the video stops, instead of pointing out or reacting to the already present hazard. Horswill et al. (2020) found that drivers that have not been involved in traffic accidents and experienced drivers made more valid predictions than drivers that have been involved in accidents and inexperienced drivers.

The overarching goal of any hazard perception or prediction test with a purpose of certification of learning drivers is thus to successfully differentiate between drivers proficient in hazard perception and those lacking in the skill. Typically, this validation is achieved by relating test outcomes to driving experience , as experienced drivers (fifteen or more years of experience) have been shown to exhibit greater proficiency in hazard perception skills compared to novice drivers (three or less years of experience) (Manley et al., 2020).

Underwood et al. (2011) conducted a study on hazard perception in a driving simulator to assess its comparability to on road driving behaviour. In this study, big differences were observed between experienced and inexperienced drivers, especially in the glancing times and techniques between novice and experienced drivers, with differences in variance of horizontal search of the front window especially large on dual carriageways according to Underwood et al. (2011). Underwood et al. (2011) claimed that novice drivers were busy with maintaining appropriate road positions, while experienced drivers “showed sensitivity to the demands of the roadway”. They found that the difference between novice and experienced drivers differs per scenario, depending on the demands of the roadway and the tasks at hand. They

recommended further research into what conditions allowed situations to successfully discriminate between experienced and novice drivers.

In line with these findings, Stanton & Salmon (2009) reported that the differences between novice and experienced drivers is the result of higher order cognitive skills, in which novice drivers lack. This includes Hazard Perception, but also Hazard Prediction. A study by Wang et al. (2022) revealed that high task complexity inhibits drivers from dividing full attention to predicting on-road hazards, resulting in a reduction of a drivers' ability to predict dangerous situations. Understanding task complexity and the factors that influence it is thus necessary for Hazard Perception and Hazard Prediction tests.

To construct reliable and valid hazard prediction tests, a theoretical framework is necessary. The Task-Capability Interference (TCI) Model by Fuller (2005) (Figure 1) is used as a theoretical foundation, as the probability of a subject to successfully perform in any given scenario is deemed a function of the balance between the subjects' competence on the task and the demands of the task at hand. Modern test theory shares this notion of difficulty in a task being the balance between capability and demands of the task. This allows the TCI Model to work in line with the modern test theory, applying the theory to the driving framework. It is widely used to conceptualize task difficulty and demands in traffic. More in depth, the model states that task demands are dependent on the speed and trajectory of the vehicle of the driver, the characteristics of the vehicle itself, other road users, and the environment. With the characteristics of the vehicle itself, the "operational features" of the vehicle are considered. For example, the information displays, the headlights, and its ability to brake sufficiently et cetera. Environmental factors include weather conditions, state of the road driven on, road signs and markings et cetera.

According to the TCI model, the interaction between the drivers' capabilities and the task demands at hand determines the result of the situation; either control or a loss of control,

which can result in collision or a “lucky escape”. The perceived difficulty is thus based on the balance of driver capability and task demand (R Fuller et al., 2008).

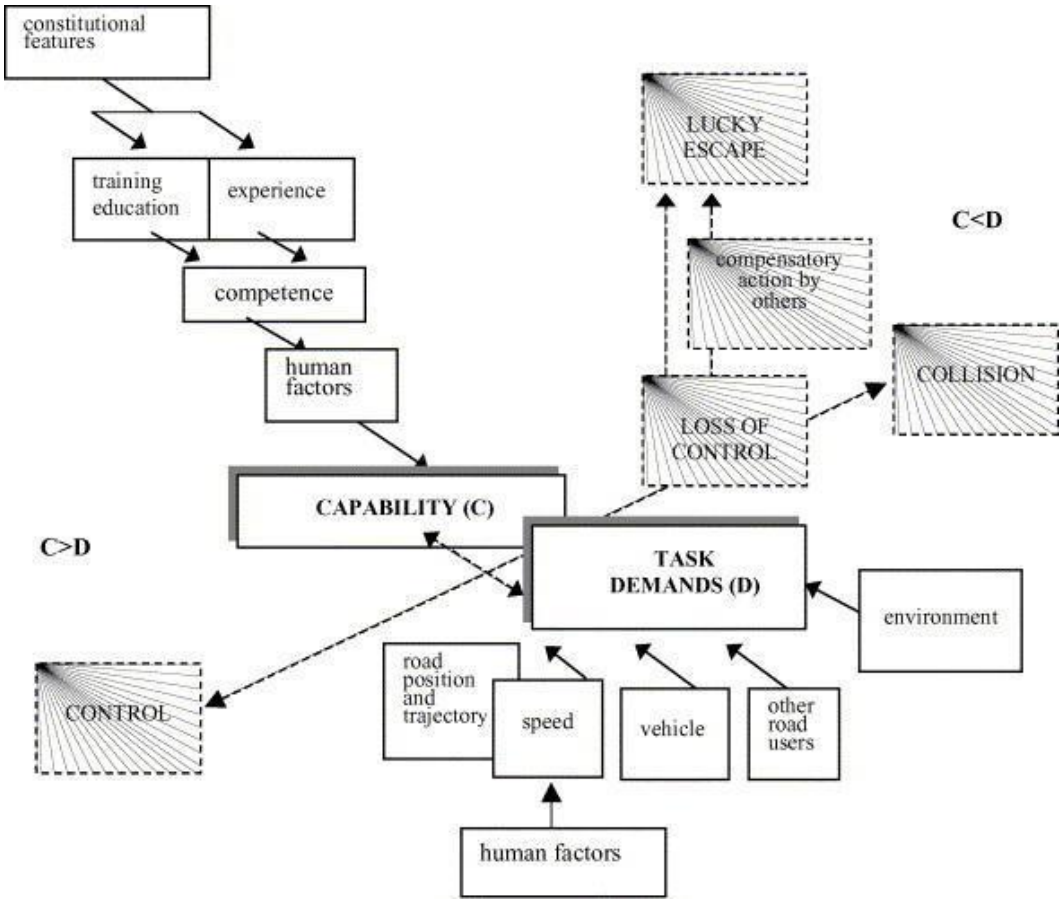

Figure 1: The Task-Capability Interference Model (Fuller, 2005).

In this thesis, the assumption is made that this conceptualisation for task demand in traffic is upheld in video-based hazard perception tests. This assumption is supported by research, as Underwood et al. (2011) for example showed that video-based hazard perception tests can deliver comparable results for several measures of hazard perception compared to real life situations and driving simulation scenario’s.

In order to create a test that covers a range of traffic situations of varying complexity, knowledge on what task characteristics affect task complexity is necessary. A study by Roelofs et. al (2011) created a tool to assess the difficulty of a traffic scenario based on a variety task features derived from a study by (Stanton & Salmon, 2009). They outlined

several situational factors, or task characteristics, that hinder specific mental processes in driving.

They found that the vision and visibility, the number of other road users, the number of vulnerable road users, the available room for action, the speed differences, the time pressure and the traffic regulations could influence the perception, evaluation and decision-making processes, the same processes involved in Hazard Perception and Hazard Prediction.

With that said, the aim of this study is to explore how several task features within a video-based hazard perception test attribute to the quality of the test. The two psychometrical indices that show the quality of the test are item difficulty and the discriminative power per item. Therefore, the difficulty levels and the discriminatory abilities of each test item (scenario) will be assessed, to lead to a better understanding of task difficulty of video-based hazard perception tests. Furthermore, recommendations will be made for further improvements to the test design.

To perform this study, data will be analysed from a prototype of a video-based hazard prediction test by Compaan, Vissers, Tsapi, and Roelofs (2023), developed for the CBR, the Dutch national Institute for transportation exams. The study consists of three parts:

Part 1: Description of task features within the scenario's:

All task features within the test scenario's will be described. The hazard perception test at hand consists of 12 main scenario's, that will each be described in terms of their task difficulty, based on the model of (E. Roelofs et al., 2011). To assess the task difficulty, a performance score will be constructed, containing the ratio of correct to incorrect performance on each scenario. To assess the discriminatory ability, the performance of experienced drivers will be compared to those of novice drivers. The descriptive task features are therefore conceptualized as any situational factor that.

Part 2: Assessment and correlational analysis of task difficulty and discriminatory ability per scenario:

Next, the task difficulty and discriminatory ability per scenario will be assessed, and subsequently checked for correlation to the task features of each scenario.

Part 3: Conclusion

Lastly, conclusions will be drawn based on the found relationships or the lack thereof. This will be aimed at answering, and guided by, the following questions:

1. How do task features influence the empirically found difficulty per scenario of a video-based hazard prediction test?
2. How do task features influence the discriminatory power of a video-based hazard prediction test?

Based on the existing literature, the task features are expected to have a significant positive correlation on both the found difficulty and discriminatory power of the video-based hazard prediction test (Cito B.V., 2015).

Methods

Participants

A convenience sample aimed at Belgian and Dutch drivers who already hold their driving license was assembled, consisting of the data from 77 participants who completed the test and the survey, of which 44 participants (57.2%) were Dutch, 32 (41.5%) were Belgian, and 1 (1.3%) were German. 43 men (55.8%), and 33 women (42.9%) participated in the study. One participant (1.3%) preferred not to disclose their gender. The ages of the participants ranged from 19 years to 78 years, with an average of 40 years (SD = 20.6 years).

Most participants (N=35) were retired, a further 21 were employed, 12 were unfit for work, and for 9 their employment status is unknown. On average, participants had their drivers' license for 18,7 years (SD = 18,6). 24 Participants fit into the "novice driver" category (up to 3 years of driving experience), and 32 participants fit into the "experienced driver" category (15 years of driving experience or more). 2 Participants were filtered out because of inconsistencies in their respective demographic data.

Procedure

From all participants the responses per scenario were collected and analysed, along with demographic data and a questionnaire regarding their involvement in traffic accidents, behaviour in traffic and how they obtained their drivers' license. All participants gave explicit informed consent and the data was anonymized before it was received by the research team.

After participants entered the questionnaire, they first were asked for informed consent. After this was obtained, the participants were shown an example exercise of the hazard prediction test, before starting the full hazard prediction test. The participants subsequently were asked to perform all 12 hazard prediction scenario's to the best of their ability. Afterwards, the participants were asked to fill in a questionnaire regarding demographic data and several questions regarding their driving experience. The full questionnaire can be found in Appendix A.

Materials

The questionnaire in its entirety was conducted using Qualtrics. The hazard prediction test by Compaan, Vissers, Tsapi, and Roelofs (2023) was used for the collection of the hazard prediction data. All data was gathered and exported in Qualtrics, before being analysed in R,

with the addition of RStudio and the additional packages Broom, Tidyverse, Janitor, Corr, GGPlot2, Foreign, Dplyr, Psych, Effsize, and Hmisc.

A coding scheme (see Table 1) was used to describe each scenario on a scale of one to five, based on its task features: vision and visibility, the number of other road users (cars, trucks etc.), vulnerable road users (bikes, mopeds, pedestrians), room for action, speed differences, time pressure and traffic regulations within the scenario. The coding scheme was developed by (E. Roelofs et al., 2011). It is based on the aforementioned task characteristics identified by (Stanton & Salmon, 2009). A total score for task complexity was made based on the sum of the scores of the aforementioned factors. All task features are thus subjective in nature. The scores within the coding scheme were estimated by two people to assess and ensure inter-rater reliability which was assessed by means of Intra Class Correlation (ICC) scores.

The Hazard Prediction Test

The hazard prediction test itself contained twelve items, and one exemplary exercise in advance. After the exemplary exercise was shown, the actual test started. A short animated video from the view of the driver of a car in traffic is shown once, without the ability to pause, rewind or skip. The video contains a latent hazard that would only reveal itself after the video is cut off, which the participant is asked to predict based on the video. The video starts with a black screen featuring the word “start, and ends with a black screen featuring the word “end”. As the video ends, the participant is shown a black screen (see Figure 2). Subsequently the participant is shown 3 possible still image responses, of what they predict happens next (see Figure 3).

Figure 2 - Example item from the hazard prediction test

Figure 3 - Response options of the example item

Used variables

1. **Calculated variables**

The variable “difficulty” is the percentage of participants correctly completing a specific scenario. A higher percentage of correct answers is seen as a lower difficulty. This is measured in the variable “performance”, which is calculated by dividing the number of right answers by the total amount of answers times one hundred. Novice drivers are seen as drivers

with up to three years of licensed driving experience, while experienced drivers are defined as drivers with more than 15 years of licensed driving experience. The experience based definitions are based on a similar study by Horswill et al. (2020), who used these exact cut-off points. The discriminatory ability of a scenario is operationalized as the standardized difference between mean scores of experienced and novice drivers on each item (Cohen's d).

2. Subjectively coded task feature variables of The Hazard Prediction scenarios

In order to investigate the task features of the 12 scenario's within the test, seven subjectively coded variables were created, that add up to a total score for task complexity. The first variable "vision and visibility" is the average of a subjective rating on how visible the threat is to the participant scored from one to five given by the two raters, with five meaning the highest complexity. Examples of this include blocked vision due to parked cars, trees, camouflage and certain whether conditions (low sun or fog).

The number of road users is not the absolute number of road users, but rather the subjective complexity within the scenario resulting from the amount of non-vulnerable road users arriving at the location at the same time, also scored from one to five. This includes cars, trucks, busses, tractors and potential other non-vulnerable road users. The same method applies to vulnerable traffic participants, differing in the type of participants, as the vulnerable participants includes pedestrians, mopeds, motors and bikes.

Room for action is judged based upon the available space for possible courses of action for the vehicle of the user, thus the vehicle of which the driver's field of view is shown, judged from one to five. As the room for actions becomes smaller, the task complexity rises.

Speed differences are judged based upon the difference between the speed of the vehicle of the user and the speed of the direct surrounding traffic. For example, the user could

be driving 120 kilometres per hour, but if the surrounding traffic is also driving 120 kilometres per hour, the resulting speed difference score would be one.

Finally, the traffic regulations score is a subjective judgement of the complexity deriving from the presence or absence from traffic signs, road markings and layout. More road signs in this case does not necessarily result in a higher score in this case. The *less* regulated the situation is, the more the driver has to determine the right course of action himself. So, the more the situation calls for interpretation from the driver, the higher the score. All scenarios with their respective scores can be found in Table 1.

The ICC(2, k) of the raters was calculated at 0.980, this is statistically significant with $p < 0.001$, which is a satisfactory level of inter-rater reliability.

Subjective task features

Table 1: The filled in coding scheme of aggregate subjective scores for task complexity

Scenario	Vision and Visibility	Other participants	Room for action	Speed differences	Time pressure	Regulation	Vulnerable participants	Total score
1	1.5	2	1	2.5	1.5	3	2	13.5
2	3	3.5	3	2.5	2.5	3	1.5	19
3	4.5	2	2	3	3	1.5	2.5	18.5
4	2	2.5	1.5	2	1.5	1.5	3	14
7	3	2	2.5	2.5	3	2	2.5	17.5
8	3	3.5	2.5	2	1.5	2	2	16.5
9	3.5	2.5	2.5	2	2	2	4	18.5
10	3	2.5	2	3.5	2	2	2.5	17.5
12	2	2.5	2	1	1.5	2	3	14
15	3.5	3.5	2.5	3	2.5	2.5	3.5	21
20	3.5	2.5	4	3.5	3	2	3.5	22
22	3.5	2.5	1.5	1	1	2	3	14.5

3. Objective task feature variables

Apart from the variables that were used to assess and code the scenario in a subjective manner, three objective variables were also used to describe the task features of the 12 scenarios. The variable “Driving speed” is the average driving speed in kilometres per hour of the vehicle of the participant shown in the video. The variable “Time of onset critical precursor” is the time between the start of the video, until the potentially hazardous situation is starting to show in seconds. Thinking time refers to the time subjects have to prepare a prediction of an oncoming event. This time is computed by computing the time differences between the total clip time (the time when the clips turn black) and the time when a critical precursor was set on. For instance, in scenario 2, the subject has to look into the rear mirror, where he sees that a motorcyclist is gaining speed, probably because the rider is going to overtake. The thinking time in this scenario would be between when the motor cyclist is starting to gain speed, until the end of the video. All objective task features per scenario can be found in Table 2.

Objective task features

Table 2:
Objective task features of the hazard prediction scenarios

Scenario In item*	1. Location of critical precursor of hazard	2. Precursor	3. Displayed driving speed in km/h	4. Time of onset critical precursor	5. Length clip until turning black	6. Thinking time
1	Windshield	Truck appearing on entry lane	110	14.01	18.161	4.15
2	Rearview Mirror	Motorcyclist approaching with higher speed	100	2.71	8.07	5.36
3	Windshield	Deer appearing in woods left from the car	35	11.48	14.68	3.20
4	Windshield	Child on the right curb lets a ball roll onto the street	30	10.50	15.00	4.50
7	Windshield	Motorcyclist approaching with high speed	80	6.86	9.72	2.86
8	Windshield	White van approaching with high speed	30	3.44	4.60	1.16
9	Windshield	Child running on left curb, later covered from sight by a parked van	5	9.66	12.16	2.50
10	Windshield	Cyclist appearing and immediately covered form sight on crossroad left	60	5.40	8.36	2.96
12	Windshield	Whit car in parking place left showing reverse lights	15	9.63	11.03	1.40
15	Windshield	Onward cyclist approaching stopped car on right side	30	14.58	17.72	3.14
20	Rearview Mirror	Motorcyclist approaching with high speed	35	11.31	13.51	2.20
22	Rearview Mirror	Tow motorcyclist approaching with slightly higher speed	10	7.57	17.62	10.04

* Note: scenarios 5, 6, 1, 13, 14 were not used, because there was less than 1 second thinking time

Data analysis: Relation of task features to scenario difficulty and discriminatory ability

No changes were made to the dataset before importing the file into Rstudio as a CSV file. Upon receiving the file, for each participant, a 1 (answered correctly) or a 0 (answered incorrectly) was noted per scenario of the hazard prediction test. For the demographic data, all categorical responses were coded into numbers for analysis, which were re-coded back into the appropriate categorical responses in Rstudio for interpretation.

To answer Research Question 1, a variable "performance" was created, which contains a difficulty score per scenario. It is constructed as follows: number of correct responses / total responses *100. This variable was constructed into a new dataframe, as it did not fit the format of the first dataframe. This new dataframe for performance was then combined with a dataframe with the created task complexity ratings and the objective task features of the different scenarios: the speed in kilometres per hour, time to think in seconds, and the time of the appearance of the threat. For this dataframe, a correlational matrix was created using RCorr and Hmisc, including the Pearson correlation coefficient and the respective p-values. Scatterplots were also created for each variable, with the variable performance on the y-axis, and the corresponding variable on the other axis to further assess potential correlations between variables.

Descriptive measures were also gathered based on the demographic variables and questions from the questionnaire, including age, gender, country of residence and licensed years of driving.

To answer Research Question 2, the discriminatory ability of the test was assessed by looking at the difference in test scores between experienced drivers and novice drivers. Similar to the study of Horswill et al., novice drivers are defined as drivers with up to three licensed years, and experienced drivers are defined as drivers with 15 or more licensed years.

To analyse the difference, the mean and standard deviation were calculated for each group. A Welch Two-Sample t-test was used to calculate whether there is a statistically significant difference between the groups for the complete test. The discriminatory ability of each scenario separately was analysed by performing separate t-tests and calculating Cohen's d to analyse the size of the effects and their statistic validity ($P<0.05$). In these analyses, the novice drivers were put into group 1, and experienced drivers into group 2 for both the t-tests and the calculation of Cohens' d. This setup means that a potential negative t-statistic or Cohen's d would indicate that experienced drivers scored higher than novice drivers. Finally, the found discriminatory ability (operationalized as Cohen's d) were related to the task scores by correlational analyses.

Results

Research Question 1: The correlation between task features and difficulty levels.

In Table 3 all correlation coefficients and p-values for the relation between performance per scenario and the researched variables can be found. No significant correlations were found between performance on the scenarios and any variable related to task complexity. No variable neared statistical significance, and no trends were discovered for that reason.

Table 3

The pearson's correlation coefficients and p-values for all task complexity variables against performance per scenario

Variables	r	p-value
Time pressure	-0.20	0.8680
Vision and visibility	-0.05	0.5958
Other participants	-0.17	0.4852
Vulnerable participants	-0.22	0.7405
Room for action	0.11	0.7517
Speed differences	0.10	0.4332
Regulation	-0.25	0.9296
Total score	0.03	0.7054

The analysis of the objective measurements of the scenario's also did not lead to any significant correlations. All correlation coefficients and p-values for the objective measurements can be seen in Table 4.

Table 4

The Pearson's correlation coefficients and p-values for all measurement variables against performance per scenario

Variables	r	p-value
Driving speed in km/h	1.00	0.90
Time to think in seconds	0.34	0.06
Time until hazard in seconds	0.56	0.53

All scatterplots for the relationships between both the objective and subjective task features and the difficulty (performance) per scenario can be found in Appendix B. In each figure, each dot represents one scenario within the Hazard Prediction test, with the

accumulated difficulty score “performance” on the y-axis, and the score per task feature on the x-axis. Figure 13 shows the scatterplot for time to think against performance. In this instance, a regression line was added due to the nearing significance, with a p-value of 0.06.

Research question 2: the Discriminatory power of the hazard prediction test and the correlation to task features

No statistical difference ($p = 0.17$) in mean scores on the complete hazard prediction test was found between the novice drivers ($M=8.96$, $SD =1.62$, $N =24$) and experienced drivers ($M = 8.59$, $SD = 1.16$, $N = 32$). Furthermore, the 95% confidence interval for the difference in means ranged from -0.20 to 1.14. As this ranges beyond zero, this further suggests that there is no statistically significant difference between the mean hazard prediction scores for experienced and novice drivers on the complete test.

The t-tests and calculated Cohen’s d per scenario can be found in Table 5. The t-tests on scenario level resulted in significant differences in means for scenario 10 ($p = 0.02$) and scenario 20 ($p = 0.04$). In both scenarios, novice drivers score significantly higher than experienced drivers. Small effects for scenario 1, 2, 4, and 8 were found. For scenarios 1 and 8, these effects were negative, which means the average score for experienced drivers was higher than that of novice drivers.

Table 5

The t-test results and Cohen's d and per scenario.

Scenario	t-value	p-value	Mean novice	Mean experienced	Cohen's d
1	-1.4	0.17	0.5	0.69	-0.38
2	0.83	0.41	0.92	0.84	0.2
3	-0.46	0.65	0.5	0.56	-0.12
4	1.2	0.23	0.875	0.750	0.31
7	0.78	0.44	0.96	0.906	0.2
8	-1.8	0.07	0.29	0.53	-0.49
9	1.1	0.27	0.79	0.66	0.3
10	2.3	0.02	0.88	0.63	0.57
12	0.62	0.53	0.79	0.72	0.17
15	-0.15	0.88	0.54	0.56	-0.04
20	2.1	0.04	0.96	0.78	0.51
22	-0.2	0.84	0.96	0.97	-0.05

Finally, the task features were related to the discriminatory ability of each scenario, expressed as the standardized difference between mean scores of experienced and novice drivers on each item, as well as item difficulty. No significant correlations were found between the task features and the discriminatory ability. A strong significant correlation was found between item difficulty and the discriminatory ability ($p = 0.04$, $r = 0.59$). All data can be found in table 6.

Table 6

The Pearson coefficients (r) and the respective p-values for the relationship between the discriminatory ability (Cohen's d) of each scenario (n = 12) and the task features

Task feature	r	P
Time pressure	0.33	0.3
Vision and visibility	0.09	0.78
Other participants	-0.15	0.64
Vulnerable participants	0.40	0.19
Room for action	0.42	0.18
Speed differences	0.31	0.33
Regulation	-0.25	0.44
Total score	0.35	0.27
Driving speed in km/h	-0.1	0.75
Time to think in s	-0.07	0.75
Time until hazard in s	-0.12	0.71
Test difficulty	0.59	0.04

Discussion

As reported by E. C. Roelofs et al. (2021) systematically identifying task features and assessing their effect on scenario parameters such as difficulty, discriminatory ability, reliability and validity is of positive influence for future test improvement. Aligning with this goal, the aim of this study was to investigate the relationship between several task features and the difficulty, and between task features and the discriminatory ability within the video-based hazard prediction test by Compaan, Vissers, Tsapi, and Roelofs (2023).

It was expected that the results would be in line with the current literature, and thus that there would be a significant positive correlation between the task complexity scores and the performance and the task complexity and the discriminatory ability on each of the scenarios within the hazard prediction test.

To answer the research question, “How do task features influence the empirically found difficulty per scenario of a video-based hazard prediction test?”, the following can be concluded. The most important findings for were that there were no significant correlations between any of the coded task features and the performance in each scenario. Neither the total task complexity score nor any of the underlying task features showed a significant correlation to the performance.

No significant correlations were found in the objectively assessed descriptive data either (driving speed, thinking time and time until the precursor). Although all were insignificant, the correlation between the time to think and performance on each scenario is nearing significance ($P = 0.061$). Therefore, it could be argued there is a trend to be found here, taking into account the small number of investigated scenarios ($n = 12$), that leave little power for statistical tests.

It thus seems that a longer period between the start of the critical situation until the end of the video, where the predictive response is required, in this test results in a lower difficulty. People appear to have much more trouble with sudden changes and short periods in advance of the precursor. This could be because one has far less time to prepare themselves and imagine what could follow next. On top of that, if the thinking time gets too short, it would no longer test hazard prediction or perception, but simply the reaction to a sudden hazard, according to the aforementioned definition of Crundall, (2016) for hazard perception, and the definition of hazard prediction by (Pradhan & Crundall, 2016).

According to Crundall, ideally, the onset would fall into the strategic and vigilance zones. As the thinking time gets shorter, it could move into the tactical or operational zone, thus inhibiting the participant from perceiving and predicting the potential hazard, and having them simply react to it. This could explain the found trend.

These findings are in line with research done by Compaan et al. (2023), in which they found that an increase in the available thinking time leads to a higher scoring percentage and a lower difficulty rating. Subsequently, similar to this study, they also did not find a correlation between driving speed and difficulty and the location of the precursor and the difficulty.

Even with this found trend, It could be concluded that, within the context of the used Hazard Prediction test, Task features were of no significant influence on the difficulty level of each scenario.

To answer the second Research Question: “How do task features influence the discriminatory power of a video-based hazard prediction test?”, the following can be concluded. The full test has shown to be unable to discriminate between experienced drivers (fifteen or more years of experience) and novice drivers (up to three years of experience). As with the insignificant correlation, the test as used before is known to have a limited Cronbach's alpha due to the limited number of items and would need a set of nearly thirty items of similar reliability to reach a Cronbach’s alpha of over 0.80.

No scenario possessed a statistically significant ability to differentiate successfully between experienced and novice drivers in the intended direction, and on scenario 10 and 20 novice drivers scored higher than the experienced drivers.

The found effect sizes were correlated to the task features despite the lack of significant effects. This resulted in no further significant correlations between task features and the discriminatory ability. It did result in a strong positive correlation between item difficulty and discriminatory ability ($p = 0.04$, $r = 0.59$). This means that as the items get more difficult, they differentiate better between novice drivers and experienced drivers, in the intended direction. While the correlations are intriguing, the insignificance of the effect sizes impairs the robustness of the test and these findings. Therefore, these findings should lead to

further exploration of the matter, rather than taken as a definitive conclusion. It does open the door for the argument that the test items could be lacking in discriminatory ability due to the items being too low in difficulty. Future research could look into the correlation between item difficulty and discriminatory ability, to further establish appropriate difficulty levels for hazard prediction tests, ensuring their accuracy in assessing drivers' hazard prediction skills."

Based on these findings, it could therefore be concluded that task features are not found to have any significant influence within the context of the used Hazard Prediction test.

Further recommendations for further research would thus be to expand the set to a total of well over thirty items, with a setup similar to that of the one currently used and increase the difficulty level of the test. This broader setup could ameliorate the low Cronbach's alpha, and allow for a more complete data analysis, as the 12 data points that were used in this study have proven to be a limitation to the reliability, and greatly inhibit the chance of any significant findings, and use of more complex methods. Further studies might also seek to explore other methods than a video-based hazard prediction test. Especially simulation-based hazard prediction tests could provide further insight into components that contribute to task difficulty in the broader scope.

More research could also be done in deeper analysis of the components, and especially the combination of components that could lead up to a correlation. Compound effects of several variables leading to significant correlations together did not fall into the scope of this research, but could yield interesting results, nonetheless.

This study underlines the importance of broadly standardized test scenarios, as the lack of different standardized scenarios possibly contributed to insignificant correlations and implications for the found results. On top of that, the notion that items higher in difficulty could better differentiate between the experienced and novice drivers warrants careful

deliberation of item difficulty in future test designs, even if these findings are inconclusive. Although the study did not directly yield results that spark research into a new innovated test for hazard prediction to be used for licensing, or significant changes in current methods, it offers insights that could guide future research in refining hazard prediction tests to be used in assessing hazard perception.

Conclusion

To conclude, the goal of this study was to gain more insight into the role of task features and task complexity on the difficulty and discriminatory ability of scenarios within a video-based hazard prediction test. The results showed significant implications, with no significant correlations across 11 investigated task features between both item difficulty and the discriminatory ability of items. No task features were thus found to influence the task difficulty or discriminatory ability. This study therefore presents no strong evidence for the influence of task features on task difficulty or the discriminatory ability. Future studies could increase the number of scenarios that are investigated, as this could potentially solve several statistical issues with the methods at hand, which could help better develop hazard perception and hazard prediction tests in the future. Based on these results, it can be concluded that the current hazard prediction test is not suitable for use, as it does not reliably differentiate between experienced and novice drivers.

Reference list

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. https://doi.org/https://doi.org/10.1016/j.trf.2009.02.001

Cao, S., Samuel, S., Murzello, Y., Ding, W., Zhang, X., & Niu, J. (2022). Hazard Perception in Driving: A Systematic Literature Review. *Transportation Research Record*, *2676*(12), 666–690. https://doi.org/10.1177/03611981221096666

CBR. (n.d.). *Leren en oefenen*. https://www.cbr.nl/nl/rijbewijs-halen/auto/theorie-examen-auto/leren-en-oefenen

Centraal Bureau Statistiek (CBS). (2023). *1 916 zelfdodingen in 2022, 54 meer dan in 2021*. https://www.cbs.nl/nl-nl/nieuws/2023/19/1-916-zelfdodingen-in-2022-54-meer-dan-in-2021

Cito B.V. (2015). *Booklet 1: knowledge base regarding on-road coaching for safe driving*.

Compaan, R., Tsapi, A., Vissers, J., & Roelofs, E. (2023). *Ontwikkeling theorie-examen gevaarpredictie*.

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

European road safety observatory. (n.d.). *The Driver Test*. https://road-safety.transport.ec.europa.eu/european-road-safety-observatory/statistics-and-analysis-archive/young-people/driver-test_en

Fuller, R, McHugh, C., & Pender, S. (2008). Task difficulty and risk in the determination of driver behaviour. *European Review of Applied Psychology*, *58*(1), 13–21. https://doi.org/https://doi.org/10.1016/j.erap.2005.07.004

Fuller, Ray. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention*, *37*(3), 461–472. https://doi.org/https://doi.org/10.1016/j.aap.2004.11.003

Habibzadeh omran, Y., Sadeghi-Bazargani, H., Yarmohammadian, M. H., & Atighechian, G. (2023). Driving Hazard Perception tests: A Systematic Review. *Bulletin of Emergency And Trauma*, *11*(2), 51–68. https://doi.org/10.30476/beat.2023.95777.1370

Horswill, M., & Mckenna, F. (2004). Drivers' hazard perception ability: Situation awareness on the road. *A Cognitive Approach to Situation Awareness: Theory and Application*, 155–175.

Horswill, M. S., Hill, A., & Jackson, T. (2020). Scores on a new hazard prediction test are associated with both driver experience and crash involvement. *Transportation Research Part F: Traffic Psychology and Behaviour*, *71*, 98–109. https://doi.org/https://doi.org/10.1016/j.trf.2020.03.016

Manley, H., Paisarnsrisomsuk, N., Hill, A., & Horswill, M. S. (2020). The development and validation of a hazard perception test for Thai drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *71*, 229–237. https://doi.org/https://doi.org/10.1016/j.trf.2020.04.011

Pradhan, A. K., & Crundall, D. (2016). Hazard avoidance in young novice drivers: Definitions and a framework. In *Handbook of teen and novice drivers* (pp. 81–94). CRC Press.

Roelofs, E. C., Emons, W. H. M., & Verschoor, A. J. (2021). Exploring task features that predict psychometric quality of test items: the case for the Dutch driving theory exam. *International Journal of Testing*, *21*(2), 80–104. https://doi.org/10.1080/15305058.2021.1916506

Roelofs, E., van Onna, M., & Brookhuis, K. (2011). *Developing driving task scenarios for developmentally tailored driving assessments: using an evidence centered design model*.

Stanton, N. A., & Salmon, P. M. (2009). Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, *47*(2), 227–237. https://doi.org/https://doi.org/10.1016/j.ssci.2008.03.006

SWOV. (2014). *SWOV Fact Sheet*.

SWOV. (2023). *De Staat van Verkeersveiligheid 2023*.

Treat, J. R., Tumbas, N. S., McDonald, S. T., Shinar, D., Hume, R. D., Mayer, R. E., Stansifer, R. L., & Castellan, N. J. (1979). *Tri-level study of the causes of traffic accidents: final report. Executive summary.*

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

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. https://doi.org/https://doi.org/10.1016/j.trf.2013.12.013

Wang, L., Li, H., Guo, M., & Chen, Y. (2022). The Effects of Dynamic Complexity on

Drivers' Secondary Task Scanning Behavior under a Car-Following Scenario. In *International Journal of Environmental Research and Public Health* (Vol. 19, Issue 3). https://doi.org/10.3390/ijerph19031881

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

Appendix A: the questionnaire as seen by the participants:

Thank you for participating in this survey. The aim of this survey is to test your Hazard Prediction skills, and to see how this relates to your driving behaviour. To participate it is important that you possess a valid B driver's license and that you are able to comprehend the test questions in either English, Dutch or German. We advice you to use a laptop, PC or tablet to complete this survey; using a mobile phone is not possible.

►The survey will take about 25 minutes and the results will be processed anonymously.
►You are free to withdraw from this research at any time.
►The collected data will be used for analysis to investigate the validity, difficulty and usability of this test. Your data cannot be traced back to you and we strive to assure the anonymity and safety of your data.
►This research has been reviewed and approved by the University of Twente's BMS ethics committee.
►If you have any questions, remarks or complaints please contact our student representative [c.a.biester@student.utwente.nl], our supervisor [e.c.roelofs@utwente.nl] or the UT BMS ethics committee [ethicscommittee-hss@utwente.nl] if necessary.

	Yes	No
I have read and understood the study information as provided above.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that my Hazard Prediction Test scores will be used to assess the test itself, and that they will be analyzed in relation to the other information I provide.	○	○
I understand that the data provided by me will be stored in accordance to the guidelines for data collection and storage	○	○
I understand that the answers and perspectives I provide during this pilot study will be used to improve the survey and that this information will be processed under a random participant number and cannot be traced back to me	○	○
I possess a valid B driver's license; either full or provisional.	○	○
I speak English, Dutch or German well enough to comprehend the test questions.	○	○

What device are you currently using to complete this survey? It is recommended that you use a large screen (laptop/pc or possibly tablet) to complete this survey. The use of a cell phone is not possible for this survey

Where are you currently living?

What is your gender identification?

What is your year of birth?

What is your work status (multiple answers possible)?

What is your highest educational attainment?

When did you get your (provisional) driver's license?

Where did you get your driver's license?

What type of B driver's license do you have?

During your driver training, how many clock hours did you drive the car under the following conditions? (Move the slider to the appropriate number. If you want to select the number 0, drag the slider to the right and back to 0)

How many exam attempts have you made to pass the sections of the driver training listed below? (Move the slider to the correct number. If you want to select the number 1, drag the slider to the right and back to 1)

On average, how many kilometers do you drive by car per year? Spread across categories: commuting, personal and business (Move the slider to the appropriate number. If you want to select the number 0, drag the slider to the right and back to 0 again)

How many **active** collisions have you been involved in since obtaining your driver's license? These are collisions where you, the driver, hit another road user or an obstacle and were therefore at fault. (Move the slider to the appropriate number. If you want to select the number 0, drag the slider to the right and back to 0)

Did 1 or more of these collisions involve:

	Yes	No
Damage to own vehicle	○	○
Damage to another vehicle	○	○
Minor injuries to yourself	○	○
Minor injuries to another person	○	○
Serious injuries to yourself	○	○
Serious injuries to another person	○	○

In the last 12 months, how often did you slow down so early while driving that you held up the other traffic?

In the last 12 months, how often did you brake so late that you ended up very close to another car or an object?

In the last 12 months, how often did you forget to check the rear view mirror while driving, to notice other traffic driving close behind you?

In the last 12 months, how often have you looked at the road ahead but you then you were suddenly surprised by a cyclist riding in front of you?

In the last 12 months, how often did another vehicle merge into your lane when you did not expect this to happen?

In the last 12 months, how often did you only have little space for a left turn to avoid an oncoming vehicle driving into you?

In the last 12 months, how often did you drive within one second distance from a car in front of you on a motorway?

In the last 12 months, how often did you want to change to the left lane and almost overlooked a car that was overtaking you at that moment?

In the last 12 months, how often have you missed a motorway exit because you were focussed on something else?

How fast do you tend to drive on 50 Km/h roads?

How fast do you tend to drive on 80 Km/h roads?

Over the last 12 months, how often:

	Hardly ever	Occasionally	Regularly	Rather often	Very often
Did you really find it difficult to keep your eyes open during a long drive because you were tired?	○	○	○	○	○
Did you get less sleep than you needed the night before you had to drive?	○	○	○	○	○
Did you drive between midnight and 6am, without having rested during the day?	○	○	○	○	○
Did you drive after working for more than a 6-hour shift?	○	○	○	○	○

[The hazard prediction example question and test]

These were all the video clips. We have one more question for you.

While answering the video clips, were you distracted which may have prevented you from choosing the correct answer for one or more clips? 0 = Not distracted at all, 10 = Very distracted (Drag the slider to the appropriate number)

Thank you for participating.

If you have any questions, comments or complaints feel free to contact us at [c.a.biester@student.utwente.nl].
If you would like to receive an answer sheet, feel free to reach out to us.

Appendix B: The scatterplots for the correlation between the task features and the task difficulty score "performance".

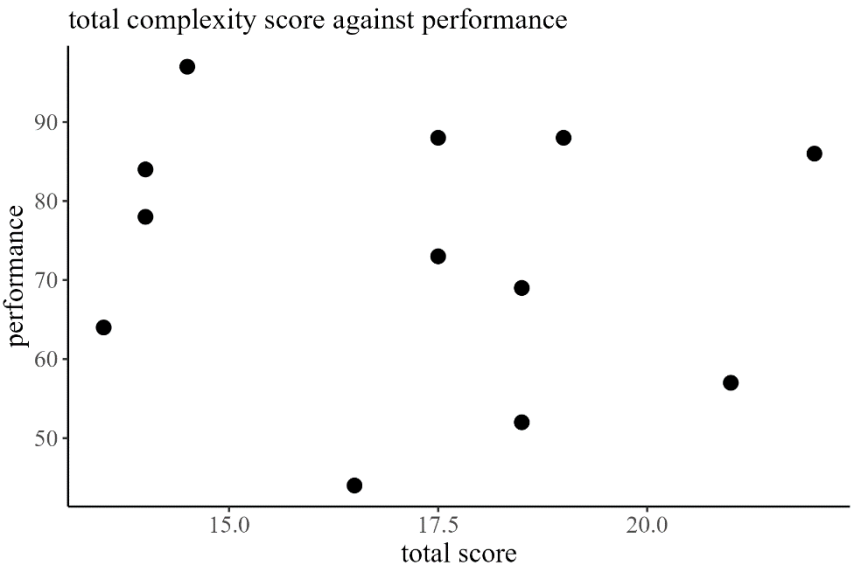

Figure 4: The scatterplot of the relationship between the total complexity score and performance on each scenario.

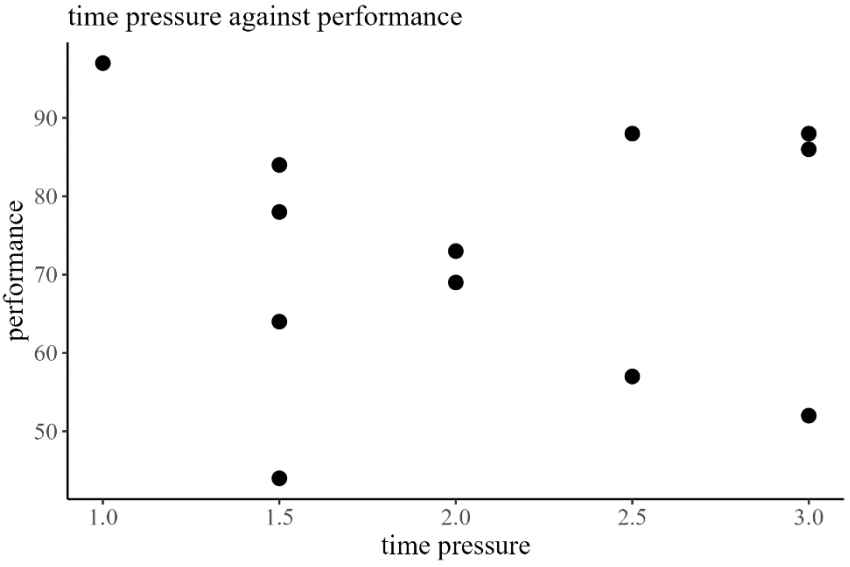

Figure 5: The scatterplot of the relationship of time pressure and performance

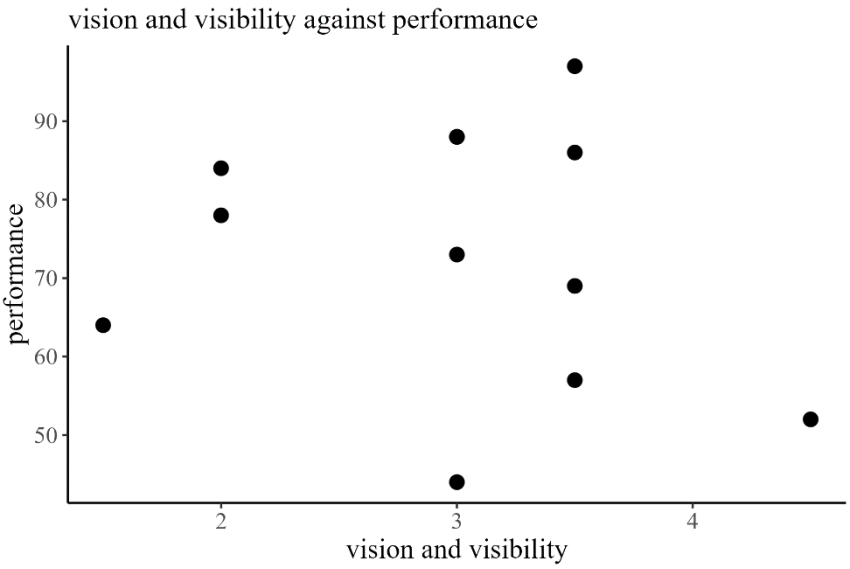

Figure 6: The scatterplot of the relationship between "vision and visibility" and performance

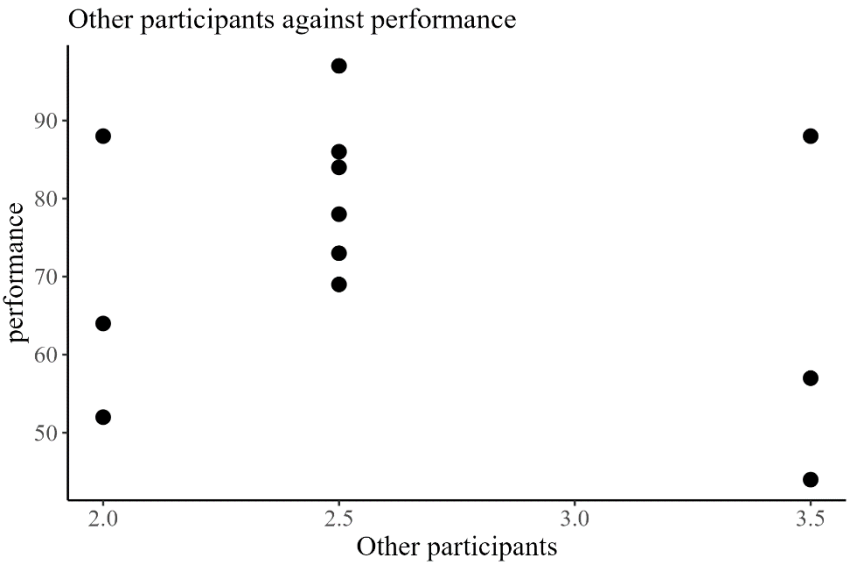

Figure 7: The scatterplot of the relationship of other participants on performance.

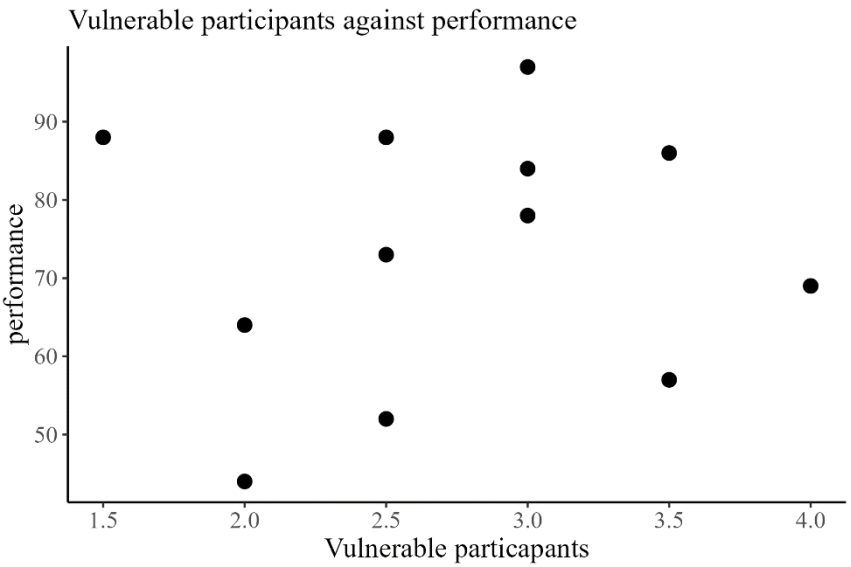

Figure 8: The scatterplot of the relationship of other participants on performance.

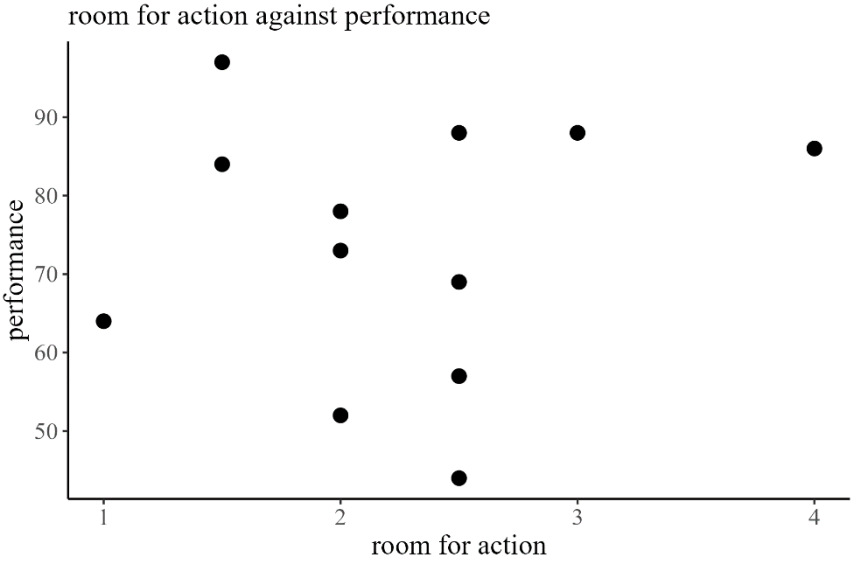

Figure 9: The scatterplot of the relationship between the room for action and performance

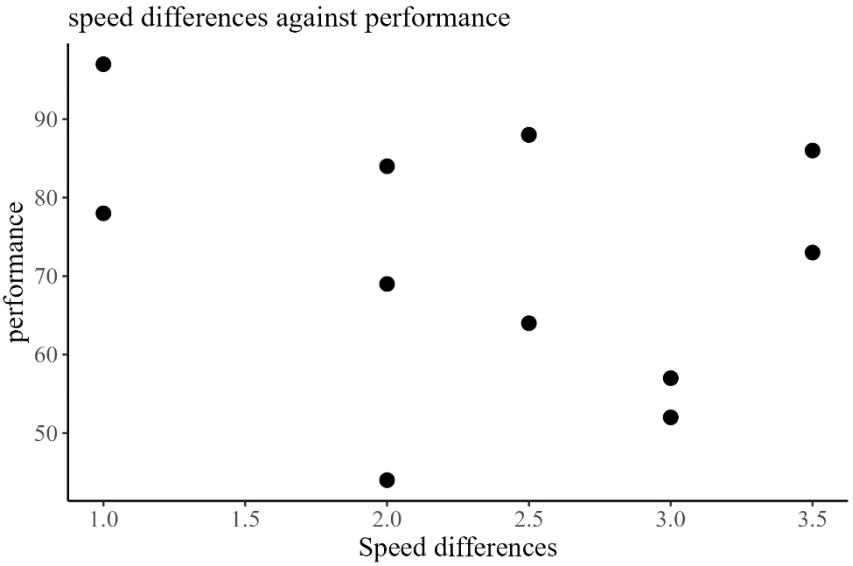

Figure 10: The scatterplot of the relationship between speed differences and performance

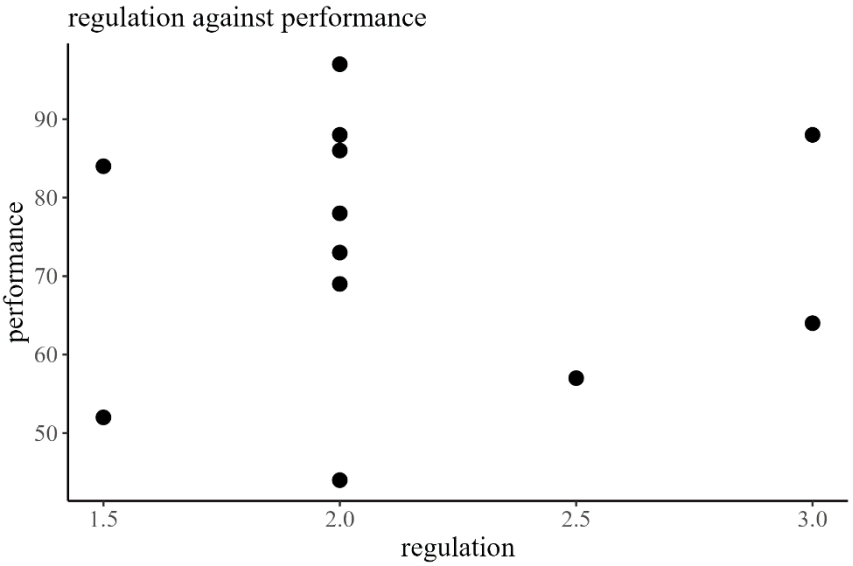

Figure 11: The scatterplot of the relationship between speed differences and performances

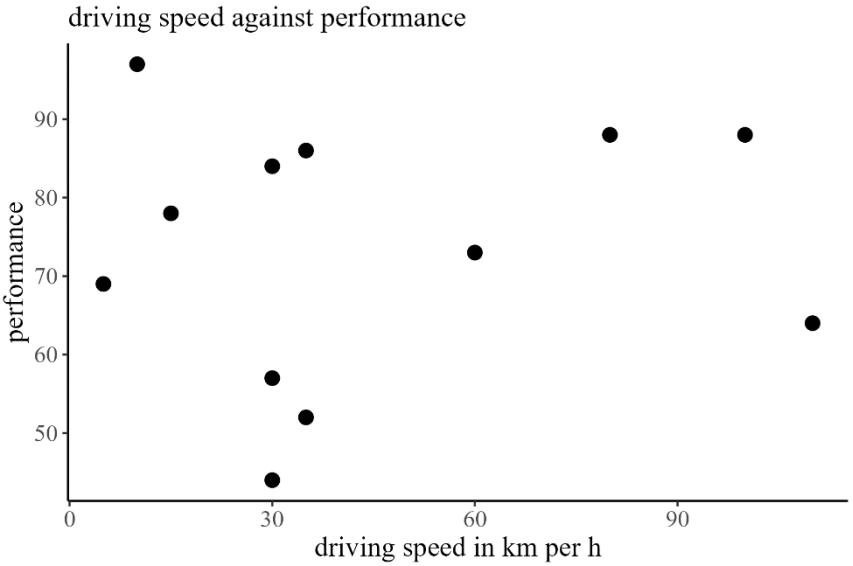

Figure 12: The scatterplot of the relationship between the driving speed and performance

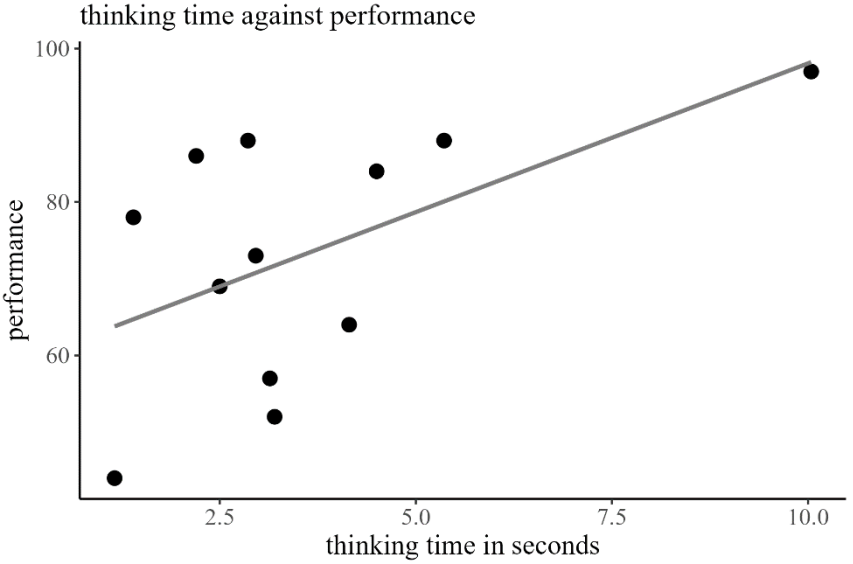

Figure 13: the scatterplot for thinking time against performance

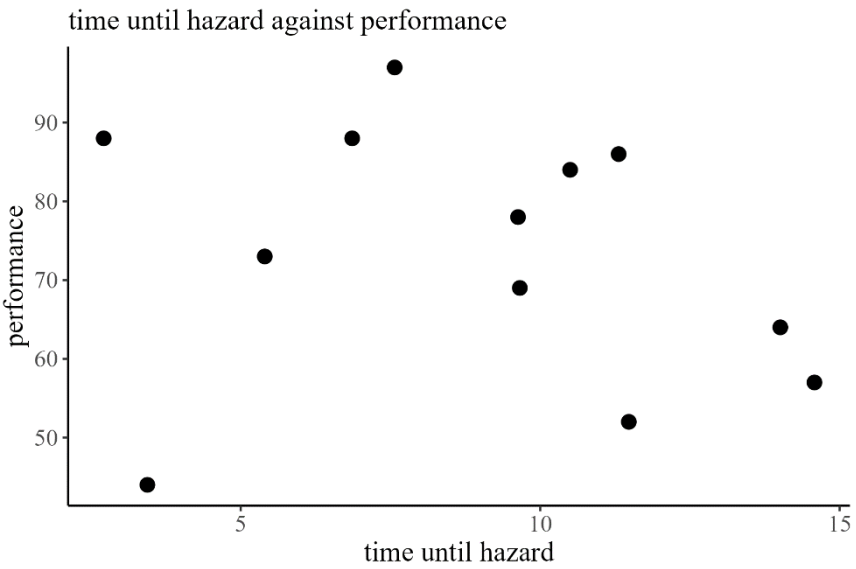

Figure 14: The scatterplot of the relationship between the time until the hazard is shown since the start of the video, and performance on each scenario.