Final Year Project

Final Report

Driving Behaviour Analysis at Approaching Yellow Traffic Light Signal

Yan Yi Cheng

17109505

School of Engineering and Technology

Sunway University

Supervisor

Dr. Richard Wong Teck Ken

School of Engineering and Technology

Sunway University

Friday, 1 July 2022

Abstract

Yellow-light-running phenomenon which often times result in red-light-running violation is leading cause of traffic light accidents. Traffic lights rotates the right of way at an intersection by taking turns to give each stream of traffic the right of way to cross the intersection. When the phase of the traffic light is constant, the behaviour of road users are predictable and any conflicts are minimized. However, during traffic light phase changes, particularly from green to yellow to red, a decision is demanded from the driver to either stop or go when traffic light signal is yellow. The zone where drivers do not have a safe clearance to either stop or go is known as the dilemma zone. A misjudgement of one's capability may result in a red-light-running violation and possibly traffic light accidents. Therefore, there is a need to understand driving behaviour when approaching yellow traffic light intersection, to mitigate and possibly reduce traffic light intersection accidents.

To understand driving behaviour, data has to be first collected in order to be analysed. Two data collection methods are implemented to collect data pertaining to driving behaviour. The first method is simulator study, where participants are recruited to drive in a driving simulator with different scenarios. The data that were collected is the initial speed, final speed, distance to intersection, brake or gas pedal usage and decision to stop or go. The second data collection method is questionnaire survey where questionnaires were shared to the masses with question portraying driving scenarios, prompting the respondents for their response. The data collected was the decision to stop or go.

With the data collected, measures of central tendency were calculated and data distribution graphs were plotted to understand the data. From there, k-means clustering was performed on the data in order to observe any characteristics from the clusters. The clusters revealed 2 main types of driving style, careful being the majority and risky being the minority. Age, gender, traffic intensity and presence of rain have all been found to have influence on an individual's driving behaviour.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1. Background / Motivation

In Malaysia, news about traffic accidents constantly plagues the pages of news outlet and worries the heart of Malaysian alike. According to data by the Ministry of Transport Malaysia, the total number of road accidents across Malaysia increase steadily in the past decade, with 414421 report road accidents in 2010 to a figure of 567516 in 2019 [1]. This amounts to a total of 4.94 million reported road accidents with the actual figure possibly being higher as minor accidents and near-misses, specifically at traffic light junctions go unreported. The year 2019 is taken as an example to understand the severity of this issue. As of October 2019, there are 16.71 million registered vehicles in Malaysia [2], therefore, indicating that there are 33.9 reported road accidents per 1000 registered vehicles. Although the exact proportion of traffic light accidents are indistinguishable, these numbers still pose as an alarming figure.

Traffic light accidents are defined as accidents that happens at a traffic light intersection between 2 or more vehicles regardless of the vehicle type. Traffic light accidents are mostly due to the red-light-running (RLR) phenomenon, where road users do not obey traffic rules and continue in motion during red light or road users overestimating their ability and the duration of yellow lights.

There are generally 2 main types of traffic lights in Malaysia; one is the more traditional timer-based traffic light whereas the other is a traffic actuated traffic light. Of the former, there are also 2 variants, one with a countdown timer shown, whereas the other without the countdown timer. Although the timer might vary between 45-120 seconds [3] for the signal timing cycle, this variation is implemented based on different intersections and not cycle to cycle. On the other hand, the traffic actuated traffic light, or better known as sensor-based traffic light aims to maximize traffic flow efficiency by allowing more green light timing to a particular direction upon detecting increased traffic load from the direction through its sensors. This type of traffic light is usually implemented at intersections with only 2 main stream of vehicles.

Traffic lights are installed with the purpose to control and maintain a safe environment for road users when approaching cross intersections as it is able to minimise conflict points within the intersection by sharing and rotating the right of way between competing traffic streams [3]. However, traffic lights accidents still do occur, especially during phase changes from yellow interval which accounts for more than half of traffic light accidents [4]. This is due to the fact that a decision is demanded from the driver when traffic light changes its phase, as the status of right of way for the current direction will be change.

The decision made by the driver would then consequently affect the possibility of getting into a traffic light accident, as is the worse case scenario after making the wrong decision (to either stop/go). This decision-making process is further complicated when drivers are in the dilemma zone (DZ). Dilemma zones are defined in either time or space, as zones where some drivers may decide to proceed through an intersection while others may decide to stop at the onset of a yellow indication [5]. An illustration of DZ is shown in Figure 1 for better clarification:

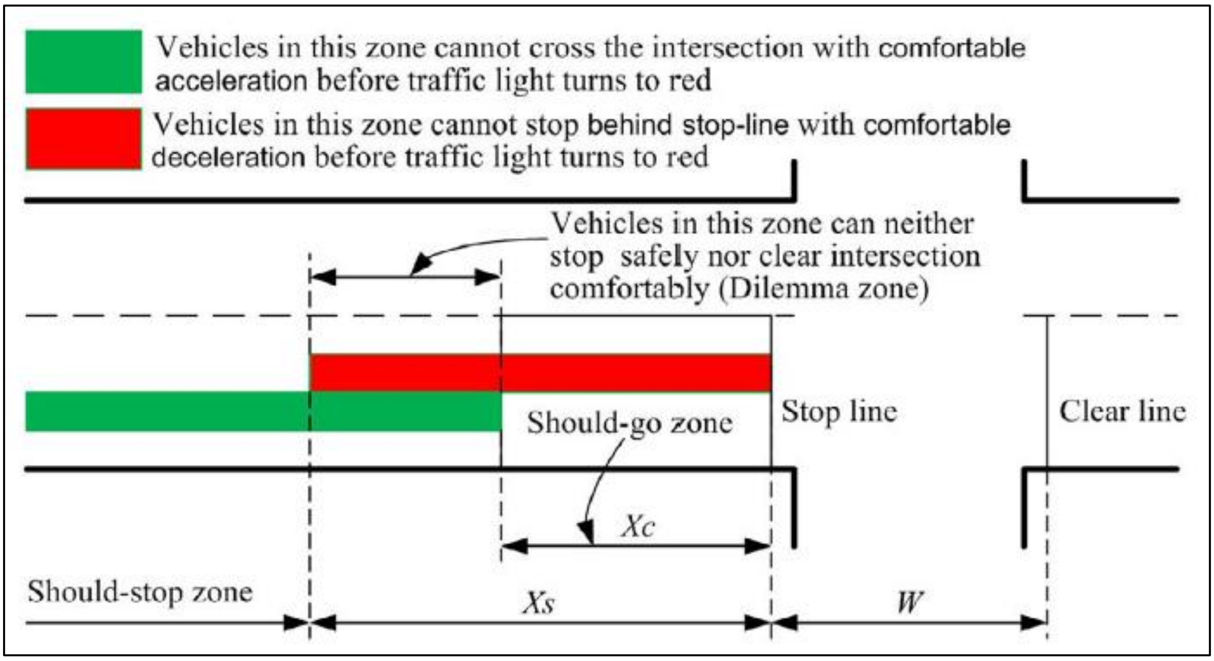

Vehicles in this zone cannot cross the intersection with comfortable acceleration before traffic light turns to red
Vehicles in this zone cannot stop behind stop-line with comfortable deceleration before traffic light turns to red
Vehicles in this zone can neither stop safely nor clear intersection comfortably (Dilemma zone)
Should-go zone
Stop line
Clear line
Xc
Should-stop zone
Xs
W

Figure 1: DZ Illustration courtesy of [6]

1.2. Problem statement

Considering all of the above, one of the biggest contributors to traffic light accidents, which remains a concerning issue to all Malaysian road users, is the decision making in dilemma zones. Therefore, there is a need to understand driving behaviour when drivers approach traffic light intersections with yellow signal. This is done so to better understand driving behaviour and the possible need to improve traffic light designs in order to reduce traffic light accidents.

1.3. Aim/Objective

In order to solve the problems above, the aim of this project is to understand driving behaviour when approaching yellow traffic signal and cluster these behaviours into careful or risky drivers, if possible. In order to achieve this aim a few objectives need to be addressed.

First, data collection methods have to be identified. Data pertaining to driving behaviour have to be first obtained before analysis can be done. Next, factors that might influence driving behaviour should be identified. This is so that these factors can be tested and relevant data collected during the data collection phase.

Secondly, suitable pre-processing approaches need to be identified to clean and extract relevant data from the collected data. The data will then be explore and understood through statistical methods, such as the measure of central tendency to have an overall understanding of the data including its distribution.

Third, data analysis techniques need to be determined to understand and identify patterns of behaviours from different individuals under similar conditions. The end result of this analysis may be numerous, distinct behaviours being identified or no driving styles produced at all due to there being no observed patterns of driving behaviour.

1.4. Project Scope

As this is a student project, naturalistic observation and controlled road studies are not expected to be used as means of data collection due to time and financial constraints. Therefore, the advantages of either of these data collection method should not be expected from this project.

Besides that, for the purpose of standardization and reduced cognitive load, the test scenarios used within the simulator study would only consist of cars as the sole vehicle type. Motorcycle, trucks, buses and other vehicle types will not be present in the test scenarios. In the same vein, pedestrians, cyclist and other non-vehicle objects but objects that might be encountered on/along the road would also be excluded from the study.

Lastly, the data that will be collected will include demographics of the participants along with their driving experience. Independent variables will also be tested through different scenarios, where the dependent variables stop/go decision will be recorded. Other values such as the approaching speed and distance to intersection will also be collected. As a whole, the dependent variable would represent an individual's driving behaviour, with emphasis given to the stop/go decision.

2. Literature Review

This section discusses on the review of literature that past researchers have done to achieve the same or similar aim as this project. This section will contain 2 sub-sections which are respective to the objectives of this project; data collection at yellow light intersections and driving behaviour analysis based on the data collected.

2.1. Data Collection at Yellow Light Intersections

As it has been mentioned, in order to perform driver behaviour analysis, a sufficient amount of data has to be available in order to perform the analysis. This section will cover the data collection (inclusive of methods use, parameters collected and the scenario) at yellow light intersections. There is a total of 4 distinct methods used based on the literatures that have been reviewed.

2.1.1. *Simulator Studies*

Rittger et al. [7] conducted a driving simulator experiment aimed at understanding unassisted driving behaviour when approaching traffic light intersections. The driving simulator used to conduct the study was a static simulator located in Würzburg Institute for Transport Science. As such, the driving simulator is rather high-end with a 300° horizontal field of vision, auditory output of 5.1 Dolby Surround System and powered by 9 PCs connected via 100Mbit Ethernet. The driving simulation software used was SILAB which is a software developed by the institute itself.

In order to increase knowledge of driving behaviour, specific parameters deemed to describe driving behaviour has to be collected when participants are driving within the scenario of the test environment. Rittger et al. has collected the driving speed, distance to intersection, acceleration (both positive and negative) and pedal usage (both throttle and brake) when participants are approaching the traffic light intersection.

Besides that, since this is a simulator study, the test scenarios where participants drive in can be designed and controlled by the experimenter beforehand. This allows flexibility in which the experimenter is able to test and understand driving behaviour under different external circumstances. Rittger et al. did this by varying factors such as the traffic light phase (solid green, solid red, green to red, red to green), presence of lead vehicle and visibility (controlled by presence of fog). This allowed them to understand driving behaviour under different traffic light phases, different levels of visibility and whether the presence of lead vehicle affected a person’s driving behaviour all from a test participant through the simulator. It should also be noted that this is one of the main advantages in using a driving simulator.

2.1.2. *Controlled Road Studies*

Taking a step away from simulation and closer to real driving is controlled road studies. Rakha et al. [5] conducted controlled road studies as a means of data collection for their study to model driver behaviour within a signalized intersection when approaching decision-dilemma zone. The test bed used for the purpose of controlled road study and data collection was a signalized intersection at the Virginia Department of Transportation’s Smart Road facility located at Virginia Tech Transportation Institute

(VTTI). This test bed is a 3.5 km private highway limited to test vehicles for the purpose of research.

In order to collect data, the test vehicle was fitted with a real-time DAS that is concealed from the driver’s view. The DAS was developed and built by the VTTI for the purpose of collecting data for experiments. Similar to the simulator study above, the collected data includes vehicle speed, traffic light signal phase as well as brake and throttle pedal values. The brake and throttle pedal values were particularly used to determine perception and reaction time for the study.

Due to this being a controlled road study, the experimenters do not have the flexibility and freely controlling the external conditions. However, the experimenters made sure that all participants were tested individually under controlled good conditions. These good conditions are defined as presence of daylight, good weather and dry pavement. Therefore, the subsequent modelling done by the experimenters would only be able to explain driving behaviour under good conditions.

2.1.3. Video Surveillance Data

Video surveillance, also known as naturalistic observation, is a form of a data collection method with no experimenter intervention as opposed to the 2 aforementioned methods. Li et al. [8] utilized a Vehicle Data Collection System(VDCS) to collect data through video surveillance in order to build machine learning models to predict driver behaviour during yellow light intervals. Since the data is collected through video surveillance, a certain degree of computer vision is involved in order to extract useful data from video footages in which the VDCS is used achieve this.

The VDCS is a system containing 6 modules, developed in C++ with an OpenCV package used to capture drivers’ behaviour data. The data that were captured are vehicle approaching speed, acceleration, distance to intersection and occurrence of red-light-running violation. This is all made possible through the 6 modules present in the VDCS. Firstly, a Gaussian Mixture Model is used for motion detection in order to detect vehicle approaching the intersection. Next, vehicle tracking and trajectory are extracted using Kalman Filter algorithm. Lastly, with motion detection and vehicle trajectory in place, the vehicle speed and acceleration are then calculated by checking the displacement and speed change of 2 successive frames.

Lu et al. [6] also used high-resolution traffic data for their research on an analysis of yellow light running at signalized intersections. The high-resolution traffic data was collected by the SMART-SIGNAL system developed at the University of Minnesota. However, data collected by the SMART-SIGNAL system is only raw data and the details on the processing of the data was not mentioned. Similarly, Yang et al. [4] research driver behaviour in yellow interval at signalized intersections based on camera image data. The system used to collect the data as well as the methods of processing the data were also not mentioned. However, this proves that video surveillance is a well-known data collection method as it is with the 2 methods above.

2.1.4. Questionnaires

Another method of data collection is through collecting responses through questionnaires. Questionnaires are an easy way to collect mass data through means self-reporting by the participants and distributing it out to the masses. The Multidimensional Driving Style Inventory (MDSI) created by Taubman-Ben-Ari et al. [9] is one such questionnaire that allows a quick understanding to one's general driving style. Although the main focus of Taubman-Ben-Ari et al. study was to conceptualize a person's habitual driving style as a driving-specific factor that can directly explain involvements in car accidents, it still offers understanding to an individual's specific driving style. The end results of Taubman-Ben-Ari et al. study revealed 8 main driving styles: dissociative, anxious, risky, angry, high-velocity, distress reduction, patient and careful.

It should also be noted that there are many other existing questionnaires that are related to driving behaviour. Some examples of these are the Driver Behaviour Questionnaire measuring risk taking on the road developed by Furnham and Saipe [10], Driver Style Questionnaire created by French et al. [11] to describe the relation between driving style, decision-making style and accident liability and the Driver Skill Inventory created by Lajunen and Summala [12] to measure self-reported safety motive and skill dimensions.

However, of the questionnaires reviewed, none directly supports the aim of this project which is to understand driving behaviour when approaching yellow traffic light. Additionally, questionnaires are traditionally prone to social-desirability bias. More considerations have to be taken into account when using this method.

2.2. Driving Behaviour Analysis

Now that the data collection methods have been reviewed and summarized above, this section will cover 2 main approaches used by researchers to understand driving behaviour. This section contains 2 sub-sections namely the statistical approach and the machine learning approach based on reviewed literatures.

2.2.1. Statistical Approach

Statistical approach is the method where researchers aim to use statistical tests and/or functions in an attempt to describe and explain human driving behaviour based on the data that has been collected. Some of the statistical approaches covered by the literature that has been reviewed are the ANOVA test for significance, different plotting of graphs to understand the driver profile (based on speed and pedal usage) when approaching traffic light intersections and factor analysis with Varimax rotation.

Using the data collected, Rittger et al. [7] applied separate ANOVA tests between the data collected (speed) for each distance section of 10m under each of the different circumstances (with/without lead vehicle and fog). The distance mentioned here is the distance between the car to the intersection. The ANOVA test was done to test on the significant distance where drivers would start applying brakes given the test scenario.

Besides that, using acceleration and deceleration, mean speed profile under different circumstances was also plotted against distance to traffic light as shown in Figure 2. This allows the researchers to understand how drivers would react and behave when approaching.

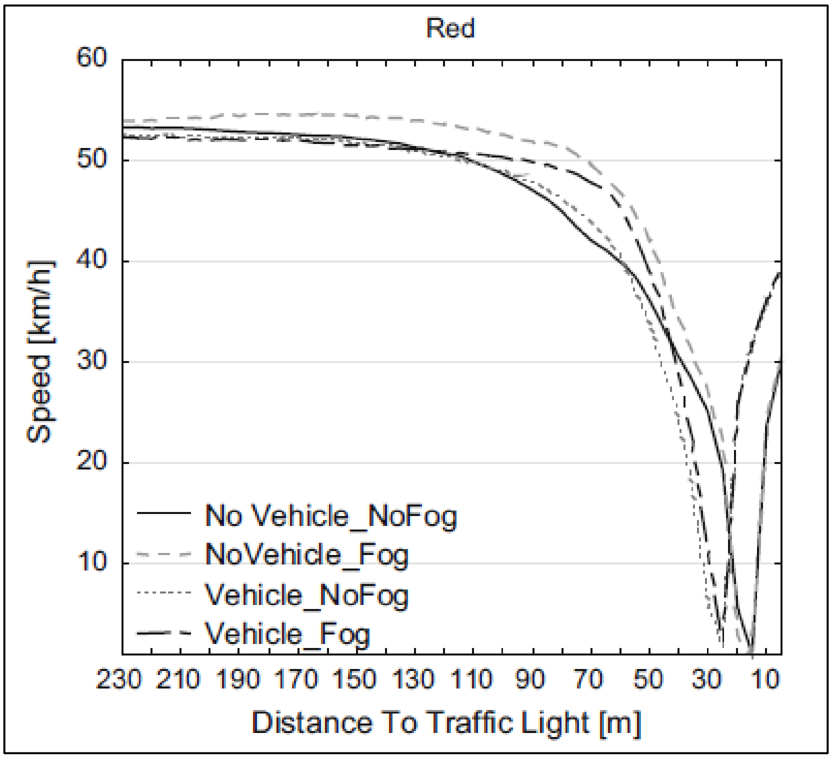


Figure 2: Example of Mean Speed Profile plotted

Similarly, Rakha et al. [5] also used graphs to analyse and understand driving behaviour. Rakha et al. plotted graphs of brake time against Time To Intersection (TTI) and stop time against TTI in an attempt to understand when the human driver would apply brakes upon yellow light onset. TTI is an attributed derived from the speed of the car divided by the distance to the intersection.

Other statistical methods used were a factor analysis with Varimax rotation by Taubman-Ben-Ari et al. [9] to determine if the 44 MDSI items fell within distinguishable domains. Varimax rotation is generally used to simplify the expression of a sub-spaces to few items each. In this case, the driving styles are the sub-spaces whereas the MDSI items are the items within the sub-spaces.

Probability is also another form of statistical approach used to analyse driving behaviour based on data collected. Van der Horst's [13] study on driver decision making at traffic signals is one such example. Using the data collected including stopping/non-stopping decisions, the probability of a road user stopping at a given time to stop line (similar to Rakha's TTI) can be calculated. This is done by taking the proportions of road users that stopped divided by the total observations for stopping decisions and vice versa for non-stopping decisions. Although this does not directly contribute to understanding driving behaviour when approaching yellow traffic signals, this would help in identifying the dilemma zone when drivers approach a yellow traffic signal. This can be observed at the time to stop line when the probability of stopping/non-stopping is around 0.5, indicating drivers are unsure if they should stop or go.

2.2.2. Machine Learning Approach

The machine learning approach has also been used by researchers to understand and predict driver behaviour. Some of the machine learning models that have been used in the past are binary logistic regression models, artificial neural network, game theory etc.

Based on the study of Yang et al. [4], a binary logistic regression model was implemented with the dependant variable being driver's choice behaviour at yellow signal's period. The outcome of the model is either vehicle chose to stop (y=1) or vehicle is driving into the intersection (y=0). The factors that were taken into account to affect this outcome were the distance between the vehicle and the stop line during yellow light onset, the vehicle speed and the presence of a countdown timer. Similarly, Li et al. [8] also applied the same method to describe a driver's stop/go decision at the onset of yellow light. However, the difference between both of these studies is that Li et al. uses a sequential binary logistic model to further validate if the "go" decisions resulted in a red-light-running violation.

Besides that, artificial neural network(ANN) was also used to approximate driving behaviour at yellow traffic light. Li et al. [14], in her study, trained an artificial neural network model to predict potential red-light-runners to serve as the basis of an innovative red-light-running prevention system. The ANN was trained using 4 input candidates: distance to intersection, travel speed, number of front vehicles before the stop line and headway from the leading vehicle. These are the parameters that Li et al. [14] has deemed to contribute to driving behaviour. The output of the ANN is a classifier, indicating whether a vehicle is a red-light-runner or not. In the literature, a vehicle is labelled a red-light-runner if it is still within the intersection upon the red-light onset. The structure of the ANN a feedforward neural network trained using the standard backpropagation algorithm. However, one key difference is that the weight of the neural network is updated only once after each epoch instead of updating after each row of input data.

It is worth mentioning that of the literatures that were reviewed, the machine learning approach in analysing driving behaviour is used mostly when the data collection method is through video surveillance. This is likely due to the data available to researcher from video surveillances, as data pertaining to kinematics of the car such as the pedal usage will be unavailable. Data that are extracted from video surveillances would usually only include speed, time and distance. These data fit in well into machine learning models to explain and predict driving behaviour. Besides that, being machine learning models, these approaches offer more in terms of predicting capabilities but comparatively loses out in terms of explanatory ability and understanding relationships between variables.

2.3. Conclusion

Taking into account all that have been reviewed and mentioned above, a simulator study is first chosen to as the data collection method for this project. With the flexibility of changing parameters, weather and traffic intensity, of test scenarios that a simulator study allow, varying conditions can tested to understand driving behaviour when approaching yellow lights under these conditions. Data collection method through surveys are also considered as they provide massive amounts of data compared to a simulator study.

The data analysis techniques will take the form of hybrid approach, taking both statistical and machine learning approach. The statistical approach will aim to understand the data for any baseline or norm behaviour. With the baseline identified, k-means clustering will be performed to possibly identify any observable patterns or characteristics within the clusters. If a characteristic is identified, the data points will be related back to their demographics and driving experience to understand the relationship between the characteristic of the clusters and their respective demographics and driving experience.

3. Method 1 (Simulator Study)

A driving simulator with different test scenarios will be used to collect data. Volunteers would then be asked to participate and drive within the driving simulator. The collected data is used to understand and analyse for possible patterns of driving behaviour when approaching yellow traffic light signal. As such, there will be 5 sub-sections to explain the flow of this method starting off with the *driving simulator selection and setup*, followed by *scenarios design*, *data collection*, *data analysis* and *results and discussion*.

3.1. Driving Simulator Selection and Setup

As was mentioned under Section 2.3, concluding all past literature that has been reviewed, a simulator study is suited for the purpose of data collection for this project. A few factors are considered before making the decision of collecting data through simulation.

One such factor is time constraints. With roughly 14 weeks for planning and another 14 weeks for execution of this project, data collection through video surveillance is not viable. This is because raw data collected from video footage within roughly 10 weeks is not only insufficient but the data extraction process and analysis that follows makes this method unsuitable for the given timeline.

Another aspect to consider is the financial and permission access. As this is a student project, there is no external fundings or organization supporting this project. Therefore, collecting data through the controlled road study method is also non-feasible as renting cars and test track would involve a significant amount of money.

Therefore, it can be said that a simulator study is suited for the purpose of this project as it is the most accessible and allows individual, independent variables to be tested. Multiple driving simulators such as CARLA, OpenDS, TORCS and Udacity's self-driving car simulator were reviewed, but ultimately CARLA is chosen to be used to design the test scenarios. One of the main reasons that CARLA is chosen over the other driving simulators is due to it being open-source, free and has a community that uses it, which provides basic teaching materials in using the driving simulator.

CARLA is an open-source simulator for autonomous driving research. CARLA was chosen as it provides all the necessary functions and tools that would help in designing and generating the test scenarios. Some of the functions that CARLA allows is for the user to generate and create their own maps, control traffic conditions and weather and many more. CARLA also provides written documentation and a rather active GitHub repository that discusses on issues faced. This would be particularly beneficial as spending time to learn a complex software, given the timeline, is not ideal.

With CARLA chosen, the next step is to setup and build CARLA in order for the it be used to design the scenarios that are to be tested. CARLA offers its driving simulator in 2 different versions, namely the Release version and the Source Build version. The release version can be comparable to that of an application that is readily available to be downloaded and is not customizable nor editable. The source build version can be understood as the developer/editable/customizable version of CARLA, where users can change and edit the world.

The release version was setup by first downloading the release file from here, followed by installing the dependencies as well as the client library used by CARLA. CARLA runs on a client-server architecture whereby the server, which is essentially the world that containing all buildings, roads, traffic lights etc. will have to be started first followed by the client, the driver of the controlled vehicle, which connects to the server.

The source build can be split into 2 main parts. The first part is to install all the prerequisites required to build CARLA. The prerequisites are shown in the list below:

(a) CMake
(b) Git
(c) Make
(d) Windows 8.1 SDK from Visual Studio 2019
(e) x64 Visual C++ Toolset from Visual Studio 2019
(f) .NET framework 4.6.2 from Visual Studio 2019
(g) Unreal Engine

The second part is the build part where the CARLA GitHub repository is cloned. Next, the assets that are used within CARLA as well as Unreal Engine is downloaded. The Unreal Engine environment variable is then set as all edits on the CARLA world is done through the Unreal Engine editor. Lastly, the Python API client and the server is complied. Once the compilation is done, editing and customization can be done through the Unreal Engine editor.

3.2. Scenario Design

This section covers the different scenarios that would be used within the CARLA simulator to test on independent variables. These variables are external factors (factors that are not controlled by the participant) which would be tested through different scenarios based on the variables tested. It is hypothesize that external circumstances have an influence on driving behaviour. Thus, the purpose of these variables is to test if driving behaviour, when approaching a yellow traffic light, of an individual would change under different external circumstances.

One of the variables to be tested is traffic intensity. It is hypothesize that heavier traffic (high intensity) would affect an individual's driving behaviour negatively. Negative behaviour is defined as driving behaviour that breaks or more likely to break traffic rules, which in this case is red-light-running. This is because an individual stuck in heavier traffic may have a higher tendency to run a yellow light as they refused to wait for another cycle to cross the intersection. However, the inverse may hold as well, since drivers might run yellow light if there is minimal to no traffic around as it is commonly observed with motorcyclist.

Weather condition is another variable that would be tested through the simulated scenarios. Weather conditions can be varied through different means, however, in this implementation only the time of day and presence of rain is varied. The time of day is varied between Day and Night, which affects illumination whereas the presence of rain is varied between present and absent.

With these variables, a total of 8 scenarios were used to test on these variables. The variables that are tested along with the scenarios that are produces are summarized in Table 1 and Table 2.

Table 1: Variable Table for Simulator Study

Variable	**Variation**	
Traffic Intensity	Low	High
Time of Day	Day	Night
Rain	Present	Absent

Table 2: Scenarios for Simulator Study

Scenarios	**Traffic Intensity**	**Time of Day**	**Rain**
Scenario 1	Low	Day	Absent
Scenario 2	Low	Day	Present
Scenario 3	Low	Night	Absent
Scenario 4	Low	Night	Present
Scenario 5	High	Day	Absent
Scenario 6	High	Day	Present
Scenario 7	High	Night	Absent
Scenario 8	High	Night	Present

3.3. Data Collection

This section covers on how the data will be collected and what data are being collected. The first sub-section will discuss about the procedures to obtain the data through volunteering participants, followed by what data collection and extraction process.

3.3.1. Procedure

The study starts with recruiting volunteers to driving within a driving simulator as part of the data collection process. Volunteers (whom at this point are participants of the study) will be asked to fill in a form, with questions pertaining to their demographics, driving experience and how close does the simulator simulate real-world driving scenarios.

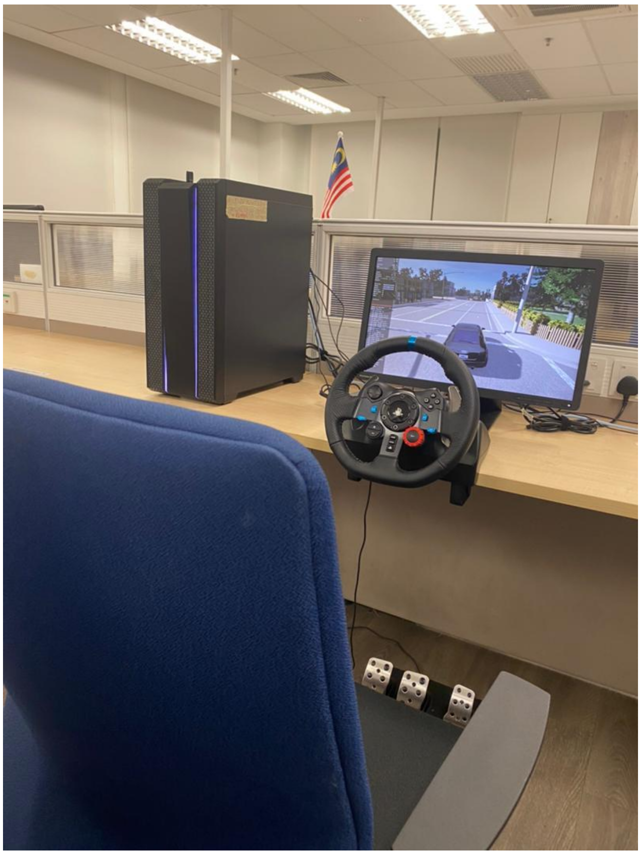


Figure 3: Driving Simulator Used

Figure 3 shows the driving simulator that was used for all participants. The car used within the simulator was also kept constant.

Participants will be briefed about the study, but not the aim and objective specifically. This is to maintain integrity of the data collected as participants may not reflect their normal driving behaviour if they are told the specified objectives prior. Participants will then be allowed to drive freely using the driving simulator for 5-10 minutes to get themselves accustomed to the driving simulator. Then, participants will be asked to drive within the world varied with the aforementioned 8 scenarios. Each of the scenarios are tested at least 3 times and then averaged in order to minimize the effect of possible outliers. This would result in a total of at least 24 data points from a single participant.

In the meantime, the entire duration of the participant driving within the simulator is screen recorded using the Open Broadcaster Software (OBS). This is done in order for the experimenter to rewatch the video afterwards to extract the required data.

Furthermore, in order to prevent participants from inferring the objectives through multiple runs at a traffic light intersection, the entire course of driving will include roundabouts and normal city driving with some degree of lane changing. Another reason that “noise” scenarios are added as part of the driving course is to reduce anticipation and expectation of the participant. This is because the expectation of drivers approaching traffic light might results in different decision-making behaviour in general [13], and as a result affect the data collected.

3.3.2. Data Extraction

A total of 6 participants were recruited for the purpose of data collection. This subsequently produced 6 video recordings of the participants driving in the driving simulator through the 8 aforementioned scenarios. The experimenter then goes through these recordings and extract the necessary data into an Excel sheet whenever the participant encounters a yellow traffic light intersection.

The data that are extracted are the vehicle speed, distance to intersection, brake/gas pedal usage and the stopping/non-stopping behaviour. All of these data are collected as these are the factors would generally be affected when a decision is demanded from the driver, that is when there is a change in the phase of the traffic light.

The **vehicle speed** (in km/h) is collected twice, the initial speed of the vehicle and the final speed of the vehicle. The initial speed of the vehicle is defined as the speed of the vehicle upon yellow light onset. The final speed of the vehicle is defined as the speed of the vehicle when it completely crosses the intersection. The final speed will be 0 if the driver decides to stop upon yellow light onset.

The **distance to intersection** is recorded in the form of car lengths. This is because when people drive, the depth perception and distance is not processed in terms of meters. Units are used mainly for calculation and reporting purposes. Therefore, the car lengths can be converted to meters whereby 1 car length is 4.5 meters.

Brake/gas pedal usages are also recorded as these are the main factors that control the motion of the vehicle. Therefore, these dependant variables are also important in understanding driving behaviour. It should be noted that the pedal pressures were not recorded but instead whether the participant decided to accelerate more, brake, or lift the gas pedal upon yellow light onset.

Lastly, **stopping or non-stopping behaviour** is also recorded as a binary value, indicating whether the participant has stopped or decided to cross the intersection. If the participant decides to run the yellow light, another value is also recorded to understand if that particular yellow light running behaviour resulted in a red-light-running violation. This would indicate that the driving overestimated and should have stopped when within the dilemma zone.

3.4. Data Analysis

This section explains how the data that has been collected and extracted are used. The data will first be cleaned as pre-processing. Then, statistical methods will be implemented in order to understand the data.

3.4.1. *Data Cleaning*

With the data that was extracted into the Excel sheet from the videos, data cleaning has to be performed in order to prepare the data for analysis. Some categorical variables such as the brake/gas pedal usage and stopping/non-stopping behaviour was recoded into ordinal data. Numerical data such as the vehicle speed and distance to intersection remain as they are.

Since brake/gas pedal usage can only take 3 values and have some degree of ordinance to it, the data can be recoded to 0, 1 and 2 whereby,

0 represents, acceleration
1 represents, lift of gas pedal/no pedal usage
2 represents, brake.

In general, the value from this responding variable can then be understood as higher value representing higher tendency to slow down and stop upon yellow light onset.

Similarly, the stopping or non-stopping behaviour can also be recoded into values of 0 and 1. With 0 representing non-stopping behaviour and 1 representing stopping. This is to correspond to the brake/gas pedal usage whereby higher values represent high tendency to stop.

3.4.2. *Data Understanding*

To understand the data, statistical approaches were used. For each of the scenarios, the average for each parameters will be calculated and recorded. This is done to have an overview of the data collected.

3.5. Results and Discussion

The average initial speed, final speed, distance to intersection and brake/gas pedal usage for each of the scenarios are tabulated in Table 3. The numbering of the scenarios follow that of Table 2.

Table 3: Summary of Results from Simulator Study

Scenarios	**Initial Speed**	**Final Speed**	**Distance**	**Brake/Gas**
Scenario 1	35.89	23.83	1.86	1.39
Scenario 2	36.11	32.06	1.67	1.11
Scenario 3	38.83	42.06	1.67	0.72
Scenario 4	36.06	34.50	0.89	1.00
Scenario 5	41.94	28.78	2.08	0.83
Scenario 6	37.44	23.33	1.58	1.22
Scenario 7	37.61	32.94	0.83	0.94
Scenario 8	39.33	27.61	1.06	0.78

As mentioned above, only a total of 6 participants participated in the simulator study. These participants have an age range of 21-34. Due to the small sample size and narrow range of demographics, these results should be noted to be unrepresentative of the driving population.

One of the main reasons for the small sample size is due to the difficulty in recruiting participants. This is because a single session of data collection will require around 1 hour (5-10 minutes to familiarize with the simulator, 40-50 minutes to data collect) from the participant. The required time to be committed by participants for data collection became a barrier to recruit participants.

Besides that, the extra time needed to extract the data from the video recording of each participant's simulator driving was also a contributing factor to the small sample size. Given the timeline, this method proved to not feasible to have the initial target sample size of 30 participants. As such, an alternative method of data collection was considered. Taking into account of the review on past literatures, questionnaire survey was chosen as the second, and ultimately primary, data collection method.

4. Method 2 (Questionnaire Survey)

This section discusses in detail the alternative method used for data collection as well as the results of the analysis from the data collected. A questionnaire containing images of driving scenarios that are approaching a yellow traffic light intersection was created and sent out to the masses for data collection. The collected data was process and analysed in similar methods as those used in Section 3.4. Finally, the result of the analysis is presented and discussed. Therefore, there are 4 main sub-sections to explain this method, the *questionnaire, data collection, data analysis* and *results and discussion.*

4.1. Questionnaire

To create a questionnaire used for data collection, Google Form was used. This is because it is a common tool used nowadays for questionnaires and provides comprehensive types of questions with checks in place. There are a total of 4 main sections to the questionnaire, starting with the **consent form**, **demographics**, **driving experience** and **yellow traffic light scenarios**. The complete form can be found here. The scenarios questions and their respective responses in the form of histograms can be found in Appendix.

Under the **demographics** sections, respondents are asked of their age, gender and whether if they hold a valid driving license. Age and gender are asked as it is used during data analysis to identify if age and gender has any correlation to driving styles. Respondents who do not hold a valid driving license is also filtered out through this section. The survey will end for respondents who do not hold a valid driving license.

For respondents who hold a valid driving license, they will be brought to the next section, **driving experience**. Respondents will be asked a few questions related to their driving experience and their perceptions of speed (what is slow, moderate and fast) in this section. Years driving, frequency of driving within a week and rough estimate of lifetime mileage were asked to gauge the respondents driving experience. Respondents are also asked on a speed range that they consider as slow, moderate and fast in an urban city setting i.e., within Kuala Lumpur city. Lastly, respondents are asked if they have ever been involved in a traffic light intersection accident.

The questions mentioned above are simplified and summarized into Table 4 and Table 5.

Table 4: Demographics Questions

Question Number	**Question Details**	**Sample Response**
1	Age	22
2	Gender	Male
3	Holds a license?	Yes

Table 5: Driving Experience Questions

Question Number	**Question Details**	**Sample Response**
1	Years Driving	5
2	Frequency of driving in a week	6 to 7 days
3	Lifetime mileage	50,000km – 100,000km
4	Slow speed range	20 – 40
5	Moderate speed range	40 – 70
6	Fast speed range	60 – 90
7	Involved in traffic light accident	No

Once respondents complete the demographics and driving experience section, they would be brought to the final section, **yellow traffic light scenarios**. These yellow traffic light scenarios aim to replicate those of real-world driving where people would approach a yellow traffic light. The images used in all of the scenarios are screenshots taken from the CARLA simulator at a first person perspective.

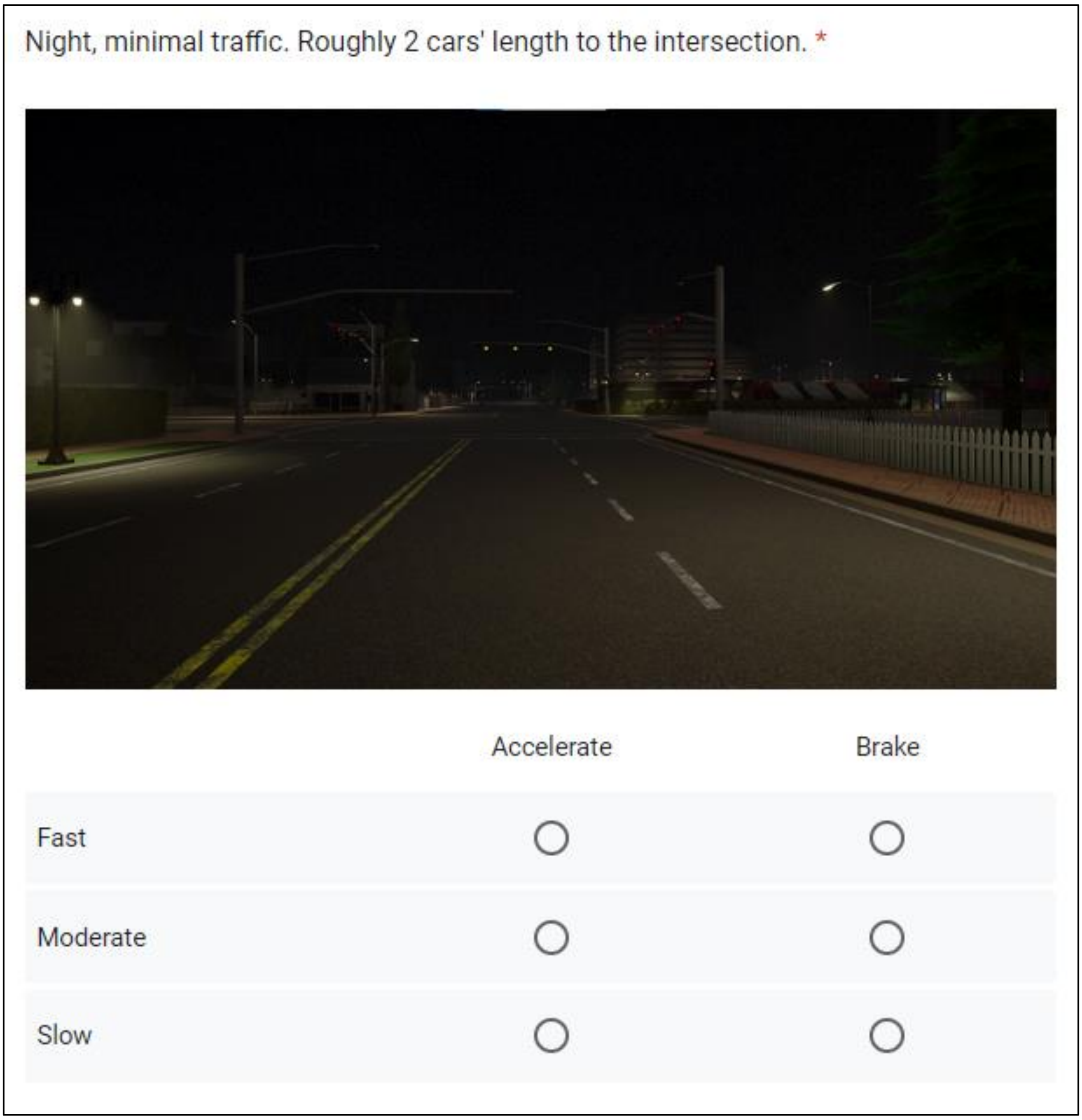
14 scenes were created to test the respondents reaction to different external circumstances (weather, traffic intensity, distance to intersection). Within each of these scenes, participants would have to assume that they are approaching each of these scenes at different approaching speeds (low, medium, high), as they have defined themselves in the previous section. This resulted in 42 scenarios/features (14 scenes, each with 3 approaching speeds) to be analysed. The 14 scenes are summarized in Table 6.

The reason that the approaching speed is not fixed for each scenes is because the assumption that everyone goes through a yellow light intersection at the same speed is not valid. Besides that, humans tends to feel the speed of approaching a yellow light intersection instead of noticing the actual value (in km/h). Hence, the questionnaire design of different approaching speeds range, defined by the respondents, is justified.

Table 6: Scenes Summary

Scene	**Time of Day**	**Rain**	**Traffic**	**Distance**	**Others**
1	Day	Absent	Low	1	-
2	Day	Absent	High	1	-
3	Day	Absent	High	2	-
4	Day	Present	Low	1	-
5	Day	Present	High	1	-
6	Day	Absent	High	0.5	Cars ahead running yellow light
7	Day	Absent	High	2	Cars ahead running yellow light
8	Night	Absent	Low	1	-
9	Night	Absent	High	1	-
10	Night	Absent	Low	2	-
11	Night	Present	Low	1	-
12	Night	Present	High	1	-
13	Day	Absent	High	2	Countdown timer of 3 seconds
14	Day	Absent	High	1	Countdown timer of 3 seconds

There are only 2 responses available to the respondents in each scenario, to either accelerate through the yellow light intersection or brake. An example question is shown in Figure 4.



Night, minimal traffic. Roughly 2 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○

Figure 4: Example Question

4.2. Data Collection

The questionnaire is then shared through contacts and social medias such as Facebook and Instagram for better reach.

A total of 347 responses was collected over the span of 10 days. After filtering and validation, only 334 responses was valid to be carried forward into data analysis. It should be noted that to ensure data integrity, each respondent was required to sign into the Google account before being able to complete the questionnaire. This is to ensure that each respondent can only complete the questionnaire once, assuming that they only have one Google account.

4.3. Data Analysis

This section discusses on how the data collected from the questionnaires are processed and analysed to extract any possible information. The data is first cleaned by converting all categorical/ordinal data into integers. This is followed by data understanding to understand the mean demographics and driving experience of the respondents. Clustering was also done to identify any possible clusters driving behaviour.

Python and Pandas were used in the process of the data cleaning and analysis. Matplotlib was used as a visualization tool to visualize the results which will be further discussed in Section 4.4.

4.3.1. Data Cleaning

Similar to the data cleaning process mentioned above in Section 3.4.1, categorical data that have some degree of ordinance to it is converted to numerical codes. This is to allow for processing afterwards, as mean calculation and clustering models are limited to numeric inputs only. The categorical data that were converted are gender, frequency of driving in a week, lifetime mileage and responses to each scenarios. These categorical data and their recoded values are summarized in Table 7, Table 8, Table 9 and Table 10.

Table 7: Gender Recode

Gender	**Recoded Value**
Female	0
Male	1
Prefer not to say	2

Table 8: Frequency of driving Recode

Frequency of driving in a week (in days)	**Recoded Value**
0 to 2 days	0
3 to 5 days	1
6 to 7 days	2

Table 9: Mileage Recode

Estimated Lifetime Mileage	**Recoded Value**
<10000 km	0
10000- 50000 km	1
50000 - 100000 km	2
>100000 km	3

Table 10: Scenario Responses Recode

Scenario Response	**Recoded Value**
Accelerate	0
Brake	1

Unreasonable speed ranges were also filtered out in this step. For example, some respondents will in a speed range of 80-999 km/h for their perceived fast speed range. Such responses are filtered and removed as current production cars are unable to reach a speed of 999 km/h.

4.3.2. Data Understanding

To understand the data, measures of central tendency were applied. In most cases, the mean is taken as a representation of that data. This can take either the form of numerical data as in the case of age, or a mean class as in the case of lifetime mileage, since the response provided in the questionnaire is a class. Wherever applicable, a histogram is also plotted to visualize the distribution of the data. These values are then taken as a baseline value to be compared to after clustering.

The perceived speed ranges for fast, moderate and slow was also further processed for information. The mode class, average upper and lower boundary for each speed range was calculated. This is to understand at what speed constitutes as fast, moderate and slow for the respondents, as this would subsequently help in understanding the different approaching speeds to the 14 scenes.

4.3.3. Clustering

Finally, k-means clustering is applied on the dataset. A total of 4 clustering were performed:

1. 14 scenes, 3 different approaching speeds. 42 features in total.
2. 14 scenes, only fast approaching speeds. 14 features in total.
3. 14 scenes, only moderate approaching speeds. 14 features in total.
4. 14 scenes, only slow approaching speeds. 14 features in total.

The reason that k-means clustering was performed on the above selection is to understand if different approaching speeds will have an influence on driving behaviour.

The value of k for each k-means clustering was optimized by calculating the within cluster sum of squared distance (WCSS) and the elbow plot. This allows the optimal k value to be determined with the right number of clusters and a minimal WCSS value. To ensure that each clustering run is the same, the random state parameter of KMeans function fixed to be 15.

Once k-mean clustering is complete, the centroid of each cluster is plotted to identify the characteristics, if any, of that particular cluster. For example, for the clustering done only on scenes with slow approaching speed, there might be a cluster with centroid of values closer to 1. This would indicate that there a group of **careful** drivers when approaching a yellow traffic light intersection at slow speed, since a value of 1 in scenario responses indicates the use of brake as mentioned in Table 10. It should be noted that k-means clustering do not guarantee an observable characteristic for each cluster.

After each cluster is identified, the demographics and driving experience of the cluster is again calculated. This is used to be compared with the baseline demographics and driving experience calculated in Section 4.3.2. If a significant difference is found between a cluster's demographics and driving experience and the baseline demographics and driving experience, demographics and driving experience can be said to have an influence on an individual's driving behaviour, particularly in those cluster that have an observed characteristic.

4.4. Results and Discussion

This section presents the data from the data analysis steps taken above and discusses on the possible reasons and justification for observations from the data. The baseline demographics and driving experience will be first presented, followed by discussions on how approaching speed, demographics and driving experience may have an influence on driving behaviour.

4.4.1. Distribution and Baseline of Data

For demographics, the mean age of the respondents is found to be 25.63 years old. The distribution for age can be seen in Figure 5. It should be noted that the data distribution for age is positively skewed with most of the respondents in the age group of 20-30.

The mean value for gender, after recoded, is 0.467. This can be interpreted as 46.7% of the respondents are male and 53.3% of the respondents are female. Although, people who prefer not to reveal their gender is recoded to be 2, the value is negligible as there are only 3 respondents out of 334 that chose not to reveal their gender.

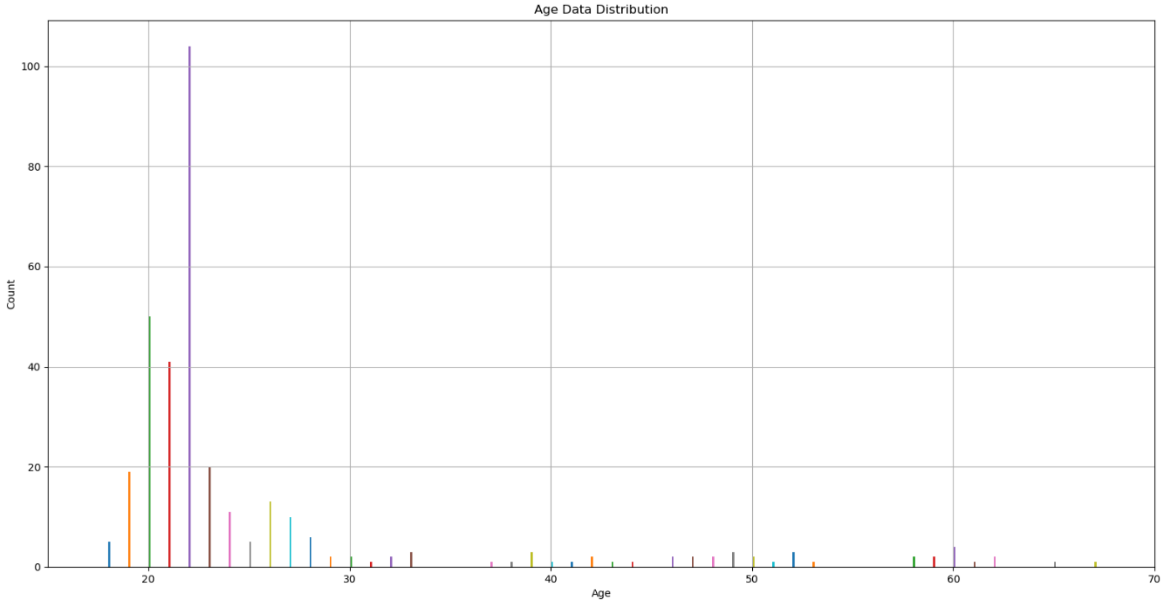


Figure 5: Age Data Distribution

For driving experience, the mean driving years of the respondents are 7 years with the distribution of the data shown in Figure 6. Similar to age, driving years are also positively skewed.

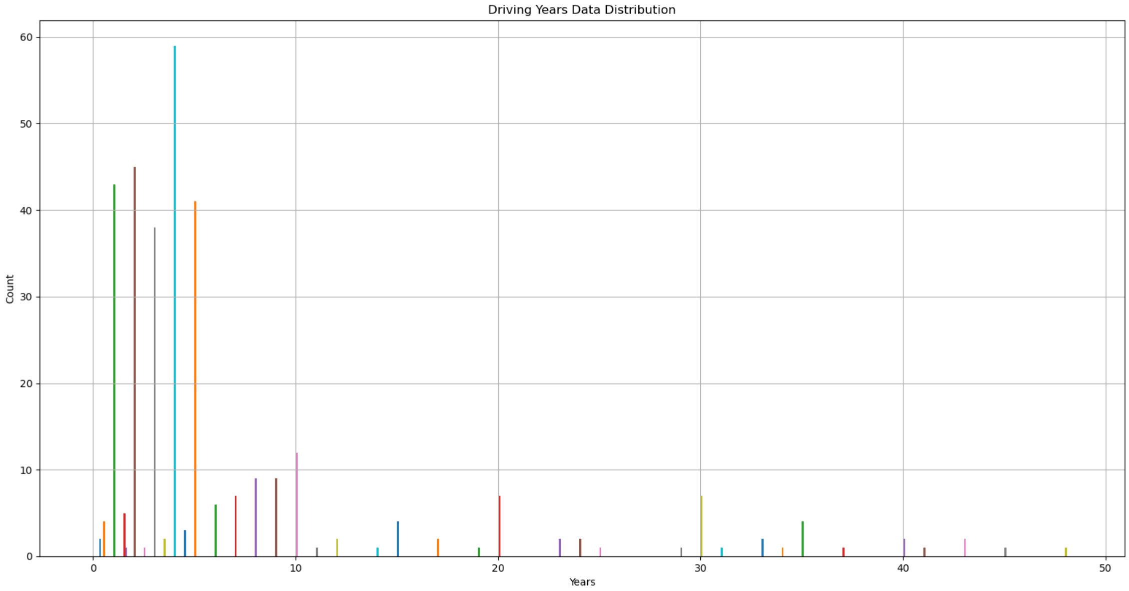


Figure 6: Driving Years Data Distribution

The mean value for frequency of driving in a week, after recoded, is 1.04. This can be interpreted as, on average, the respondents drives 3-5 days a week. The distribution is shown in Figure 7. As can be seen from Figure 7, the distribution of frequency of driving in a week is relatively even, however the value of 1.04 will still be used as the baseline for comparison.

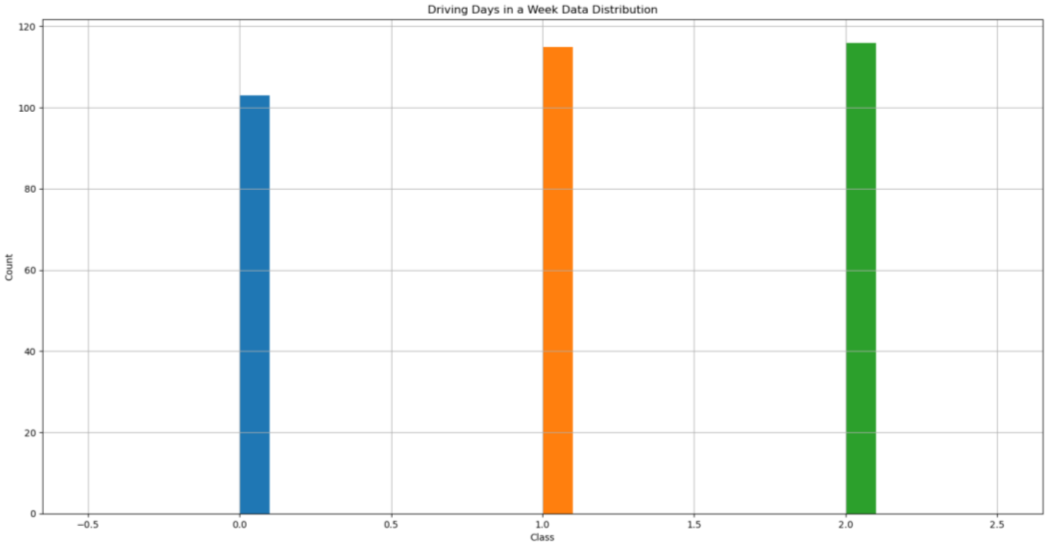


Figure 7: Driving Frequency Class Distribution

The mean value for lifetime mileage, after recoded, is 0.99. This can be interpreted as, on average, the respondents have a lifetime mileage of less than 10,000 km. The distribution is shown in Figure 8.

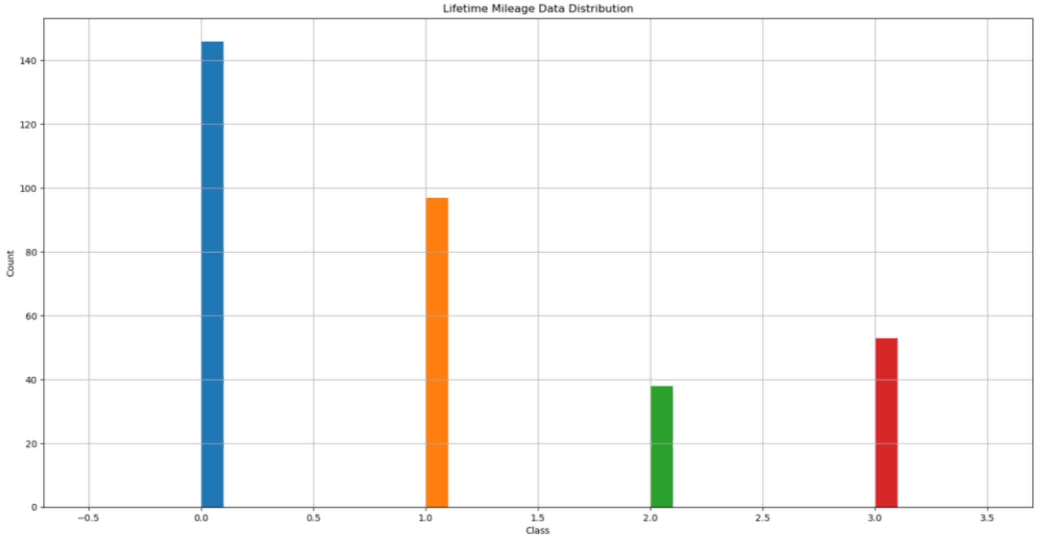


Figure 8: Lifetime Mileage Class Distribution

Based on the speed ranges of the respondents, a slow speed range is between 24 – 46 km/h whereas a moderate speed range is between 52 – 75 km/h. A fast speed range in an urban city setting based on the response are 85 – 117 km/h. All of these values are obtained by averaging the minimum speed and maximum speed of each speed range.

The distribution of the speed class for each speed range is shown in Figure 9. The size of point corresponds to the frequency. The mode class for slow speed range is 20 – 40 km/h, moderate speed range is 60 – 80 km/h and fast speed range is 80 – 100 km/h. A consistent positive gradient is also observed starting from the slow speed range (red) to fast speed range (blue).

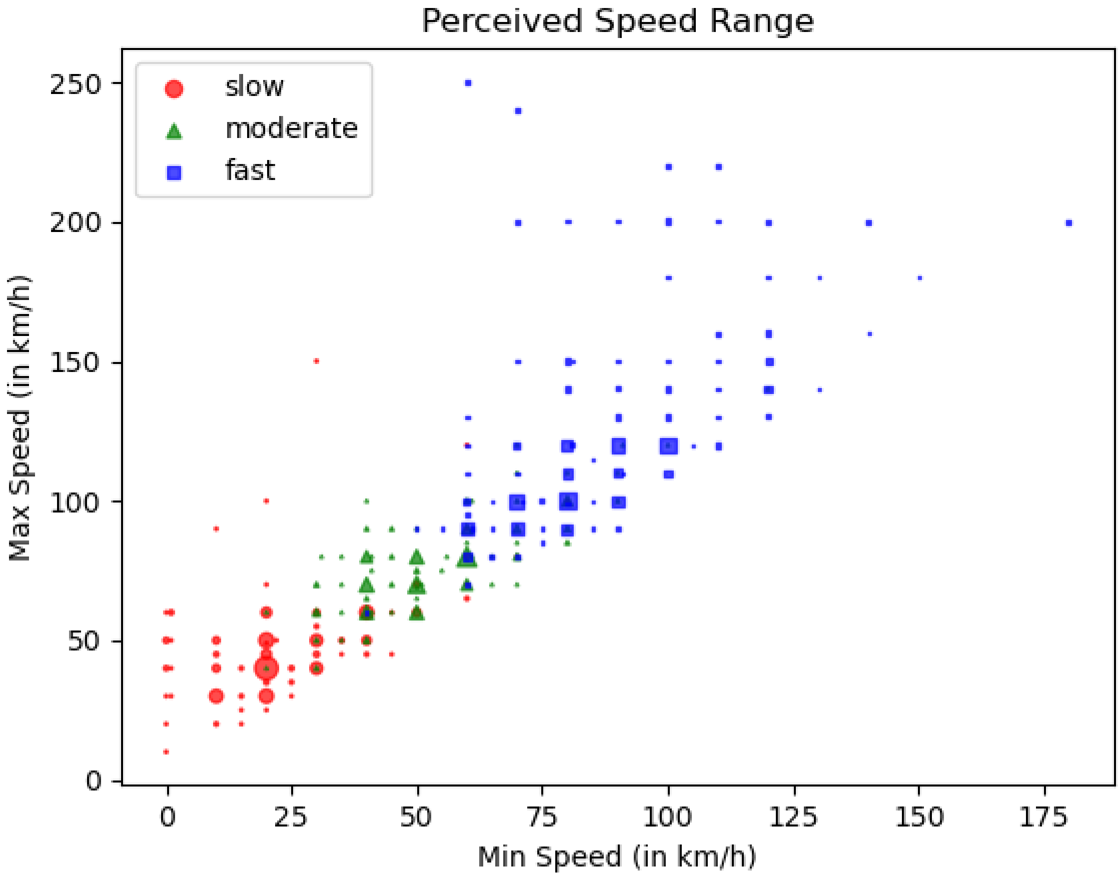


Figure 9: Perceived Speed Range Distribution

All of the above parameters mentioned above collectively forms the baseline for comparison after clustering is completed. These baseline parameters are summarized in Table 11.

Table 11: Summary of Baseline Parameters

Baseline Parameters	**Recoded Value (Actual Class)**
Age	25.63 years old
Gender	0.467
Years Driving	7 years
Frequency of driving in a week	1.04 (3-5 days)
Lifetime mileage	0.99 (<10,000 km)

4.4.2. Clustering Results

The first k-means clustering performed on all 42 scenarios had 4 clusters. As mentioned in Section 4.3.3, the k value of 4 is obtained through WCSS and the elbow graph as shown in Figure 10.

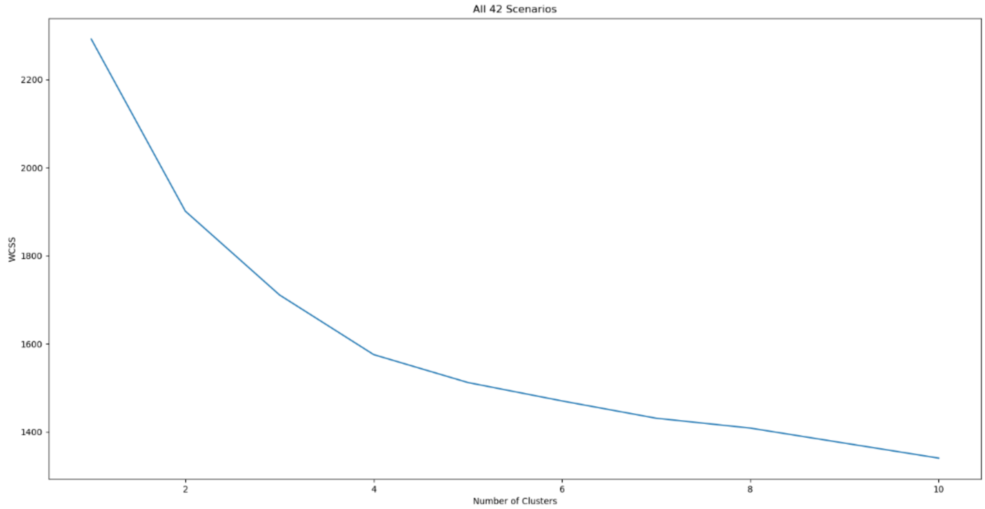


Figure 10: Elbow Graph for All 42 Scenarios

Of the 4 clusters, only one cluster had an observable characteristic. This cluster is labelled as Cluster 0 by the KMeans function found from the sklearn.cluster library. The cluster center (centroid) value for each scenarios is plotted and shown in Figure 11.

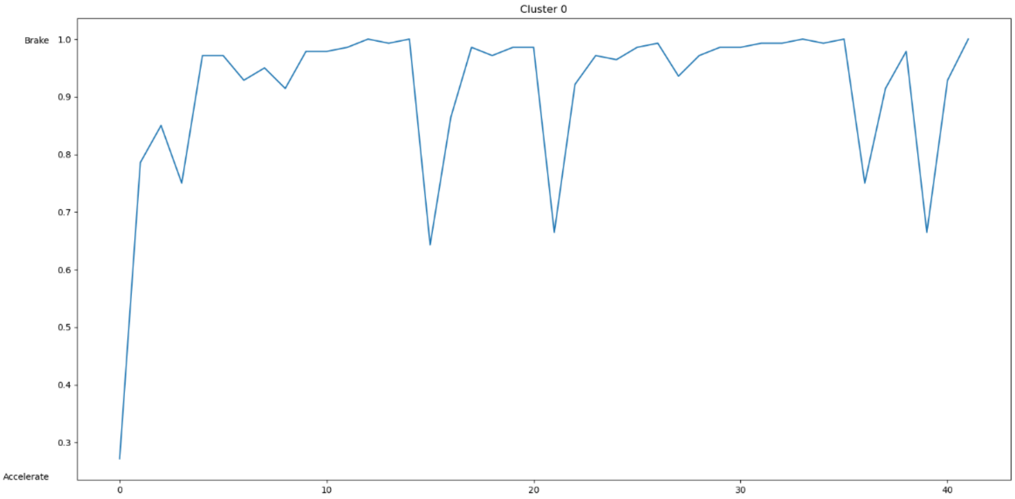


Figure 11: Cluster 0 Centroid Values

As it can be observed from Figure 11, respondents in this cluster have a high tendency to brake across almost all scenarios. Averaging all of the points in Figure 11 produced a value of 0.913, which can be interpreted, this cluster of respondents will brake 91.3% of the time when approaching a yellow traffic light. Therefore, respondents in this cluster can be labelled as careful drivers.

Further understanding of the respondents of this cluster revealed a total of 140 respondents out of 334. The demographics and driving experience of this cluster of respondents are summarized in Table 12. These values do not present significant difference when compared to the baseline parameters in Table 11, apart from the average age being 1 year old and having 10% less male respondents in this cluster. Driving experience have little to no effect on driving behaviour in this case.

Table 12: Demographics and Driving Experience of Cluster 0 (42 Scenarios)

Demographics and Driving Experience	**Value**
Age	26.69
Gender	0.379
Years Driving	7.67
Frequency of driving in a week	1.04
Lifetime mileage	0.97

Next, k-means clustering on the 14 fast scenarios also yielded 4 clusters. Of these 4 clusters, only 2 clusters have observable characteristics as shown in Figure 12.

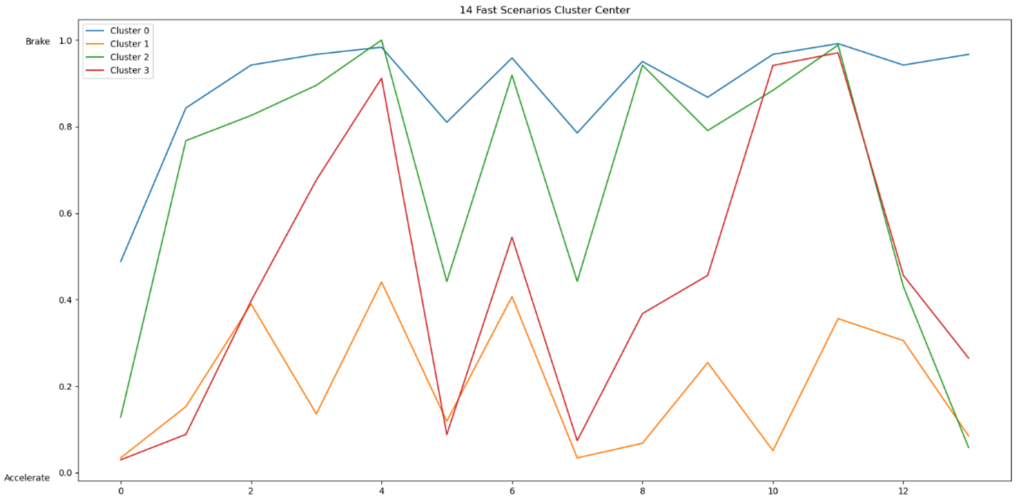


Figure 12: Cluster Center Values for 14 Fast Scenarios

These clusters are Cluster 0 (blue line) and Cluster 1 (yellow line). Both of these clusters can be described as the opposite of each other. Cluster 0 can be labelled as careful drivers with a cluster center average value of 0.890 whereas Cluster 1, having a cluster center average value of 0.202 can be labelled as risky drivers.

This indicates that Cluster 0 of fast scenarios would brake 89% of the time averagely and Cluster 1 of fast scenarios would only brake 20.2% of the time averagely when approaching a yellow light intersection. However, it should be noted that Cluster 0 have a population of 121 whereas Cluster 1 only has 59 respondents out of 334, indicating that majority of the respondents are still careful and risky drivers are a minority.

To understand these clusters in depth, the demographics and driving experience of respondents of these clusters are analysed and summarized in Table 13 and Table 14.

For Cluster 0, the discriminating factors is the age and gender whereby there are more females and having a higher average age compared to the baseline parameters. Similarly, the discriminating factors are also age and gender for Cluster 0 but on the opposite end of the spectrum whereby there are more males and a lower average age in this cluster.

Table 13: Demographics and Driving Experience of Cluster 0 (14 Fast)

Demographics and Driving Experience	**Value**
Age	26.86
Gender	0.388
Years Driving	7.61
Frequency of driving in a week	1.02
Lifetime mileage	0.95

Table 14: Demographics and Driving Experience of Cluster 1 (14 Fast)

Demographics and Driving Experience	**Value**
Age	24.47
Gender	0.678
Years Driving	6.44
Frequency of driving in a week	1.24
Lifetime mileage	1.08

For k-means clustering on the 14 moderate approaching speed scenarios, the results are similar to that of clustering on the fast scenarios. A total of 4 clusters were produced with 2 having observable characteristics on both ends of the spectrum, a cluster of careful drivers and another cluster of risky drivers. This is shown in the Figure 13.

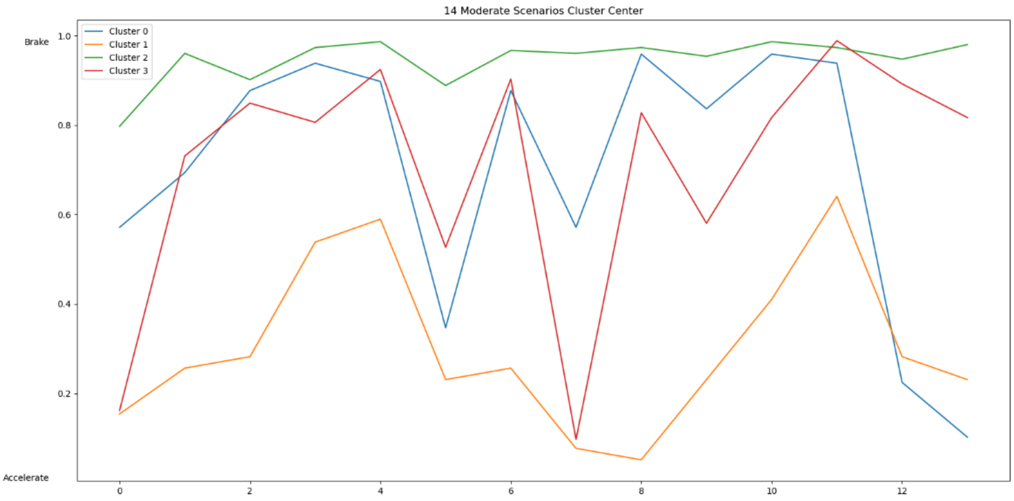


Figure 13: Cluster Center Values for 14 Moderate Scenarios

In this case, Cluster 1 (yellow line) can be labelled as the risky drivers with a cluster center average value of 0.302 where Cluster 2 (green line) can be labelled as the careful drivers with cluster center average value of 0.946. Similar to the fast scenarios, risky drivers are the minority with only a population of 39, to the careful drivers of 153 which remains the biggest cluster.

Investigating into the demographics and driving experience of these clusters revealed that the largest difference between Cluster 1 and the baseline parameters is the gender. There are 10% more males in this cluster than the baseline. On the other hand, Cluster 2's discriminatory factor is the age and driving years. However, the increase of driving years compared to the baseline may be related to the increase in average age of this cluster. These values are summarized into Table 15 and Table 16.

Table 15: Demographics and Driving Experience of Cluster 1 (14 Moderate)

Demographics and Driving Experience	**Value**
Age	25.53
Gender	0.564
Years Driving	7.38
Frequency of driving in a week	0.95
Lifetime mileage	1.01

Table 16: Demographics and Driving Experience of Cluster 2 (14 Moderate)

Demographics and Driving Experience	**Value**
Age	27.14
Gender	0.399
Years Driving	8.11
Frequency of driving in a week	1.09
Lifetime mileage	1.01

K-means clustering on the slow scenarios produced a total of 3 clusters, each with their own observable characteristics. However, Cluster 2 (green line) will not be further discussed as it is rather erratic and possibly inconsistent given the tooth-like shape of the graph. This is displayed in Figure 14.

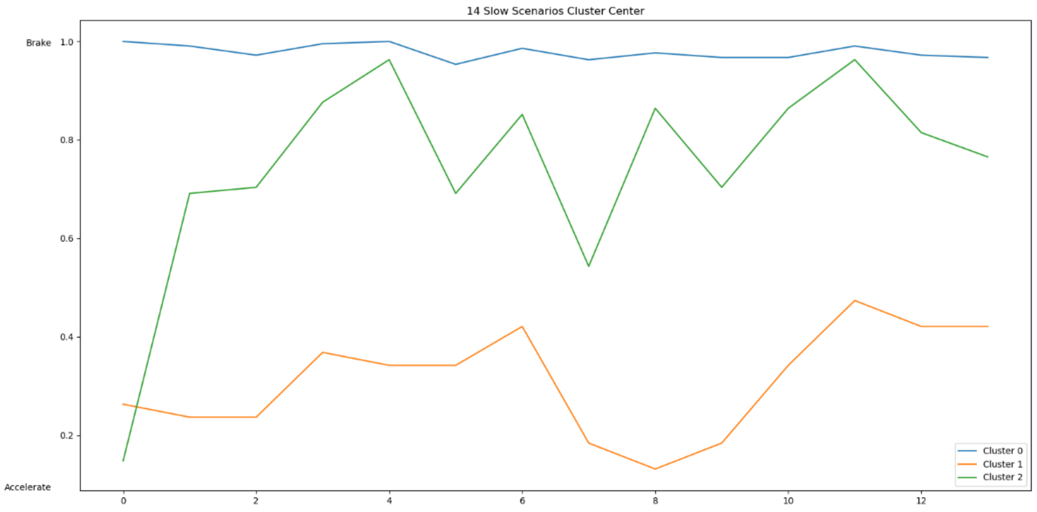


Figure 14: Cluster Center Values for 14 Slow Scenarios

Of these 3 clusters, Cluster 0 (blue line) can be labelled as the careful drivers as the cluster center average value is 0.978 for this cluster. On the other hand, Cluster 1 (green line) can be labelled as the risky drivers with a cluster center average value of 0.312. Similar to the cluster size of fast and moderate scenarios, careful drivers are the majority with 215 respondents in this category whereas risky drivers are the minority with only 38 respondents.

The demographics and driving experience of the respondents in Cluster 0 of 14 slow scenarios have little to no difference to the baseline parameters. This would indicate that this is likely to be natural response of the majority of the driving population.

Conversely, Cluster 1 (risky drivers) have lesser overall driving experience compared to the baseline and are also on average younger. Gender is the only factor that is in line with the baseline parameter.

Table 17 and Table 18 summarizes the demographics and driving experience of Cluster 0 and Cluster 1.

Table 17: Demographics and Driving Experience of Cluster 0 (14 Slow)

Demographics and Driving Experience	**Value**
Age	25.59
Gender	0.460
Years Driving	7.04
Frequency of driving in a week	1.09
Lifetime mileage	0.99

Table 18: Demographics and Driving Experience of Cluster 1 (14 Slow)

Demographics and Driving Experience	**Value**
Age	23.97
Gender	0.473
Years Driving	5.42
Frequency of driving in a week	0.68
Lifetime mileage	0.66

The significant clusters as a result of clustering is summarize in Table 19. Identifying the main contributing factors influencing driving behaviour, particularly careful driving style and risky driving style.

Table 19: Summary of clustering results

Cluster (Scenarios)	**Label**	**Cluster Center Average**	**Difference to baseline parameters (Table 11)**
Cluster 0 (All 42)	Careful	0.913	- Older - More females
Cluster 0 (14 Fast)	Careful	0.890	- Older - More females
Cluster 1 (14 Fast)	Risky	0.202	- Younger - More males
Cluster 1(14 Moderate)	Risky	0.302	- More males
Cluster 2 (14 Moderate)	Careful	0.946	- Older - More females
Cluster 0 (14 Slow)	Careful	0.978	- No significant difference
Cluster 1 (14 Slow)	Risky	0.312	- Younger - Less driving experience

From Table 19, it is apparent that demographics, both age and gender, do influence an individual's driving behaviour and style. Additionally, clusters that are labelled to be careful, regardless of their approaching speeds, tend to be older on average and consist of more females compared to the baseline. On the other hand, clusters that are labelled risky have either a younger age on average, consists of more males or both. Driving experience seems to have minimal influence on driving behaviour.

4.4.3. Approaching Speed

For approaching speeds, the number of responses for braking and accelerating will be compared between the speed range to understand if it influences driving behaviour. Table 20 tabulates the total responses of the respondents based on different approaching speeds. All scenarios are also included as a means of control.

Table 20: Total Response of Accelerate/Brake

Scenarios	**No. of Accelerate**	**No. of Brake**	**Brake Percentage**
All 42	3555	10473	74.6
14 Fast	1757	2919	62.4
14 Moderate	1080	3596	76.9
14 Slow	718	3958	84.6

Table 20 suggests that as approaching speed increases, driver's would tend to brake lesser and chooses to run the yellow light. However, it should be noted that the difference in the observed driving behaviour is most likely caused by the shift of dilemma zone (due to the changes in approaching speed) instead of approaching speed itself.

For example, given the same exact scenario, where it is a sunny day with high traffic and the vehicle is 1 car lengths away from the intersection upon yellow light onset. If the vehicle is approaching at a moderate speed, it would be in the dilemma zone. However, if the vehicle is approaching at a fast speed, it would be in the should-go zone and subsequently contribute to less occurrences of braking.

5. Relationship between collected data of Method 1 & 2

Based on the speed data collected from Method 1, and the speed range analysis done with the data collected in Method 2 it can be said that on average, the approaching speed of the simulator study are all within the slow speed range. The average distance to intersection for each scenarios were also calculated. With the knowledge of speed range and average distance to intersection, the closest scenarios in terms of similarity between Method 1 and Method 2 can be identified. From there, the response towards these scenarios can be used to further validate any possible relationships or factors relating to driving behaviour.

Table 21 tabulates the closest corresponding scenarios between Method 1 and Method 2.

Table 21: Corresponding scenarios between Method 1 and Method 2 and their responses

Method 1	Method 2
Scenario 1	Question 1
Scenario 2	Question 4
Scenario 3	Question 8
Scenario 4	Question 11
Scenario 5	Question 3
Scenario 6	Question 5
Scenario 7	Question 9
Scenario 8	Question 12

With this, we can further test if external circumstances (traffic intensity, time of day, rain) affects driving behaviour. At the same time, the responses of different data collection method can also be compared, by comparing the responses from Method 1 and Method 2.

Table 22, Table 23 and Table 24 will tabulate traffic intensity, time of day and presence of rain respectively. The values within each table should be interpreted where 1 indicates brake and 0 indicates accelerate. Therefore, values closer to 1 indicates higher tendency/percentage of drivers that would brake and vice versa. These values are obtained by summing up all the values (either 0 or 1) under the response column for scenarios that fit the criteria, then dividing by the total count.

Table 22: Traffic Intensity response between Method 1 and Method 2

Traffic Intensity	**Method 1**	**Method 2**
Low	0.48	0.81
High	0.54	0.88

Table 23: Time of day response between Method 1 and Method 2

Time of Day	**Method 1**	**Method 2**
Day	0.63	0.84
Night	0.39	0.86

Table 24: Presence of rain response between Method 1 and Method 2

Presence of rain	**Method 1**	**Method 2**
Present	0.54	0.91
Absent	0.48	0.79

From the tables above, it can be observed that between Method 1 and Method 2, Method 2 is the more reliable data collection method in this case. This is because the values from Method 1 are not indicative of either acceleration or braking, the values are mostly within the range of 0.4-0.6, which does not provide any substantial information as it is between the 2 values. This is largely due to the extremely small sample size of less than 20 data points per scenario compared to more than 300 data points per scenario for Method 2.

Another observation that can be made is the effect of external circumstances on driving behaviour, particularly traffic intensity and presence of rain. The presence of rain would result in a 12% increase of braking behaviour while approaching yellow traffic light. This may be due to slippery surface caused by rain, results in reduced friction and grip between the car tyres and road surface. Thus, drivers tend to driver more carefully and take less risk in accelerating through yellow traffic lights.

A 7% increase in braking behaviour is also observed when the traffic intensity is higher. This is possibly due to the fear of being involved in a traffic light accidents with other road users. This is because accelerating through a yellow traffic light would mean taking the risk running a red light, where the right of way would have been given to other traffic directions.

6. Conclusion

In conclusion, the aim of understanding driving behaviour when approaching yellow traffic signal was achieved along with identifying clusters of respondents (drivers) with particular driving characteristics. These characteristics are either careful drivers or risky drivers.

2 data collection methods were used as means of data collection, simulator study and questionnaire survey, with questionnaire being the primary data collection method. Comparing the 2 methods, simulator study is closer to real-world driving as different participants approached the same traffic light at different speeds and have different reaction time. Simulator study has the potential to collect a wider variety of data such as the brake and gas pedal pressure and reaction time upon yellow traffic light onset. The scenarios encountered in simulator studies are also comparatively more naturalistic, with the possibility of other cars running red light, cars ahead driving unnecessarily slowly and traffic light phase change in the middle of the intersection.

On the other hand, questionnaire response allows for mass volumes of data collection in a relatively short timeframe. However, the data that can be collected is rather limited and may not be accurately representative of a person's driving behaviour. This is due to social-desirability bias where respondents to questionnaires might give socially desirable answers. In the context of this project, it would be responding with brake more often than they naturally would.

Using the data collected, some factors that influences driving behaviour were identified. Age and gender is found to have a significant influence on driving behaviours whereas driving experience have little to no influence. Through clustering, 2 clusters of driving characteristics were identified, the careful cluster which brakes in most scenarios and the risky cluster which accelerates in most scenarios. Of the 2 clusters, the careful cluster of drivers are averagely older and also consists of more females as opposed to the risky cluster of drivers who are averagely younger and consists of more males. Based on these findings, it can be concluded that older drivers and/or female drivers tend to drive more carefully and brake more often when approaching a yellow traffic light.

Another finding from this project is the effect of external circumstances on driving behaviour. Of the 3 factors tested, time of day, traffic intensity and presence of rain, only traffic intensity and presence of rain had a significant impact on driving behaviour. For traffic intensity, a higher traffic intensity induced more braking in the respondents, indicating a more careful behaviour. Presence of rain is found to have the largest influence on driving behaviour, also inducing a more careful behaviour when rain is present.

There are no conclusive findings to state that approaching speeds have influence on driving behaviour. This is because as approaching speed varies, the dilemma zone also varies which influences behaviour. Therefore, the difference in stop/go decision for approaching speeds cannot be solely attributed to the approaching speed. However, the perception of speed ranges for fast, moderate and slow in an urban city were obtained. The general perception of slow is between 24 – 46 km/h, moderate to be 52 – 75 km/h and fast to be 85 – 117 km/h.

The main limitation with data collection using questionnaire surveys is the social desirability bias. Although the extent of the effect of this bias is unknown, it cannot be neglected. Therefore, future data collection using questionnaire surveys should look into possible ways of reducing the bias. Besides that, future work can also look in the direction of the relationship between approaching speeds and dilemma zone in order to further understand the influence if any on driving behaviour. This would allow for better traffic light designs that improves safety and efficiency.

Reference

[1] "Ministry of Transport Malaysia Official Portal Road Accidents and Fatalities in Malaysia", *Mot.gov.my*, 2021. [Online]. Available: https://www.mot.gov.my/en/land/safety/road-accident-and-facilities.

[2] "Malaysia Number of Registered Vehicles, 1996 – 2021 | CEIC Data", *Ceicdata.com*, 2021. [Online]. Available: https://www.ceicdata.com/en/indicator/malaysia/number-of-registered-vehicles.

[3] S. KULANTHAYAN, W. PHANG and K. HAYATI, "TRAFFIC LIGHT VIOLATION AMONG MOTORISTS IN MALAYSIA", *IATSS Research*, vol. 31, no. 2, pp. 67-73, 2007. Available: 10.1016/s0386-1112(14)60224-7.

[4] Z. Yang, X. Tian, W. Wang, X. Zhou and H. Liang, "Research on Driver Behavior in Yellow Interval at Signalized Intersections", *Mathematical Problems in Engineering*, vol. 2014, pp. 1-8, 2014. Available: 10.1155/2014/518782.

[5] H. Rakha, A. Amer and I. El-Shawarby, "Modeling Driver Behavior within a Signalized Intersection Approach Decision–Dilemma Zone", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2069, no. 1, pp. 16-25, 2008. Available: 10.3141/2069-03.

[6] G. Lu, Y. Wang, X. Wu and H. Liu, "Analysis of yellow-light running at signalized intersections using high-resolution traffic data", *Transportation Research Part A: Policy and Practice*, vol. 73, pp. 39-52, 2015. Available: 10.1016/j.tra.2015.01.001.

[7] L. Rittger, G. Schmidt, C. Maag and A. Kiesel, "Driving behaviour at traffic light intersections", *Cognition, Technology & Work*, vol. 17, no. 4, pp. 593-605, 2015. Available: 10.1007/s10111-015-0339-x.

[8] J. Li, X. Jia and C. Shao, "Predicting Driver Behavior during the Yellow Interval Using Video Surveillance", *International Journal of Environmental Research and Public Health*, vol. 13, no. 12, p. 1213, 2016. Available: 10.3390/ijerph13121213.

[9] O. Taubman-Ben-Ari, M. Mikulincer and O. Gillath, "The multidimensional driving style inventory—scale construct and validation", *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 323-332, 2004. Available: 10.1016/s0001-4575(03)00010-1.

[10]A. Furnham and J. Saipe, "Personality correlates of convicted drivers", *Personality and Individual Differences*, vol. 14, no. 2, pp. 329-336, 1993. Available: 10.1016/0191-8869(93)90131-l.

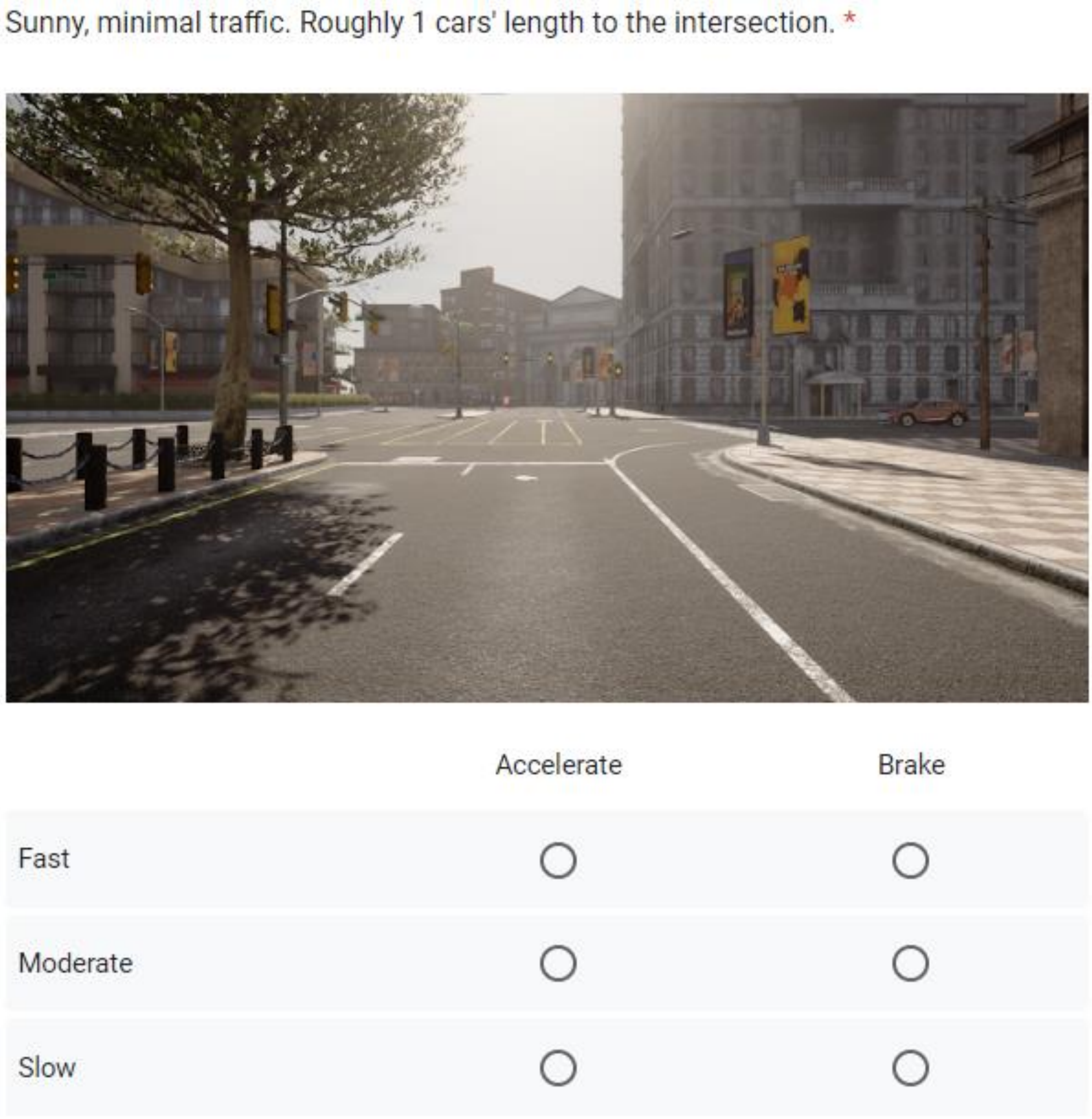
[11]D. FRENCH, R. WEST, J. ELANDER and J. WILDING, "Decision-making style, driving style, and self-reported involvement in road traffic accidents", *Ergonomics*, vol. 36, no. 6, pp. 627-644, 1993. Available: 10.1080/00140139308967925.

[12]T. Lajunen and H. Summala, "Driving experience, personality, and skill and safety-motive dimensions in drivers' self-assessments", *Personality and Individual Differences*, vol. 19, no. 3, pp. 307-318, 1995. Available: 10.1016/0191-8869(95)00068-h.

[13]R. Horst, *Driver decision making at traffic signals*. 1988, pp. 93-97. Available: http://onlinepubs.trb.org/Onlinepubs/trr/1988/1172/1172-012.pdf

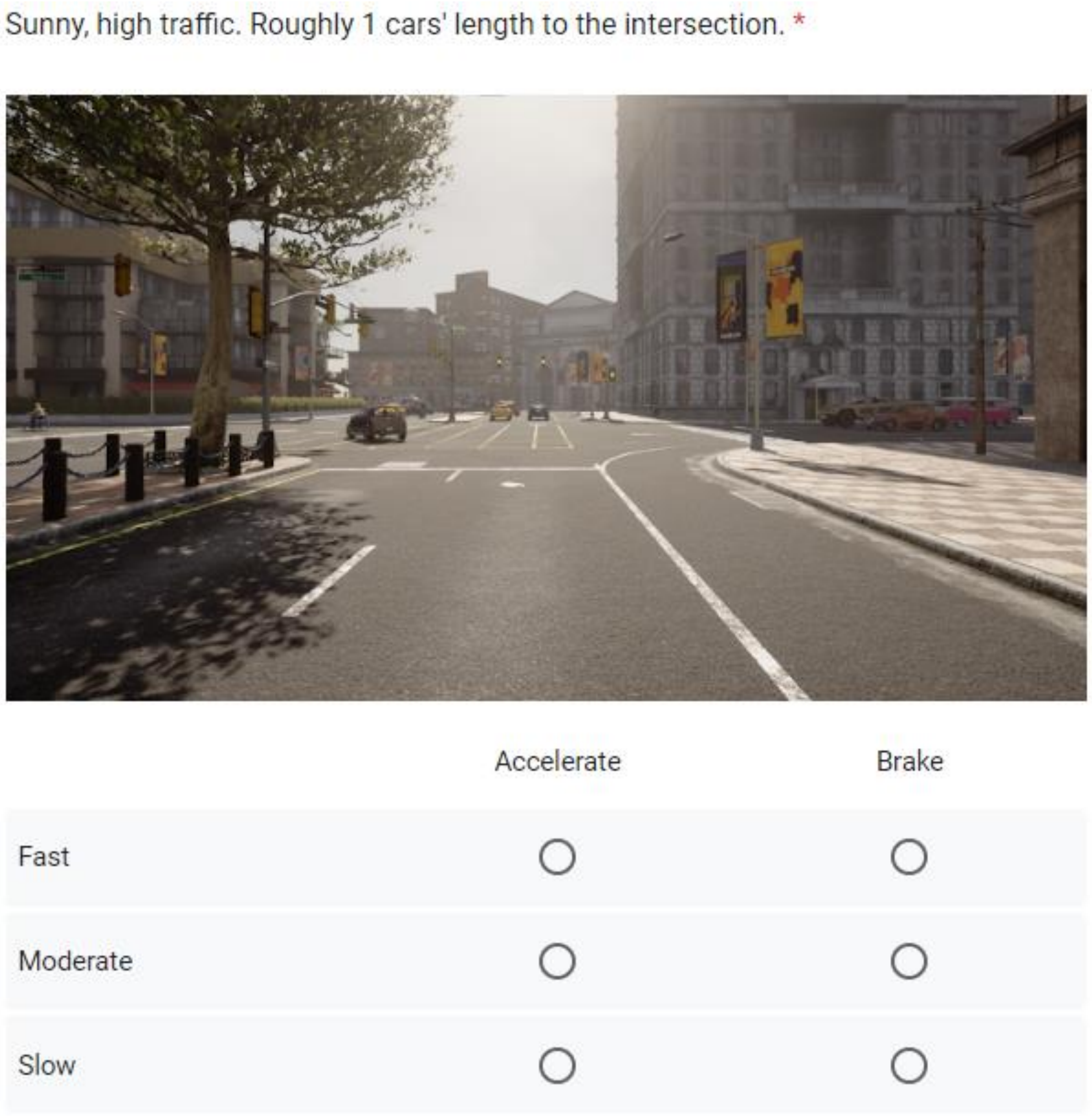
[14]P. Li, Y. Li and X. Guo, "A Red-Light Running Prevention System Based on Artificial Neural Network and Vehicle Trajectory Data", *Computational Intelligence and Neuroscience*, vol. 2014, pp. 1-11, 2014. Available: 10.1155/2014/892132.

Appendix



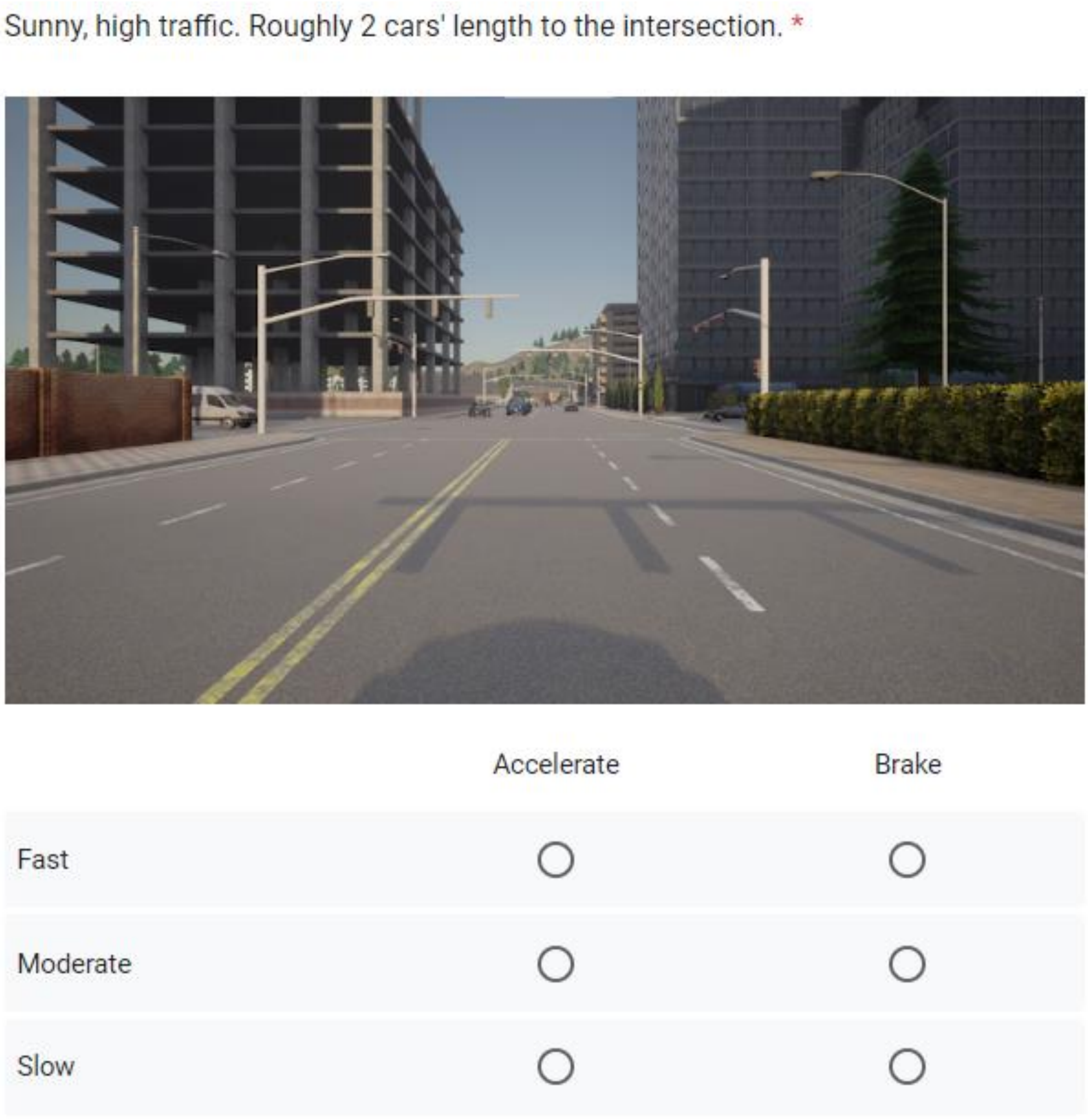
Sunny, minimal traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



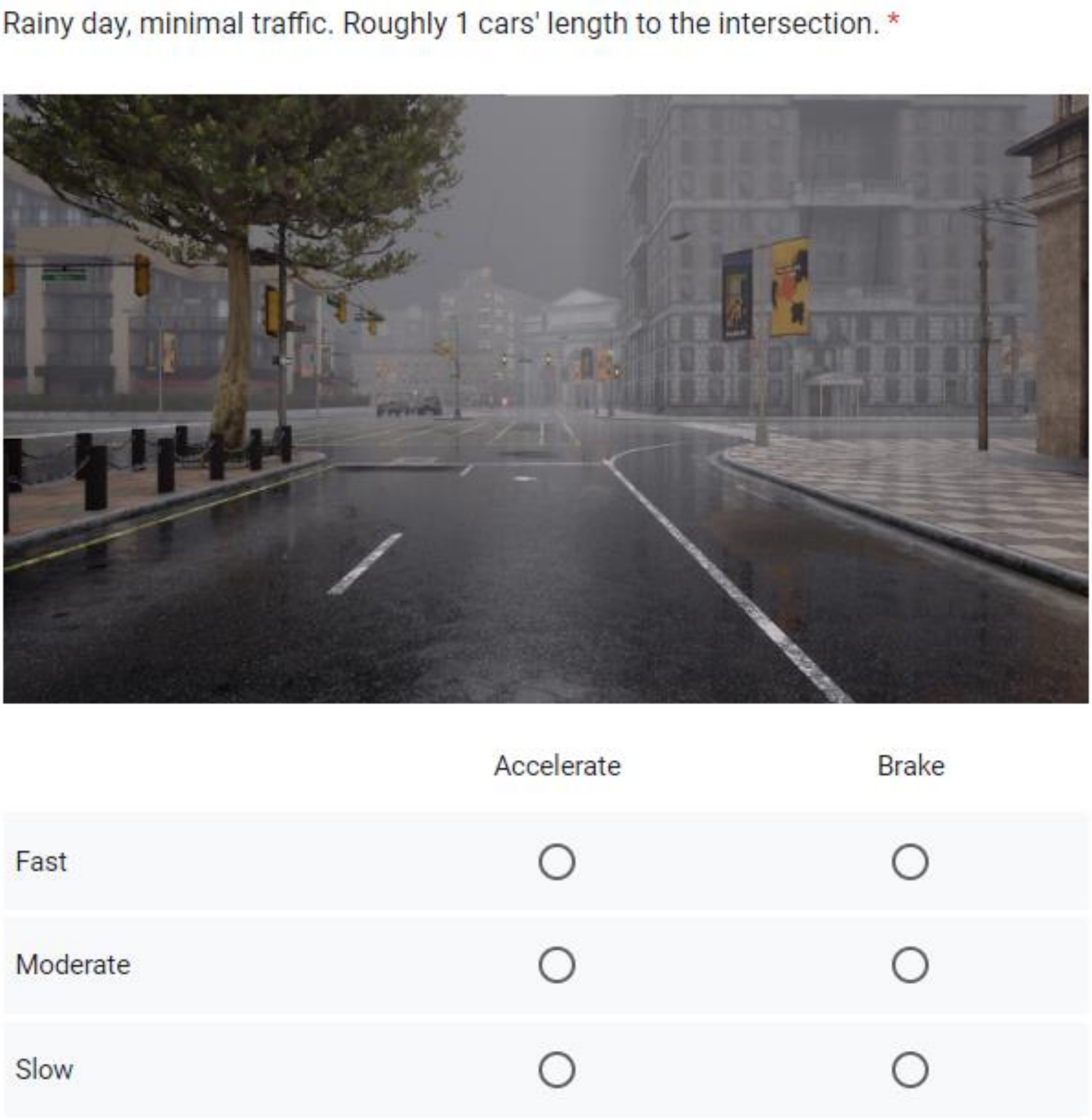
Sunny, high traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



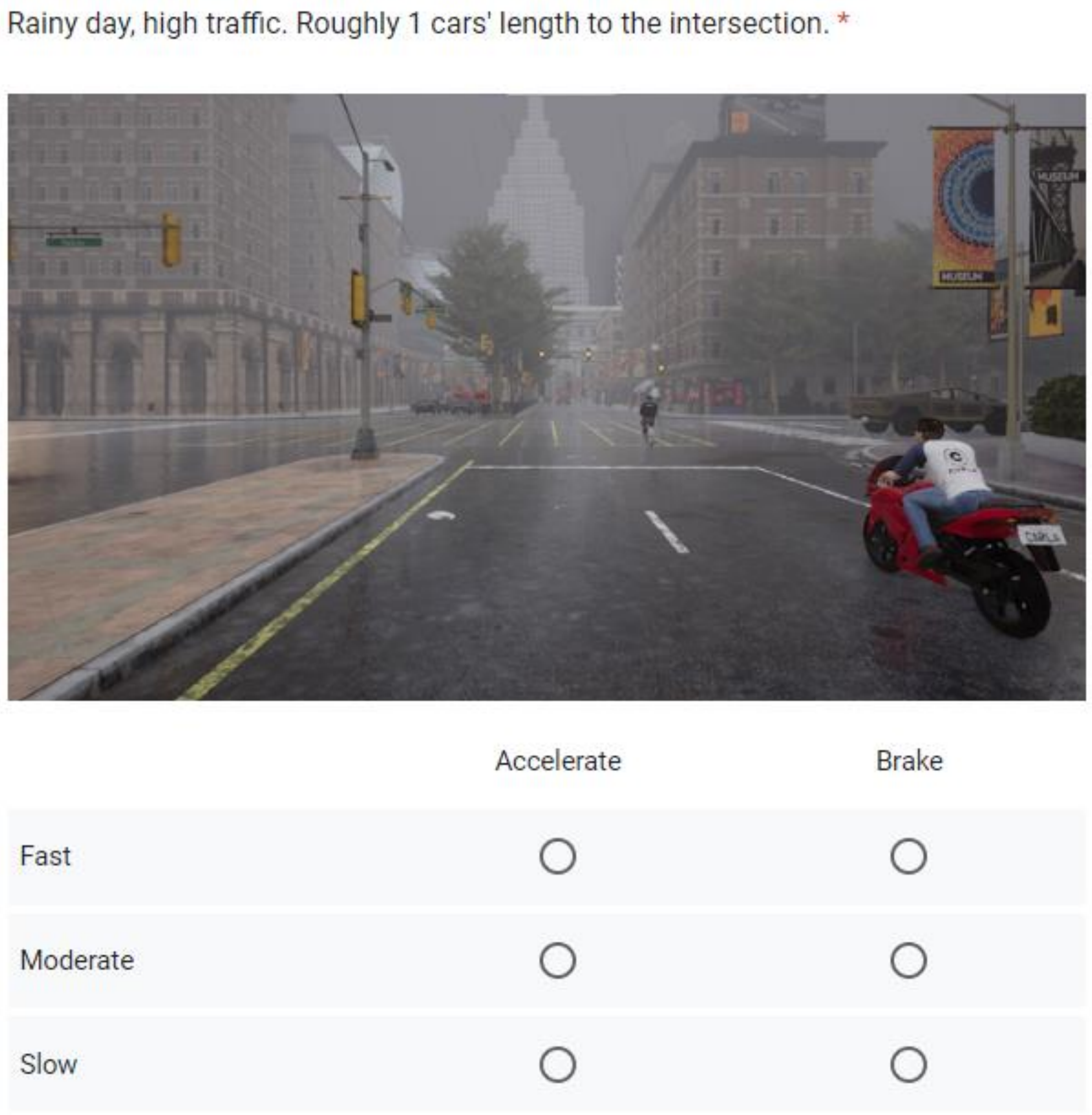
Sunny, high traffic. Roughly 2 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



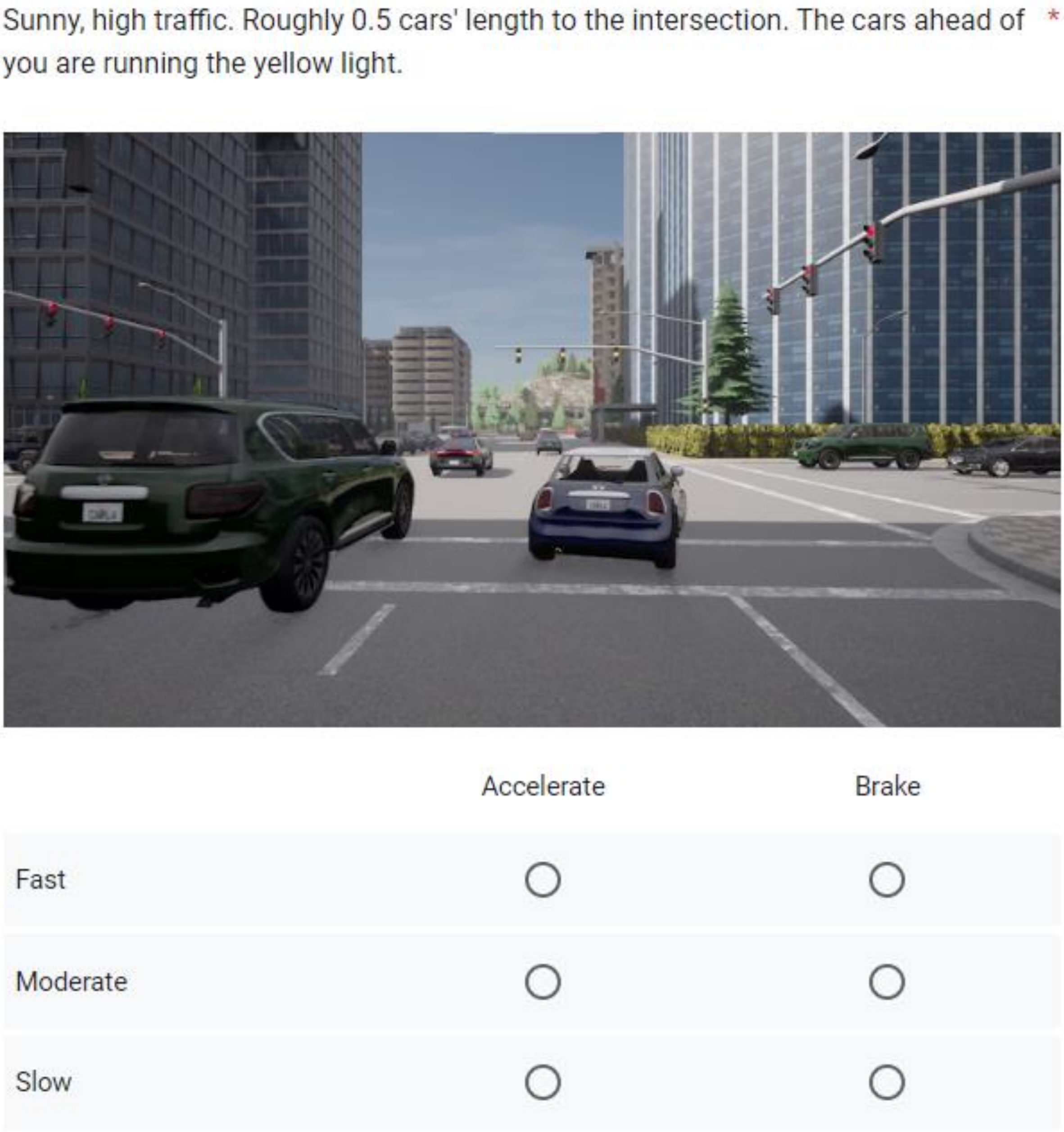
Rainy day, minimal traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



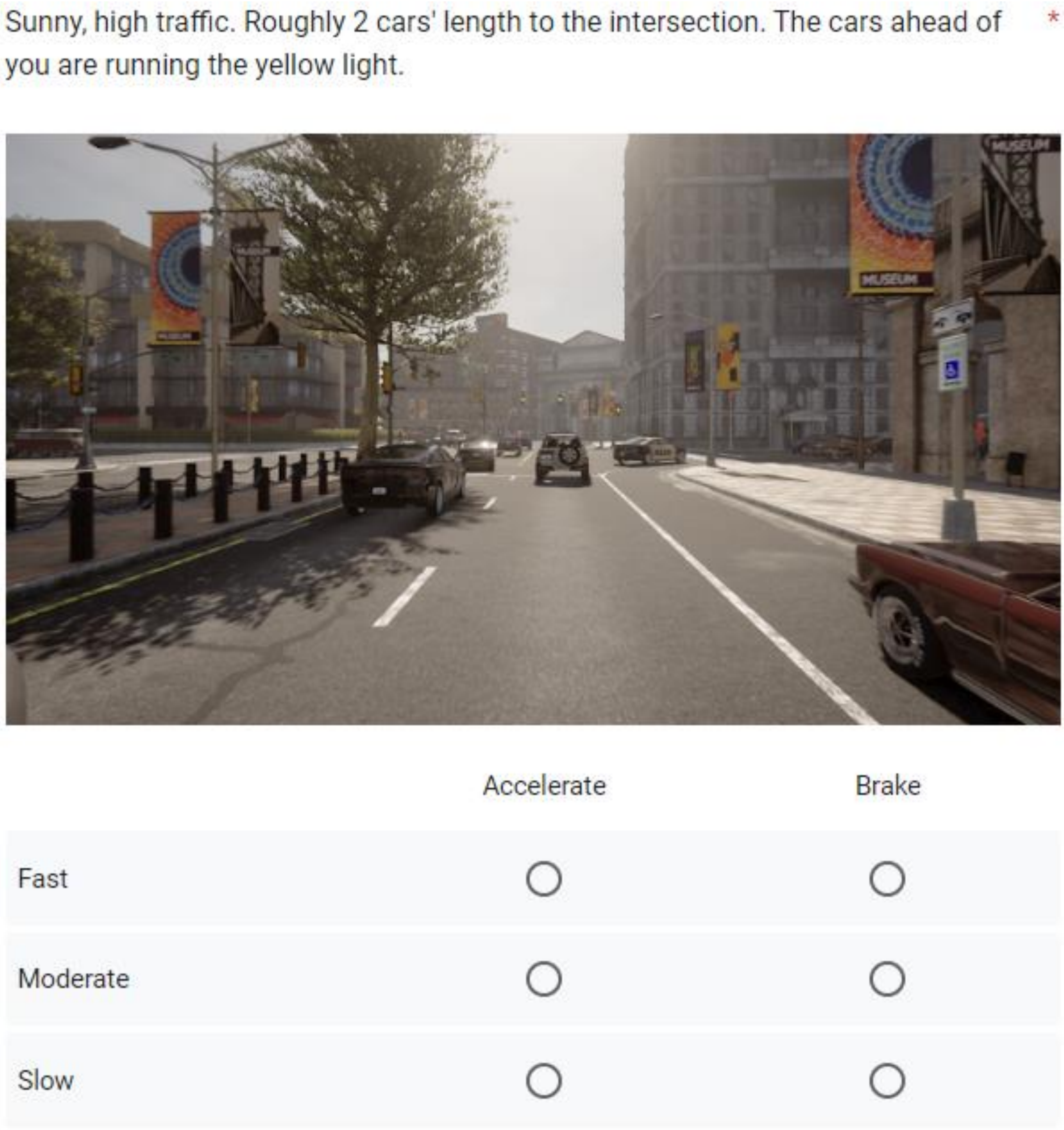
Rainy day, high traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



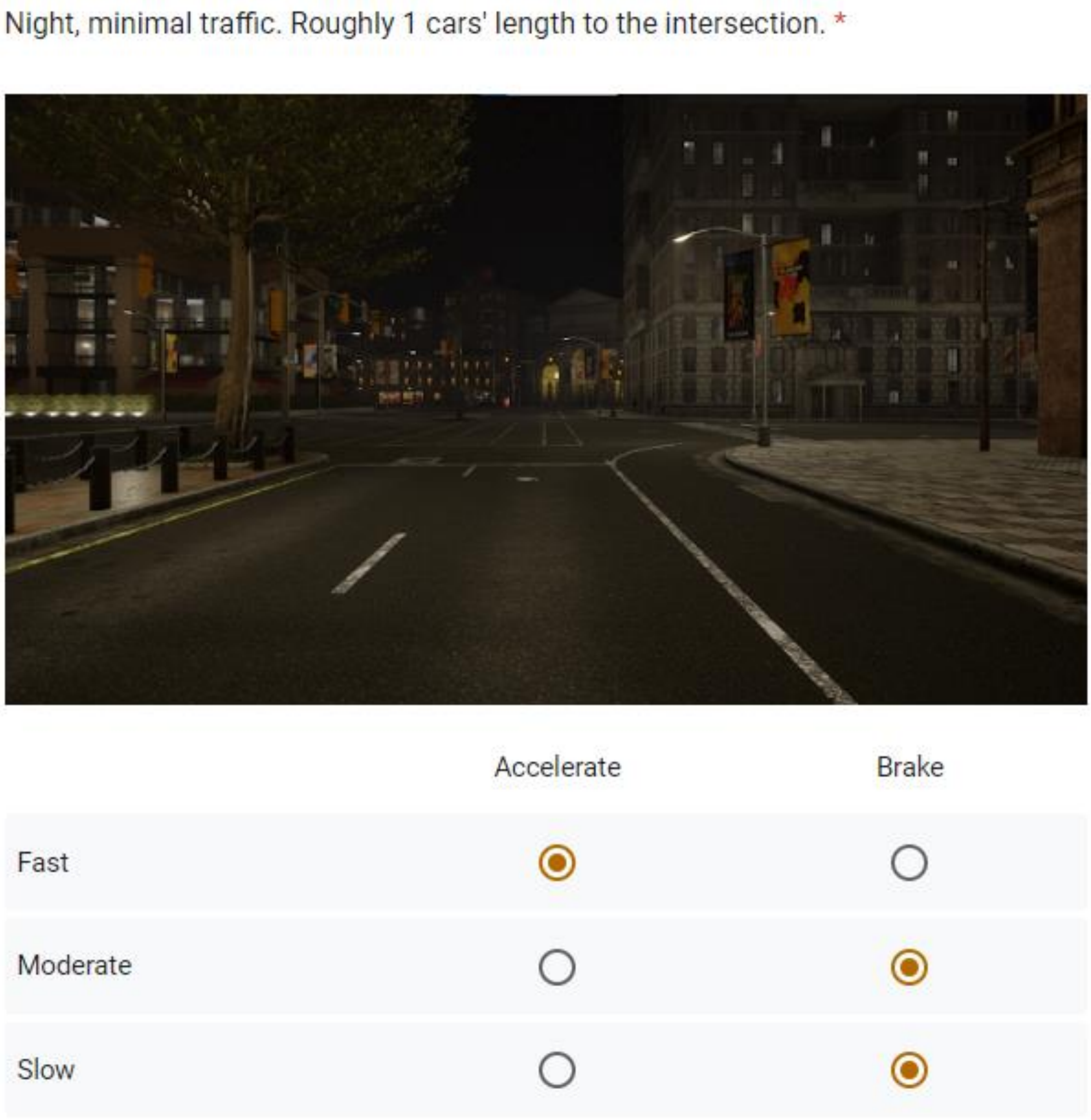
Sunny, high traffic. Roughly 0.5 cars' length to the intersection. The cars ahead of you are running the yellow light. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



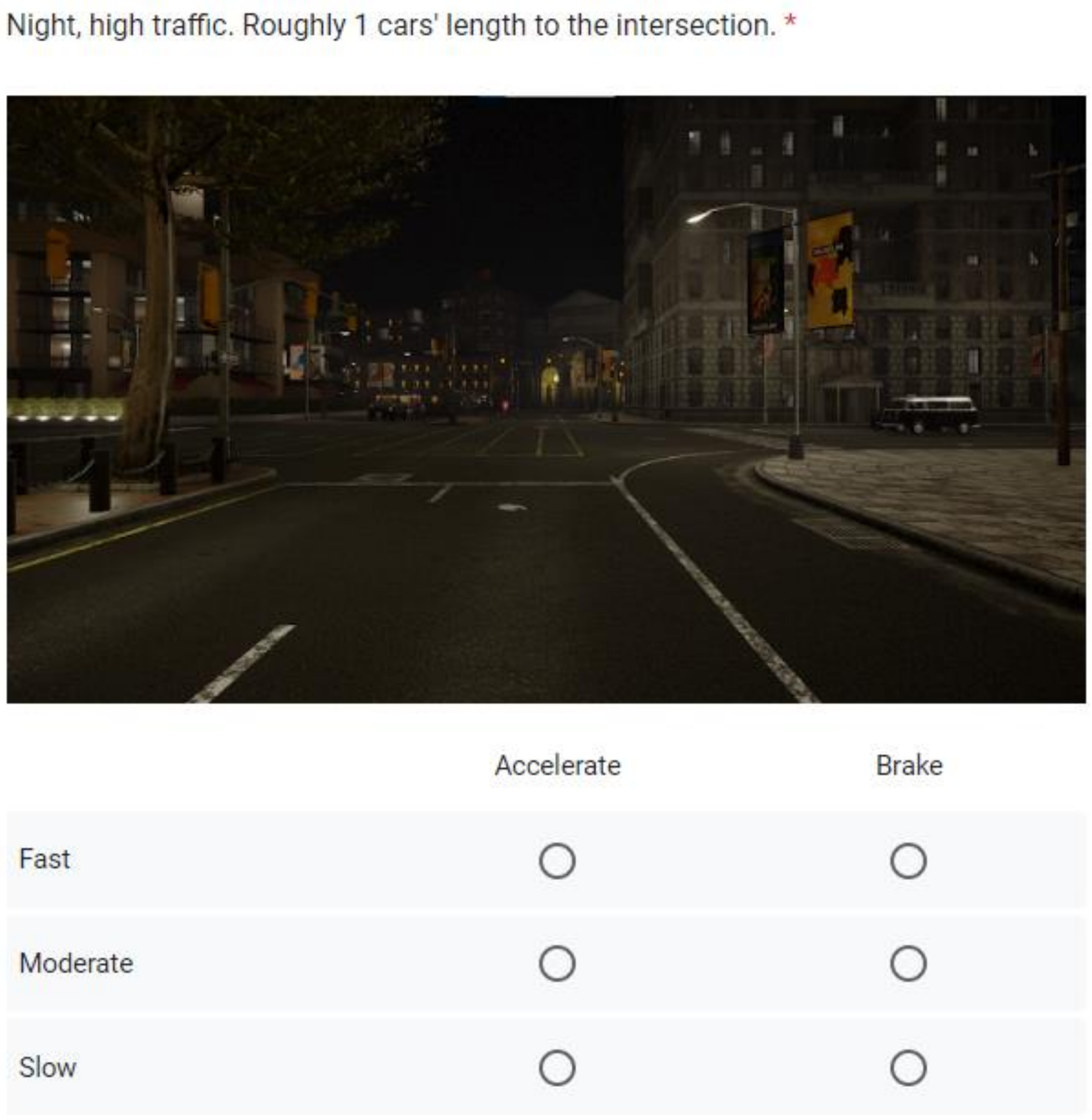
Sunny, high traffic. Roughly 2 cars' length to the intersection. The cars ahead of you are running the yellow light. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



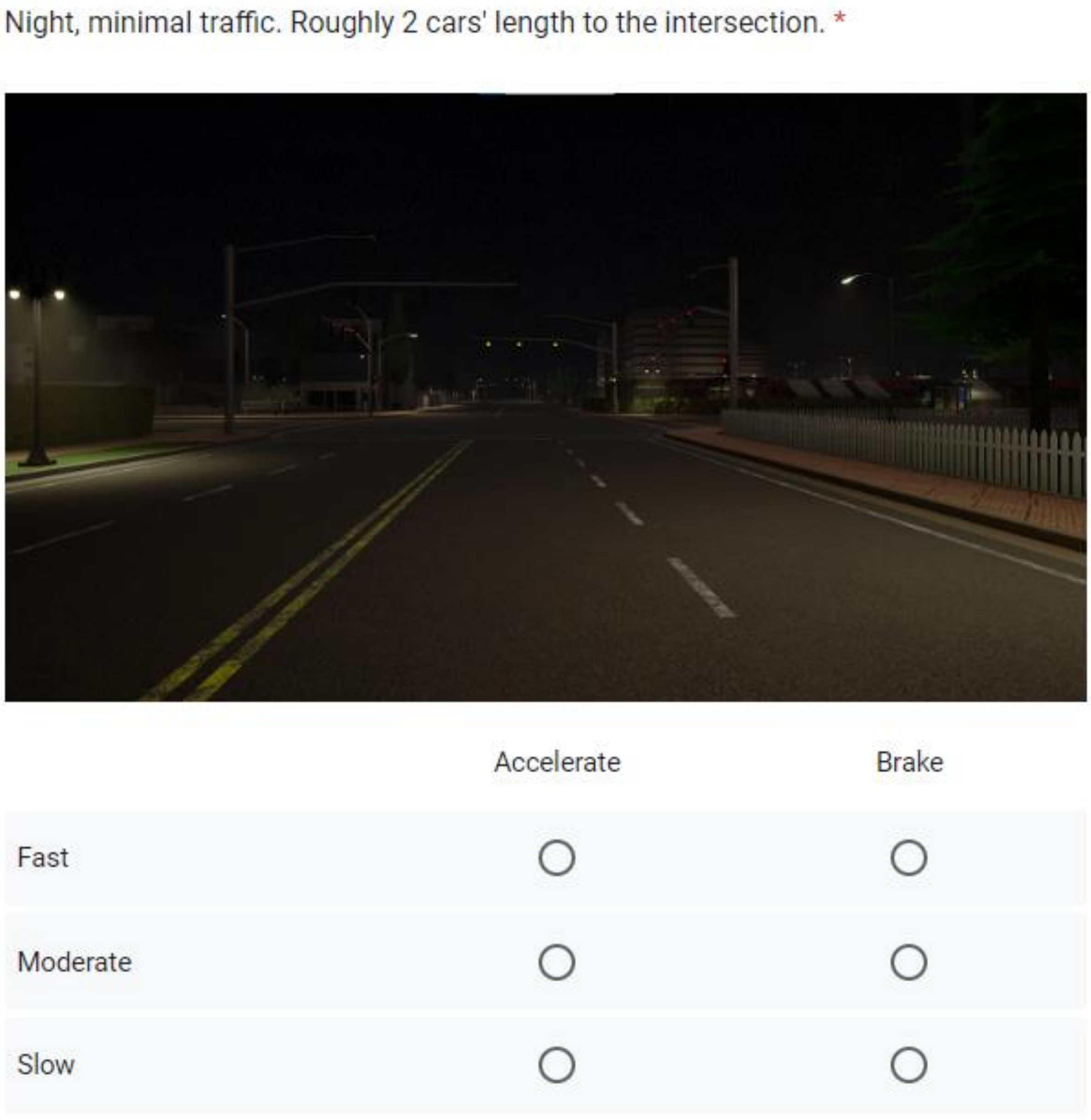
Night, minimal traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	●	○
Moderate	○	●
Slow	○	●



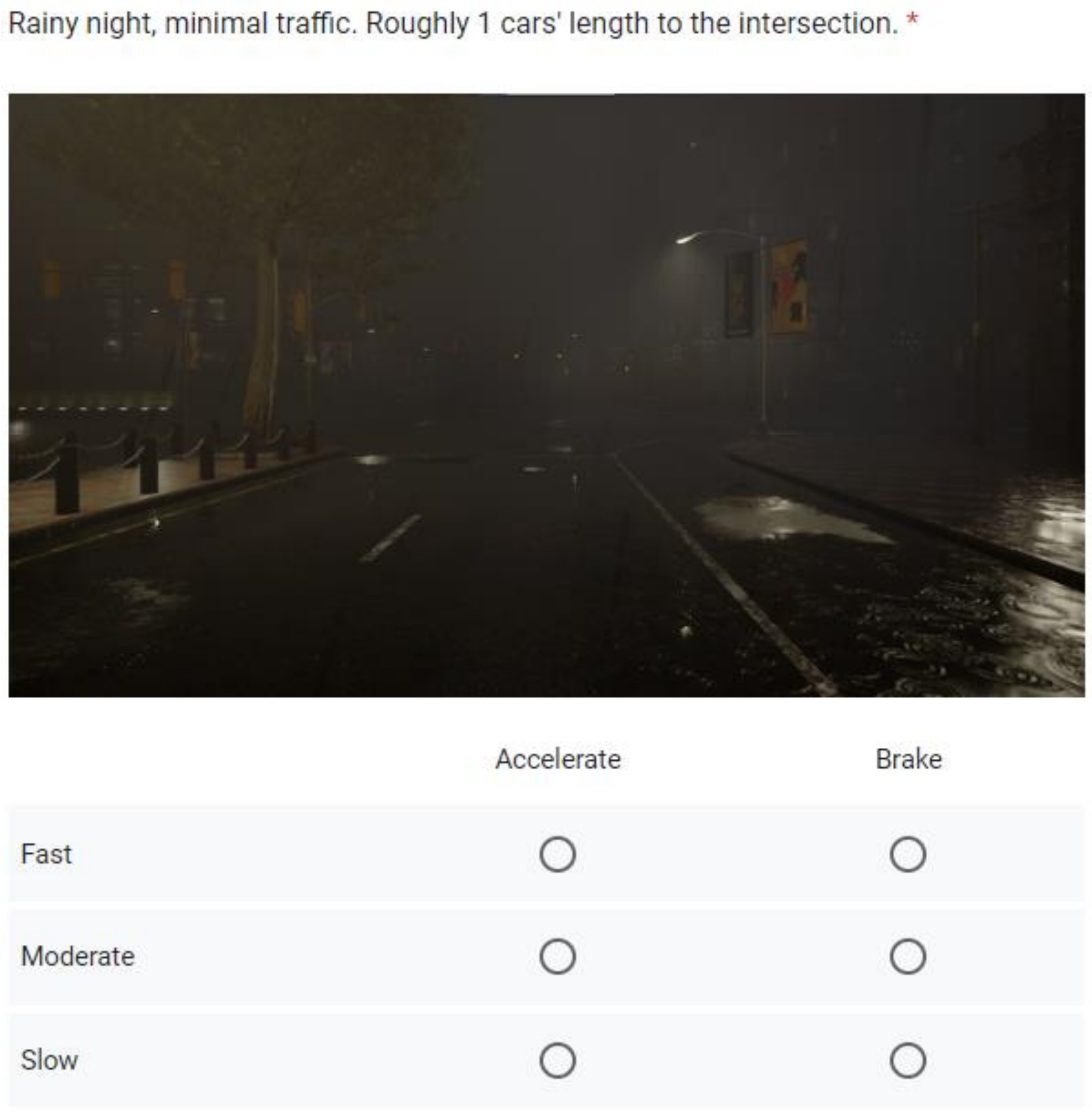
Night, high traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



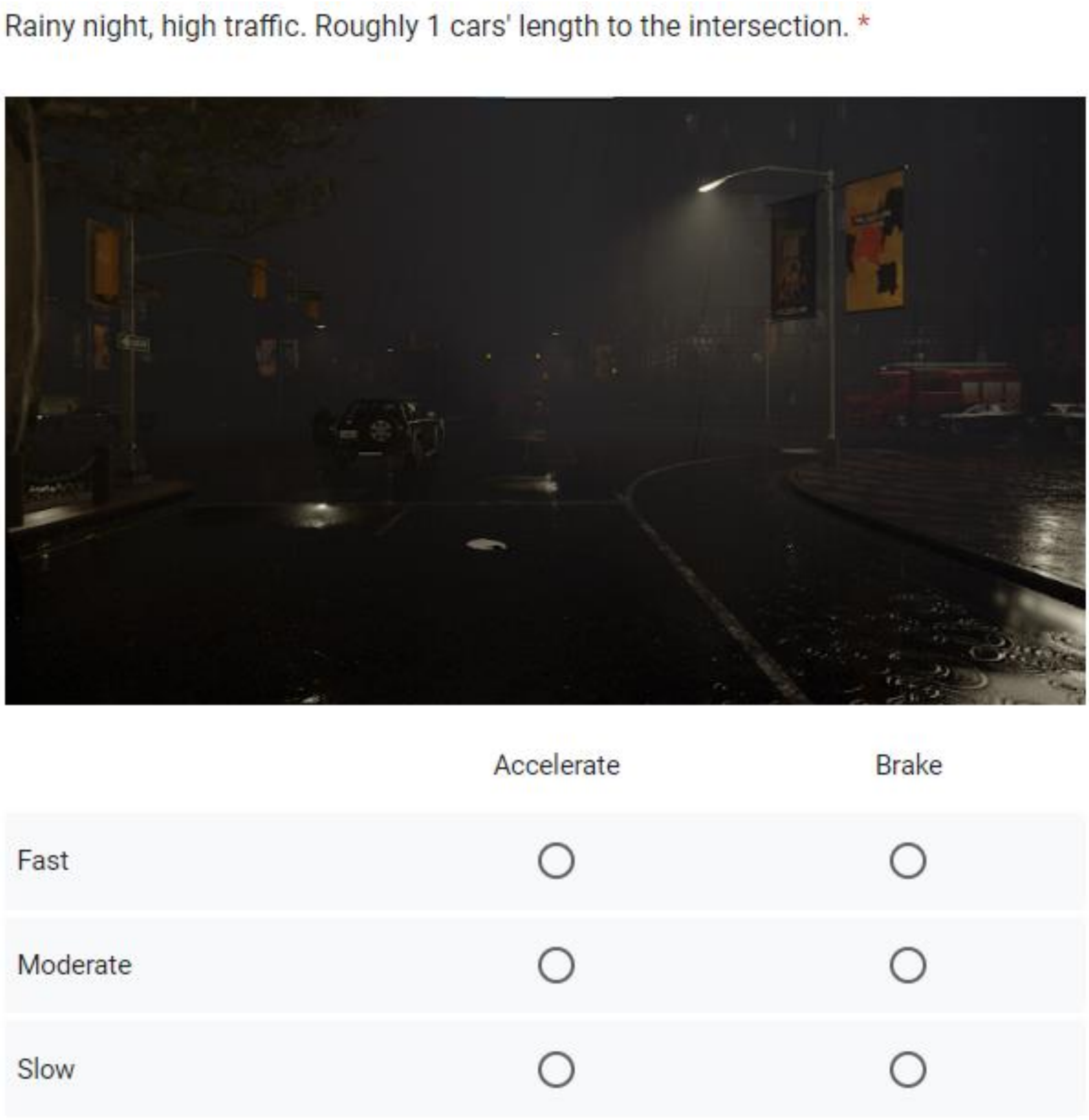
Night, minimal traffic. Roughly 2 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



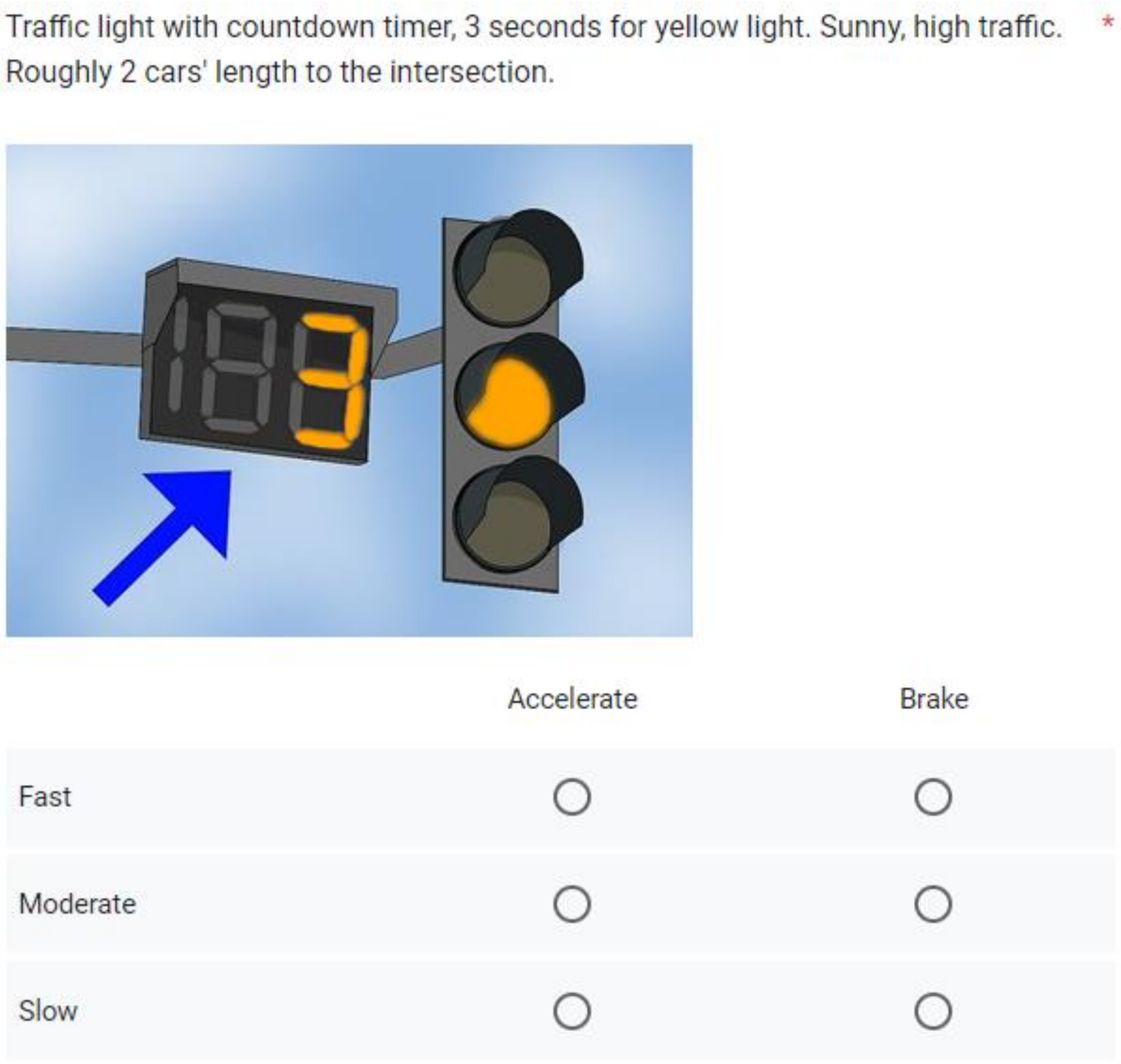
Rainy night, minimal traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



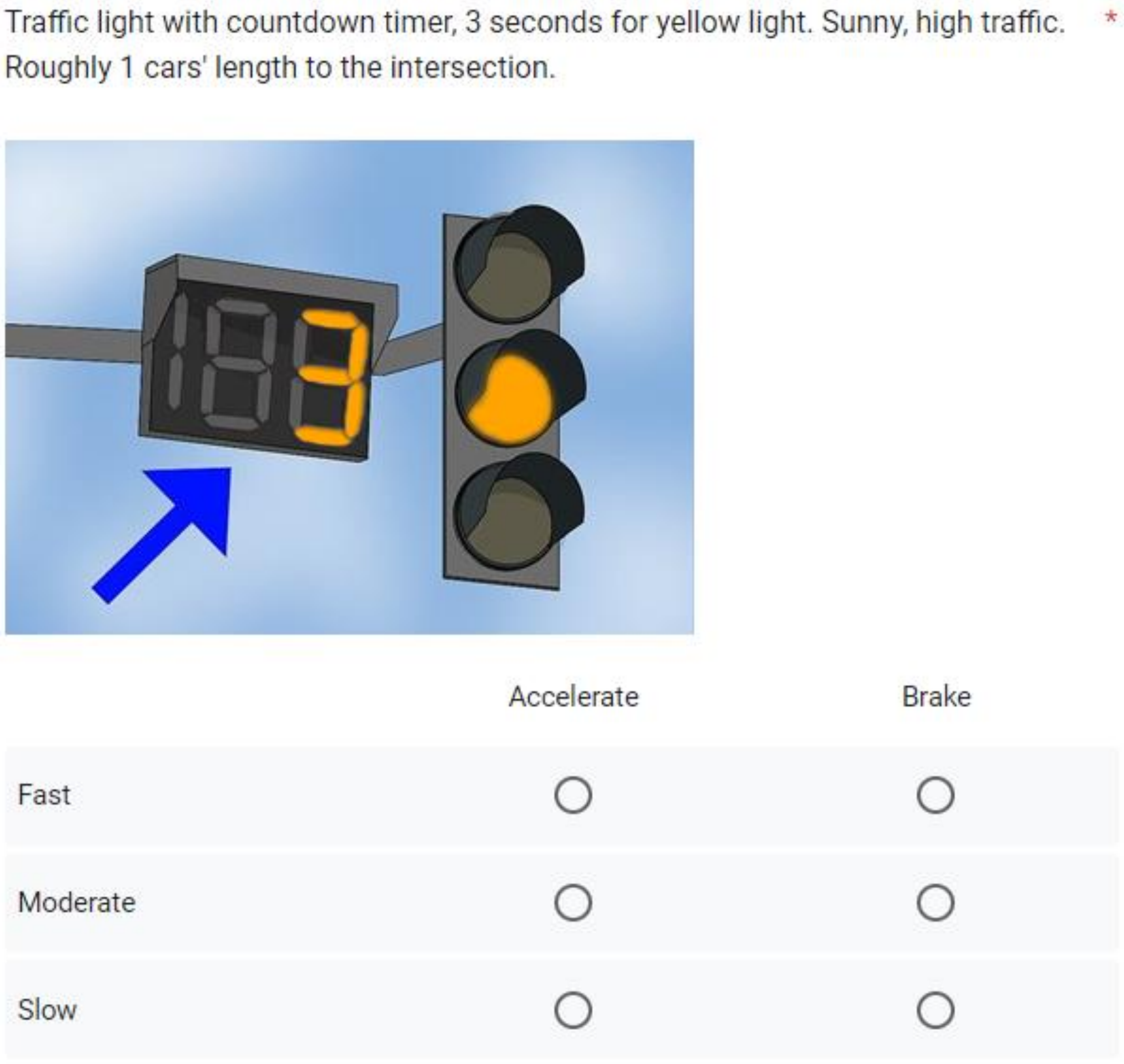
Rainy night, high traffic. Roughly 1 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



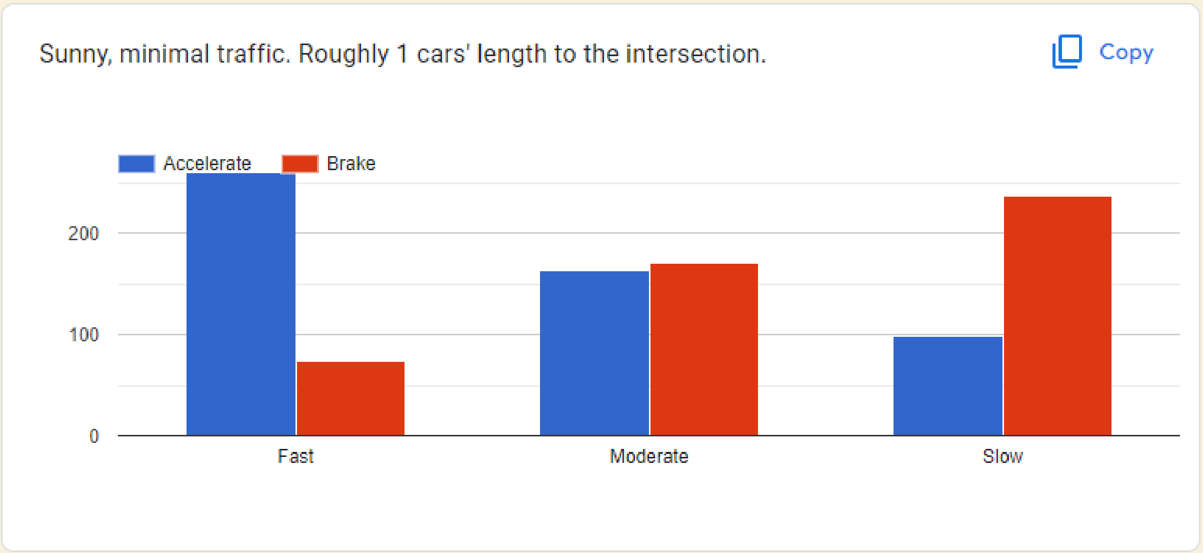
Traffic light with countdown timer, 3 seconds for yellow light. Sunny, high traffic. Roughly 2 cars' length to the intersection. *

	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○

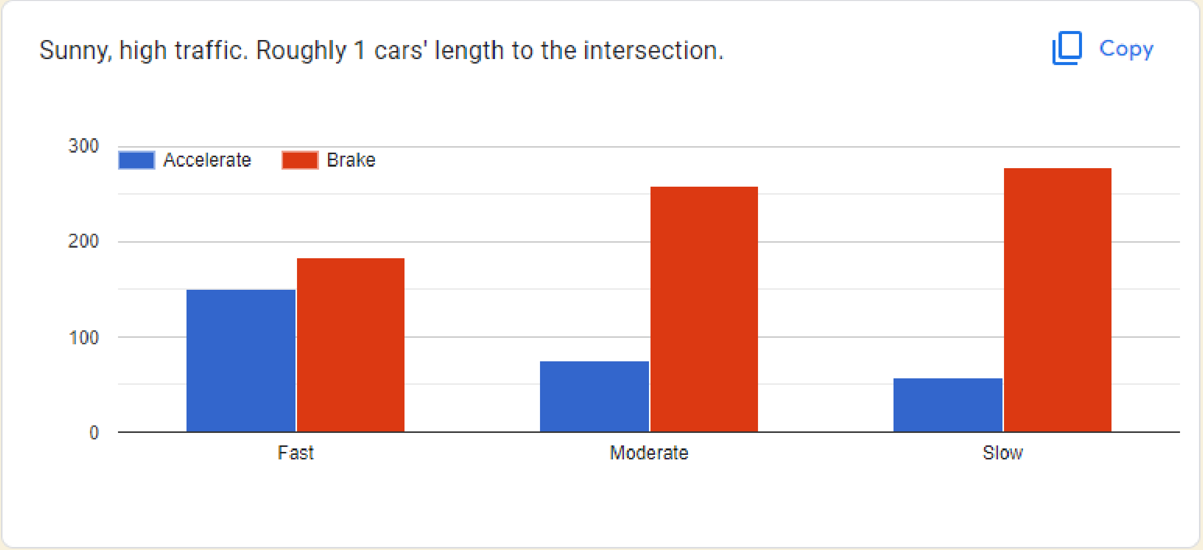


Traffic light with countdown timer, 3 seconds for yellow light. Sunny, high traffic. *
Roughly 1 cars' length to the intersection.

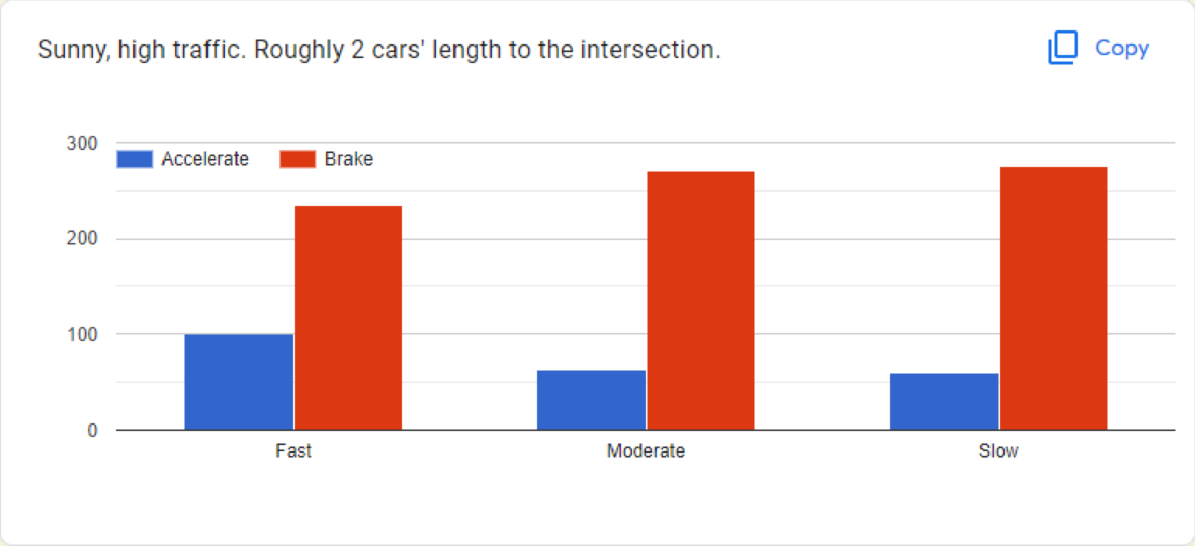
	Accelerate	Brake
Fast	○	○
Moderate	○	○
Slow	○	○



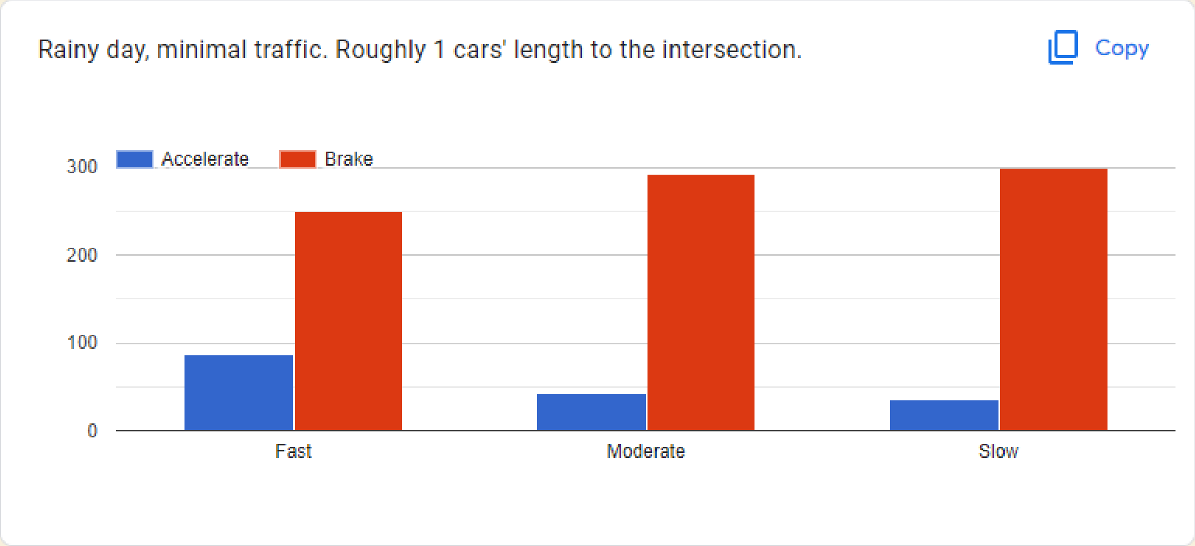
Sunny, minimal traffic. Roughly 1 cars' length to the intersection.



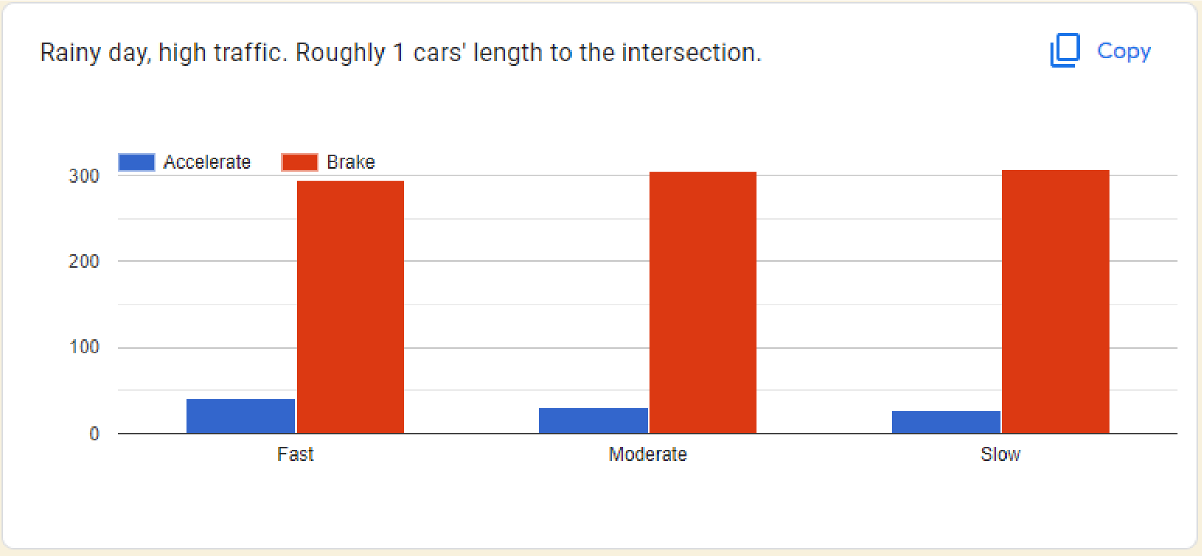
Sunny, high traffic. Roughly 1 cars' length to the intersection.



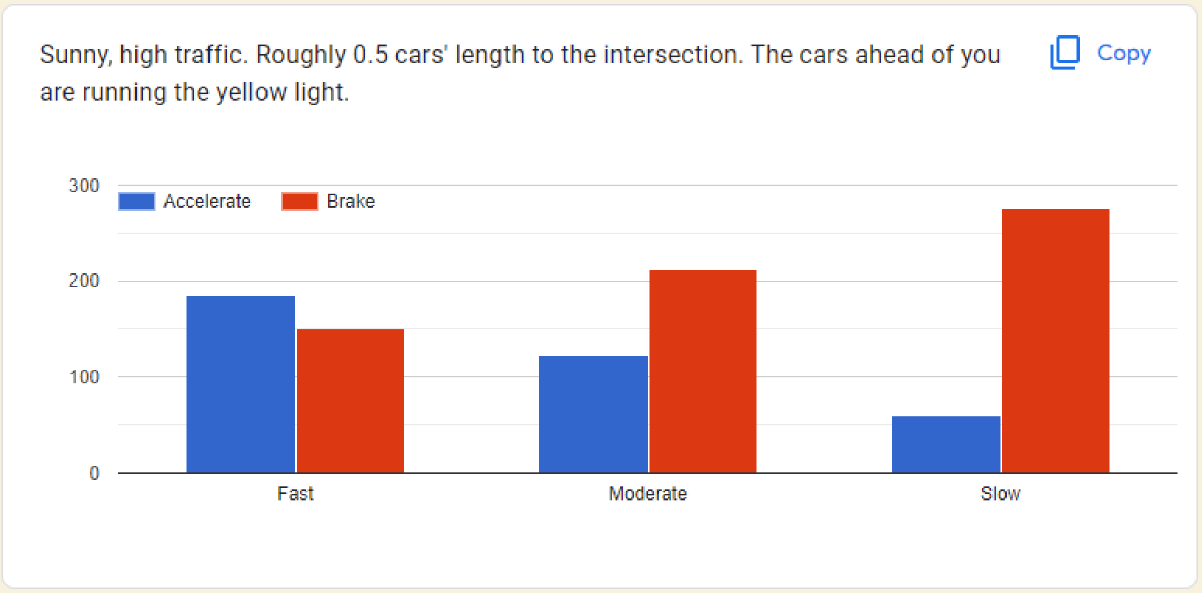
Sunny, high traffic. Roughly 2 cars' length to the intersection.



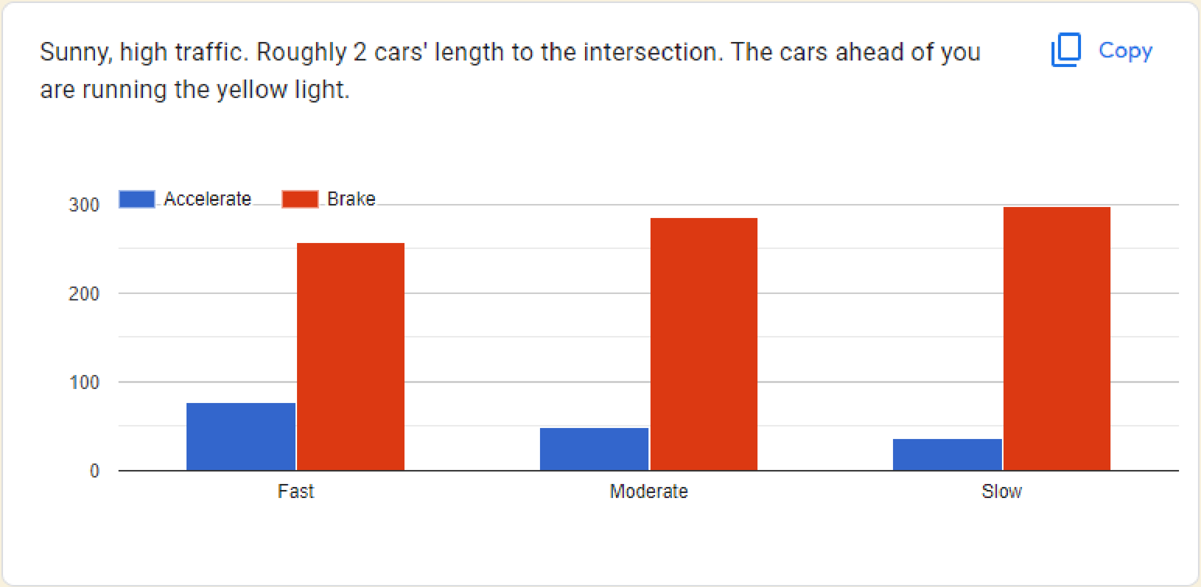
Rainy day, minimal traffic. Roughly 1 cars' length to the intersection.



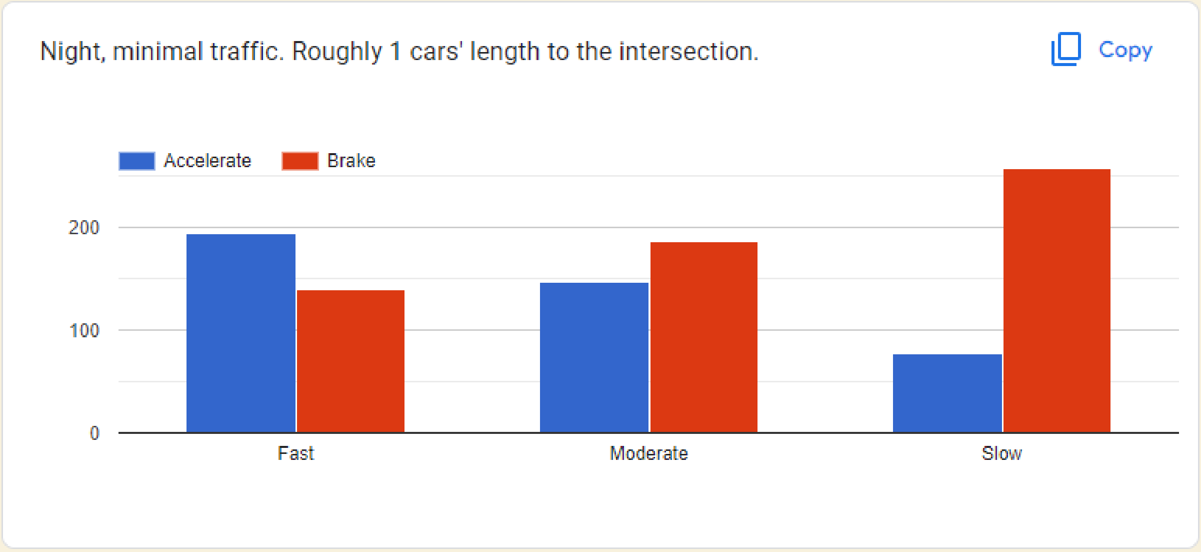
Rainy day, high traffic. Roughly 1 cars' length to the intersection.



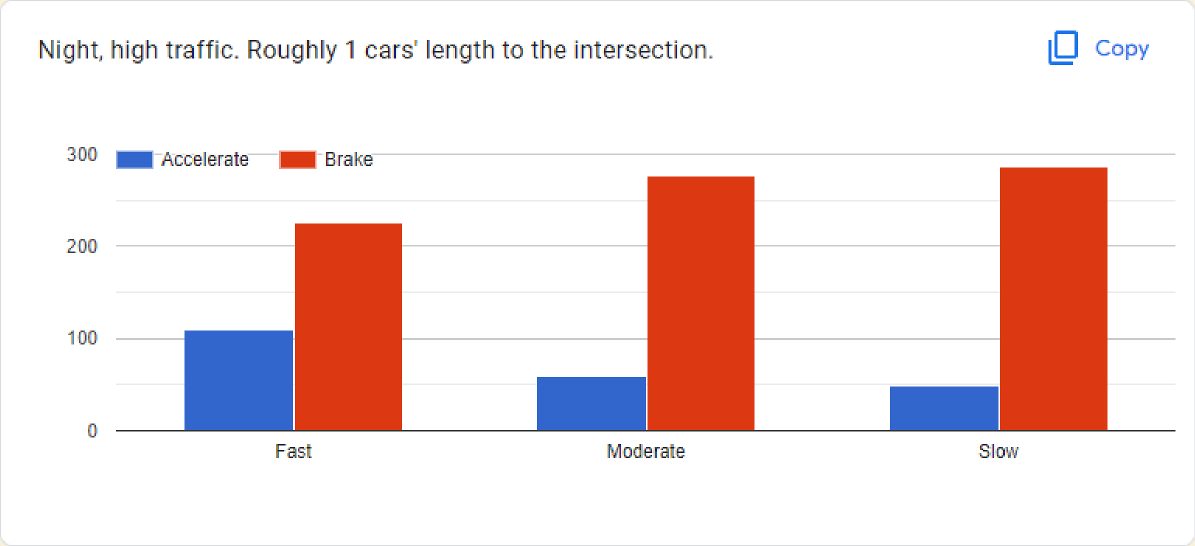
Sunny, high traffic. Roughly 0.5 cars' length to the intersection. The cars ahead of you are running the yellow light.



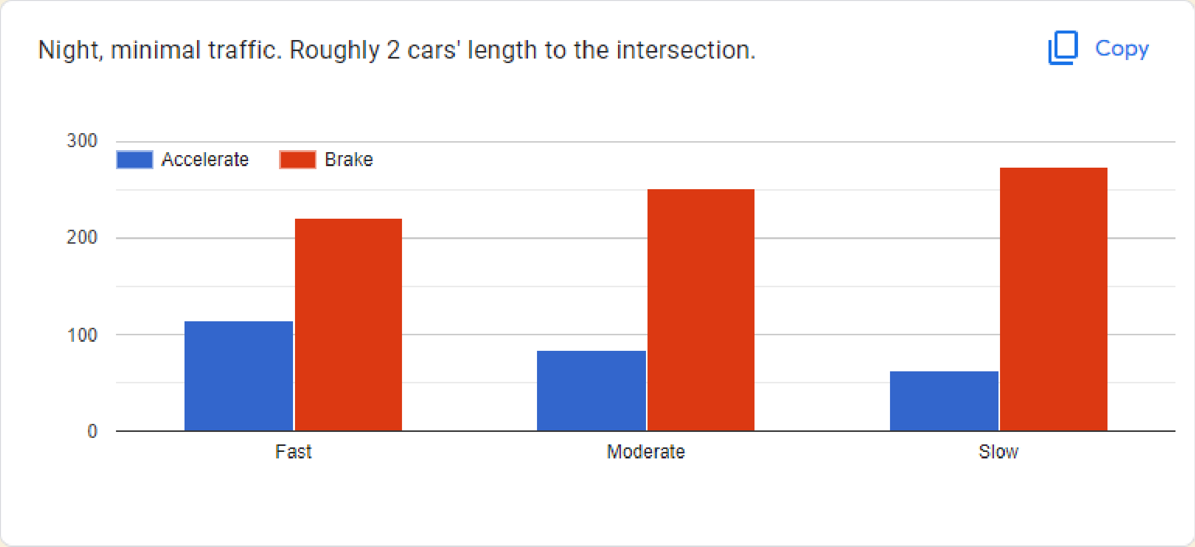
Sunny, high traffic. Roughly 2 cars' length to the intersection. The cars ahead of you are running the yellow light.



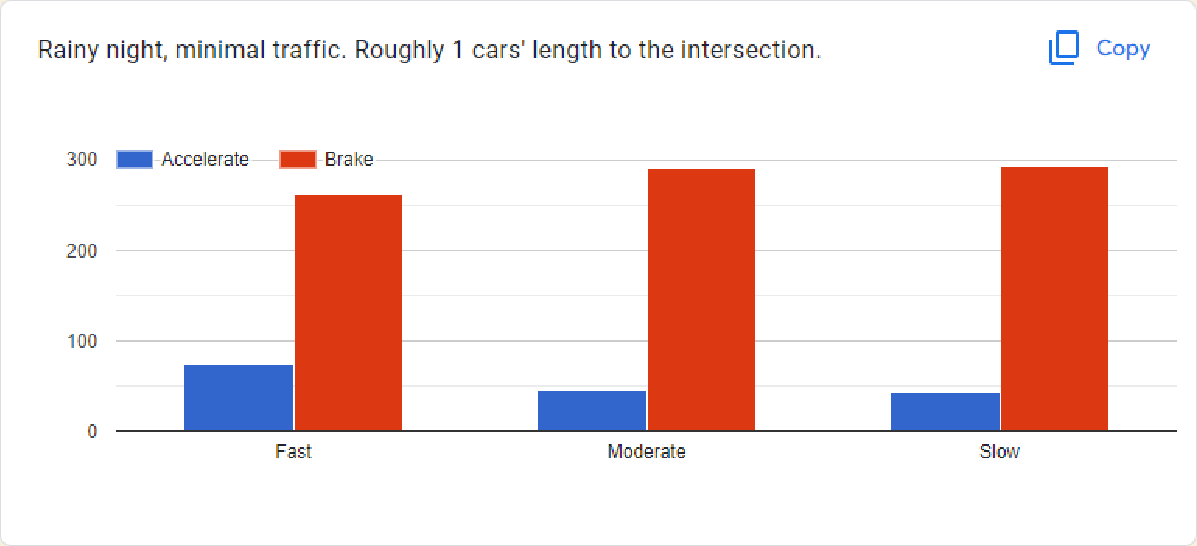
Night, minimal traffic. Roughly 1 cars' length to the intersection.



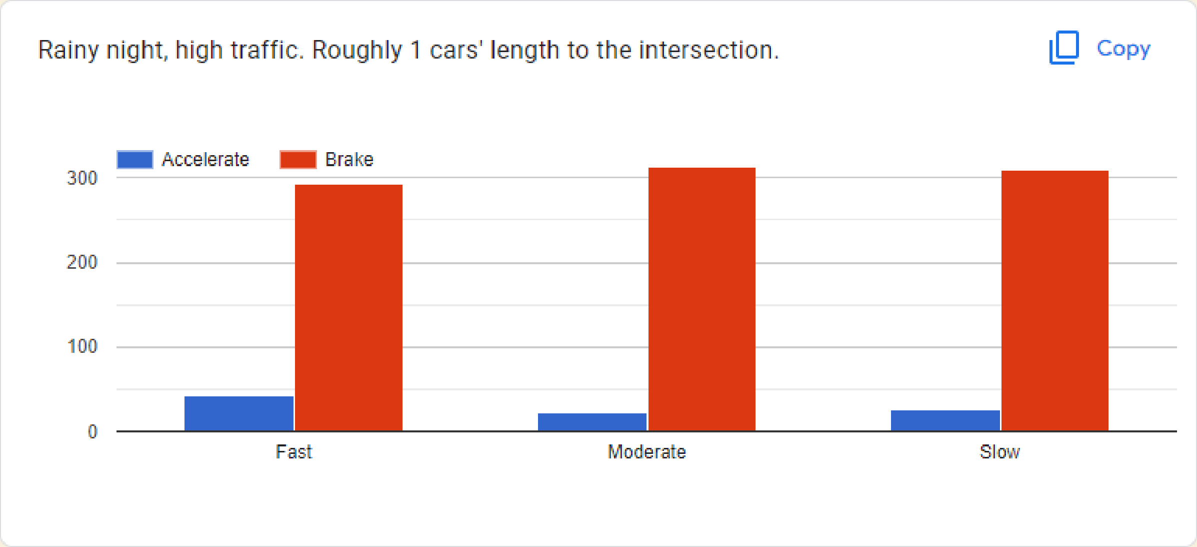
Night, high traffic. Roughly 1 cars' length to the intersection.



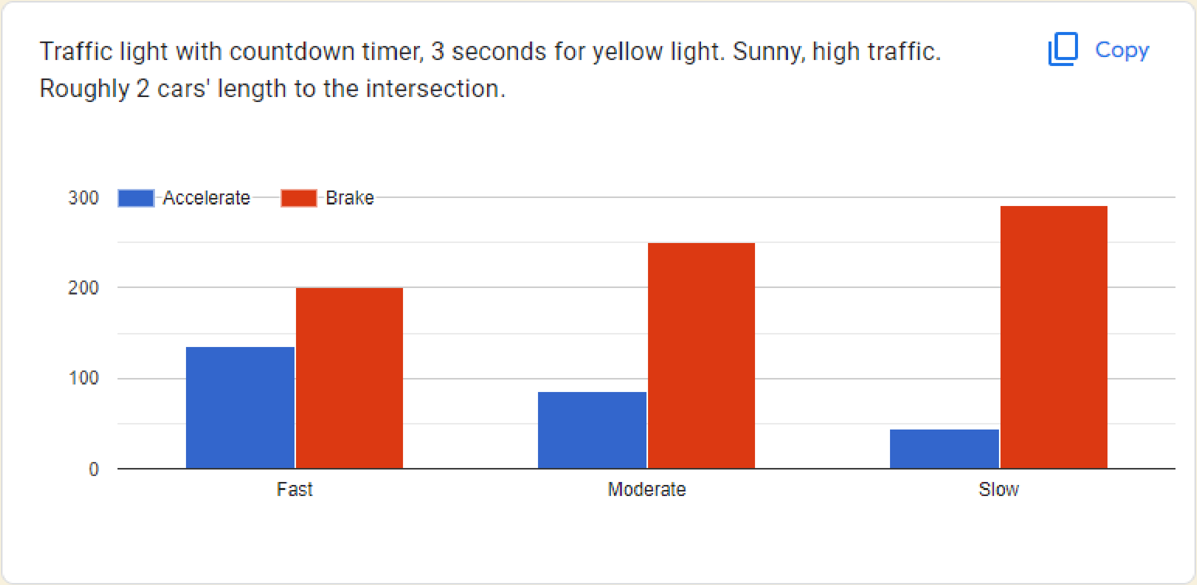
Night, minimal traffic. Roughly 2 cars' length to the intersection.



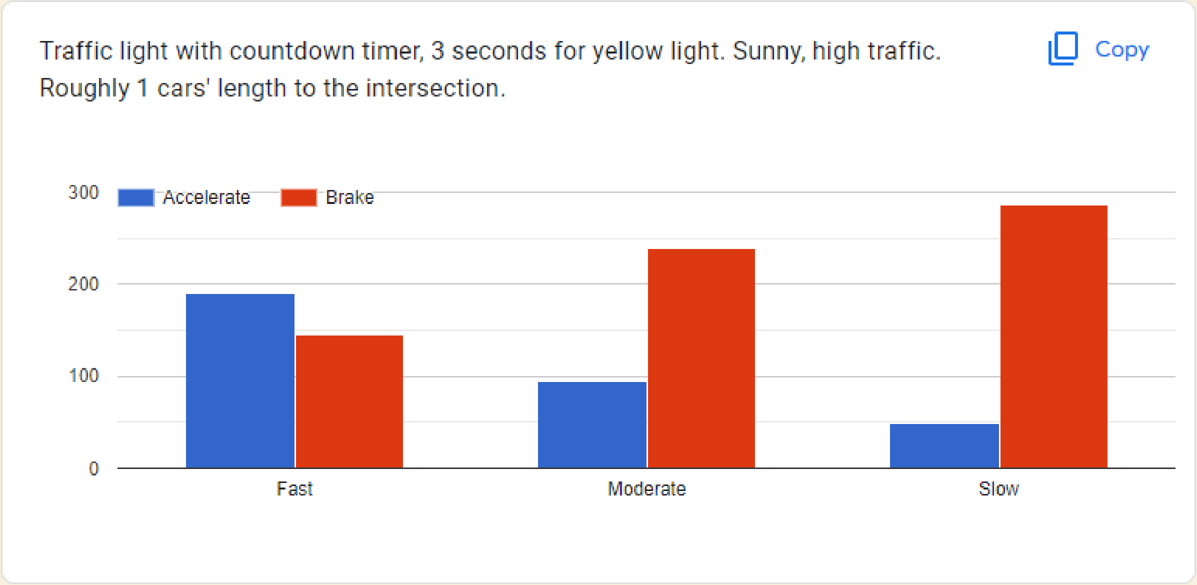
Rainy night, minimal traffic. Roughly 1 cars' length to the intersection.



Rainy night, high traffic. Roughly 1 cars' length to the intersection.



Traffic light with countdown timer, 3 seconds for yellow light. Sunny, high traffic. Roughly 2 cars' length to the intersection.



Traffic light with countdown timer, 3 seconds for yellow light. Sunny, high traffic. Roughly 1 cars' length to the intersection.