

Biostatistics Collaboration of Australia
University of Sydney

Workplace Project Portfolio

WWPA: Factor Analysis of Young Driver Perceptions, Behaviours and Driving
Experience

WWPB: Predictors of Time to First Crash for Young Drivers: Survival Analysis

Student Name: David Warr

Student Number: 200338700

Table of Contents

Table of Contents	2
Preface	4
Introduction.....	4
Student's Contribution	4
Reflections on the Learning Process	5
Ethical Considerations.....	6
Project A	7
Front Sheet: Project A	7
<i>Project Title</i>	7
<i>Location and Dates</i>	7
<i>Context</i>	7
<i>Student Contribution</i>	7
<i>Statistical Issues</i>	7
<i>Student Declaration</i>	8
<i>Supervisor Declaration</i>	8
Project Description	9
<i>Background</i>	9
<i>Aim of Project</i>	9
Overview of Study Design	12
<i>Data Management</i>	12
<i>Modeling Approach</i>	16
Results	24
Discussion	33
Conclusion.....	36
Project B	38
Front Sheet: Project B	38
<i>Project Title</i>	38
<i>Location and Dates</i>	38
<i>Context</i>	38
<i>Student Contribution</i>	38
<i>Statistical Issues</i>	39

<i>Student Declaration</i>	39
<i>Supervisor Declaration</i>	39
Project Description	40
<i>Background</i>	40
<i>Aims of Project</i>	40
Overview of Study Design	42
<i>Data Management</i>	42
<i>Modeling Approach</i>	48
Results	52
<i>Univariate Analysis</i>	52
<i>Multivariate Analysis</i>	56
Discussion	64
Conclusion.....	69
References	70
Appendix	72
<i>SAS Code for Project A: Factor Analysis</i>	72
<i>STATA Code for Project B: Survival Analysis</i>	73

Preface

Introduction

The two projects that form part of this portfolio were undertaken during the period September 2007 and November 2008. Both projects involved analysing data collected as part of the online prospective cohort DRIVE Study undertaken in NSW during 2003 and 2004 by members of The George Institute. Both projects were to form part of the wider DRIVE Study that may ultimately affect the licensing process and training applied to young drivers in Australia to lower the incidence of injuries and death related to vehicle crashes for young drivers.

The first project was a factor analysis on well known scales and including additional variables to assess whether the data in this dataset was comparable to these scales or was indicating factor loadings of a different nature. This would assist in understanding the behaviours and the inter-relationships between these behaviours of young drivers, which may ultimately be used to assess what may potentially cause increased crashes in young drivers.

The second project was also related to the DRIVE dataset and specifically related to the driver crashes. This sub-study was slightly different to the majority of studies undertaken on young drivers in that the focus was on the time until first crash rather than simply the factors associated with a crash. The general implication is that if an individual can reduce their likelihood of crashing early they may increase their driving ability to subsequently avoid crashes in the future.

Student's Contribution

The scope of both projects involved the evaluation of the collected data and ensuring the format of this data was suitable for the two planned statistical analyses.

There were no stipulated timelines for completion of either project outside submission dates for the WPP. However there was regular interaction with both

the project supervisor and statistical supervisor throughout the duration of both projects to ensure the direction of the projects remained appropriate and the statistical process was valid. A large portion of the communication centred on emailed questions and results and the occasional face to face meeting.

In both projects the SAS and Stata code to perform the statistical analyses was developed by me apart from the construction of a number of predefined measures (eg. Kessler 10 index, Socio-Economic Indexes for Area) that were supplied in the original dataset. The major component was the analysis and interpretation of results to ensure the output was relevant.

Reflections on the Learning Process

Undertaking these two projects through The George Institute allowed me to gain a better understanding of the issues and processes used in a professional setting on design and analysis of research questions.

Undertaking these two projects reinforced the notion that the actual time spent on statistical analysis is rather small relative to the time required in data management, data cleaning and more importantly interpreting and ensuring the results are ultimately logical. Unlike the datasets generally used during the BCA course that have been materially reduced to address a particular concept, the analyses undertaken from this dataset raised a number of questions on what the data was actually indicating. This resulted in numerous discussions on the output and how relevant the results were irrespective of whether they showed statistical significance.

It also became clear that defining the questions that require addressing and obtaining a very clear scope of what is required is fundamental to ensuring the project stays focused. A number of times it was easy to see how the focus could move away from the original question to be addressed once some of the results were reviewed.

One of the projects involved undertaking a statistical process that I had not previously learned in the Masters course. The factor analysis project highlighted the need to undertake sufficient research on the approach to ensure I had a reasonable understanding of the issues and on previous studies using this approach to obtain an understanding of how this information is regularly reported. There was a clear understanding that the more research undertaken at the commencement of a project on the methodologies and prior findings greatly assists in the clearer understanding of the question to be addressed.

Ethical Considerations

The projects involved data collected on young drivers that contained both personal and sensitive information. Most of the personal information had been removed from the dataset I received, although with drivers licence numbers included as a unique identifier, there remained a need to ensure this information was not released to the general public. As such confidentiality agreements were signed to ensure the specific data and results of the studies were not released in any way that would identify specific individuals.

At the time of writing, the results of the two projects have not been presented to a wider audience, although the expectation is for both to be published at a later date.

Project A

Front Sheet: Project A

Project Title

Factor Analysis of Young Driver Perceptions, Behaviours and Driving Experience

Location and Dates

School of Public Health, University of Sydney

September 2007 – November 2008

Context

The George Institute is undertaking a number of studies based on the DRIVE dataset to ultimately assist in a greater understanding of the behaviour of young drivers and their propensity to crash. This factor analysis was indicated to be of interest to The George Institute to obtain a better understanding of some of the constructed variables and to determine if particular factor structures are appearing in the data that may be different or additional to those used in previous studies. Identifying themes can also simplify and streamline educational and other intervention strategies to the most parsimonious set of messages/targets.

Student Contribution

There was regular interaction with the project supervisor, along with a number of her colleagues, throughout the duration of the project to discuss data issues, methodological approaches and interpretation of results. I carried out all the analysis after significant research into the appropriate methodology to perform a factor analysis and the interpretation of the findings.

Statistical Issues

The statistical issues encountered in this project were the understanding, proficient use and interpretation of the factor analysis method. This statistical approach had not been covered in the BCA program, and the method was new to me. Additional

to the factor analysis approach, an understanding of principal component analysis was required along with Cronbach's reliability coefficient.

Student Declaration

I declare that this project is my own work, with guidance provided by my project supervisor, Alexandra Martiniuk, and is not the result of collaboration with others and that I have not previously submitted it for academic credit.



David Warr

19 May 2009

Supervisor Declaration

I declare that David Warr has worked alone on this project under our supervision. This work has not previously been submitted for publication or academic credit.



Dr Alexandra Martiniuk, senior research fellow

19 May 2009

Project Description

Background

As a result of the reported high proportion of deaths in the 17 to 24 year age group from vehicle related crashes, the DRIVE¹ Study was initiated to investigate the effect of numerous factors on the risk of vehicle crashes and related injuries on drivers between the ages of 17 to 24 who have recently obtained their driver's licence.

The focus of the present study was the inter-relationships between factors that have been considered to contribute to a 17 to 24 year old driver's potential to crash. While numerous studies have attempted to determine which observable factors determine crashes while driving, few have considered the inter-relationships between these variables and few, if any using a prospective cohort design with such a large study subject number as the DRIVE Study.

Of those that have considered inter-relationships, one of these² centred on a factor structure for young drivers who had crashed and determined four separate factors from the data, relating to fate, environment, self and other reasons. Another³ undertook a path analysis on personality, attitudes and risk perception of young drivers and concluded that risky driving was mediated mostly through attitudes. This sub-study will contribute to further understanding of these relationships.

Aim of Project

Factor analysis was used to study the patterns of relationship among many explanatory variables, with the goal of discovering something about the nature of the predictor variables that affect them; even though those predictor variables were not measured directly.

In factor analysis, predictor (or latent) variables are composed of weightings of the observed variables and the latent variables obtained by factor analysis are necessarily more hypothetical and tentative than is true when predictor variables

are observed directly. The inferred predictor variables are called *factors*. A typical factor analysis suggests answers to four major questions:

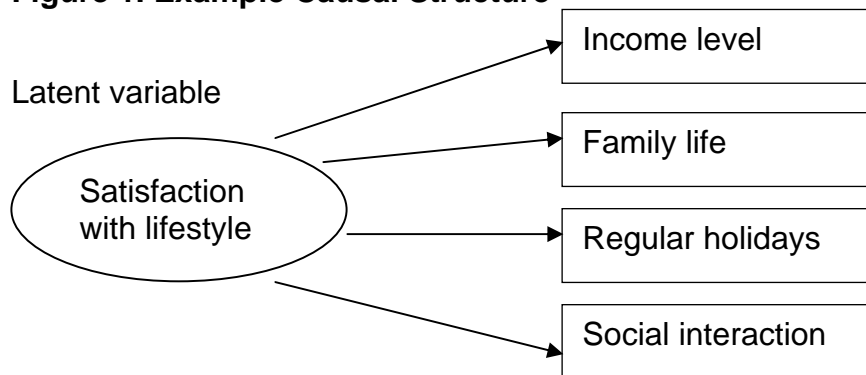
1. How many different factors are needed to explain the pattern of relationships among these variables?
2. What is the nature or underlying feature of each factor?
3. How well do the hypothesized factors explain the observed data?
4. How much purely random or unique variance does each observed variable include?

The nature of latent factors is such that they are unlikely to be easily measured by one observable variable. Therefore many measurable variables are required to adequately address the possible nuances or issues related to the particular latent variable.^{4,5}

Conceptual Approach

The general concept of the factor analysis approach is to find simple patterns in the relationship between observed variables to determine the number and nature of a smaller set of latent factors that accounts for the covariance in the dataset⁵. The factors in a factor analysis account for the common variance in the dataset rather than the total variance utilised in a principal component analysis. The unique variance, which is the difference between the total variance and common variance, is left unanalysed. The common variance is the proportion of total variance that is accounted for by the factors. A factor analysis assumes that the covariance in an observed variable is the result of causal influence from one or more latent variables. For example, the causal structure observed from the following structure shows that there are multiple influences on the satisfaction of an individual's lifestyle:

Figure 1. Example Causal Structure



The latent variable “satisfaction with lifestyle” is difficult to measure directly therefore observable variables are utilised, such as the number of holidays taken, or the income level of the individual, which are elements of the latent variable in this example. In factor analysis the latent variable is a notional variable, deemed to exist and exert influence on the observable responses.

Aim of Factor Analysis in the DRIVE Study

This sub-study presents an assessment of the latent variables determined from the output of a number of survey questions from the DRIVE dataset, to better understand how these may influence driver perceptions, attitudes, behaviours and knowledge. Specifically, the analysis was an examination of the explanatory and latent variables from a number of pre-determined questions used in known psychological scales, as well as additional variables that may also load onto these or separate factors.

The overriding aims of the factor analysis of the DRIVE Study variables were:

1. To undertake an exploratory factor analysis to determine the nature and number of underlying factors that are responsible for variation in the DRIVE dataset. The variables within the dataset used in the factor analysis related to:
 - i. Demographics (eg. age, gender, country of birth)
 - ii. Socio-economic status
 - iii. Risky driving behaviour
 - iv. Risk perception
 - v. Driver training
 - vi. Pre-licence driving exposure

- vii. Mental health
- viii. Alcohol and drug use
- ix. Sensation seeking behaviour

The analysis aimed to determine whether the factors underlying the DRIVE Study data specifically relate to the grouping listed above or relate to different latent variables and to consider the inter-relationships between these latent variables.

2. To assess the reliability of the observed variables that form each of the latent factors.

Overview of Study Design

Data Management

Original Data

Self reported risk factors on 20,822 young drivers between the ages of 17 and 24 were collected via the on-line DRIVE Cohort Study¹ during 2003 and 2004. Hypothesized risk factors considered to contribute to a young driver's propensity to crash have been formulated into a series of self report questions relating to the individual's opinions, demographics, mental health, socio-economic status, life experience, risk perception, sensation seeking behaviour, driver training and other driving experience.

A SAS dataset was received with all the original responses (baseline survey) from the participants together with outcomes collected on the participants based on police reported crashes. Specifically the data related to the date of each crash, and the number of people injured in direct relation to each crash. Internationally recognised scales were constructed from the outcomes of the survey responses. (eg. Kessler 10 psychological scale, Alcohol Use Disorders Identification Test (AUDIT-C scale for alcohol consumption) These were provided by members of The George Institute.

The variable categories used in the initial model were: age, gender, country of birth, Socio-Economic Indexes for Area (SEIFA), risk taking behaviour, risk perception, Zuckerman sensation seeking score²⁷, professional and non-professional driver training, Kessler 10 mental health index, AUDIT-C scale for alcohol use, driven on L-plates without supervision, driven a motorcycle, marijuana usage, and self harm.

Determining the number and nature of latent factors is somewhat subjective and ultimately relies upon whether the observed variables that load onto a latent factor can be adequately labeled as forming one logical group; as in the “satisfaction with lifestyle” factor in the above example. The nature of the observed variables considered in each factor analysis are initially derived from existing theories or evidenced from prior literature. However, in an exploratory analysis this does not preclude introducing additional variables not previously included in existing literature that the researcher considers plausible.

Data Manipulation

Gender

Gender remained in its original form (male / female).

Age

Based on age at the date each individual obtained their provisional licence. The range of ages was between 17 and 24 years old.

Driver Training and Experience

Based on the self reported number of hours of driver training, this was reported separately, based on the total hours spent under professional supervision (ie. licenced driving instructor) and total hours learning to drive with a non-professional supervisor (eg. parent, relative, or friend). Due to the range of self reported driving hours, these two variables were then re-categorised from continuous into discrete using the following categories:

Professional supervision		Non-professional supervision	
Continuous	Discrete	Continuous	Discrete
0 hours	Category 0	0 to 39 hours	Category 0
>0 to 4 hours	Category 1	40 to 49 hours	Category 1

>4 to 8 hours	Category 2	50 to 59 hours	Category 2
> 8 hours	Category 3	60 to 69 hours	Category 3
		> 69 hours	Category 4

A separate 5 point Likert scale (never to always) was used to record the amount of time spent driving on open, major and residential roads under both professional and non-professional driver instruction.

The length of time driven on unsealed road, during rain, at night and in heavy traffic was recorded. Due to the skewed responses these observations were re-categorised into equal quartiles.

The number of hours driving a motor vehicle prior to obtaining a learner's permit was also collected and categorised as a dichotomous variable for either never driving prior to obtaining a learner's permit or having driven at least once prior to obtaining the permit.

Supervision on L Plate

This variable identified whether an individual drove a vehicle while on their learner's permit without supervision. This variable was categorised as 0 if all driving was undertaken with supervision and 1 if some driving while on a learner's permit was undertaken without supervision.

Time on L Plate

Time on learner's permit is a continuous variable based on the number of years the learner's permit was held. This variable was also re-categorised into an ordinal variable as follows:

Continuous	Discrete
< 1 year	Category 1
1 to 1.5 years	Category 2
> 1.5 years	Category 2

Remoteness of Residence

The remoteness of residence variable was considered for inclusion however as it is not a ratio or interval based variable it was excluded from the initial analysis.

Country of Birth

The country of birth variable was considered for inclusion however as it is not a ratio or interval based variable it was excluded from the initial analysis.

AUDIT-C¹⁷

AUDIT-C is an industry standard screen used to help identify patients that may have a dependence on alcohol and potentially to assess whether an individual is at risk of an alcohol related disorder. The scale is calculated based on three questions relating to alcohol use each on a 4 point scale. The original scores from the three AUDIT-C questions are included in the original data.

Kessler 10¹⁸

The Kessler 10 is based on 10 self reported questions relating to an individual's psychological state originally developed in 1992. Questions relate to an individual's nervousness, agitation and level of depression. The 10 questions record responses on a 5 point scale. The responses from the 10 questions have been included in their original scale.

Sensation Seeking Score

This variable is based on the Zuckerman Sensation Seeking Scale²⁷ which is a series of 19 true/false questions relating to the individual's likelihood of undertaking behaviours that may be considered less safe or planned. Areas such as impulsiveness and unpredictability are covered in this variable.

Each of the original true/false responses (coded 0/1) was included in their original scale.

Risk Perception Score

This variable is based on a series of 10 questions relating to an individual's perception of how safe they believe particular behaviours are while driving. The behaviours include driving 70km/h in a 60km/h zone, driving with a blood alcohol limit slightly over the legal limit, driving while talking on a mobile phone and driving through a red light. Each question is based on a Likert scale from 1 to 4, for "always safe" to "rarely safe". This was recorded at the time of survey (post obtaining provisional licence) and could potentially lead to slightly bias results based on the individuals driving experience prior to undertaking this survey.

The recorded observations from each of the 10 questions have been used in their original scale.

Risk Taking Behaviours

This variable is based on a series of 14 questions relating to the individual's actual driving experience and ranking the occurrence of these behaviours on a Likert scale of 1 to 4 from "never" to "very often". Behaviours, such as driving 70km/h in a 60km/h zone, drag racing, driving while talking on a mobile phone, making rude gestures at other drivers are included in this variable. This was recorded at the time of the survey (post obtaining provisional licence).

The recorded observations from each of the 14 questions have been used in their original scale.

Additional Variables

Some additional variables were included at the request of my supervisor. These were: whether the individual had ridden a motorcycle recorded as true/false (coded 0/1); if the individual had ever harmed themselves recorded as true/false (coded as 0/1); the individuals use of marijuana in the past 4 weeks and general marijuana habit recorded on a 4 point Likert scale from 'never' to 'frequently'; individuals opinion of their driving ability compared to others of similar age and experience as well as compared to the general driving population recorded on a 5 point Likert scale.

Modeling Approach

Factor analysis utilizes the correlation matrix of a series of observable variables.

The mathematical approach to determining the factors is based on a similar notion to a multiple regression, whereby the sum of squares of a dependent variable can be separated into model and residual components.

The standard model for factor analysis^{30, 31} is $x = \Lambda f + v$, where $x = (x_1, \dots, x_p)^T$ is a vector of standardised observed variables, $f = (f_1, \dots, f_k)^T$ is a vector of the latent common factors and $v = (v_1, \dots, v_p)^T$ is a vector of the latent specific factors. The matrix $\Lambda = (\lambda_{ir}) = (pxk)$ is the loading matrix.

The variables x_1, \dots, x_p are standardised, therefore $E(x_i) = 0$ and $\text{var}(x_i) = 1$ for $i = 1, \dots, p$. Assuming the common and specific factors are uncorrelated and the common factors are standardised then the covariance matrix $\Sigma = (\text{pxp})$ of x is given as $\Sigma = \Lambda\Lambda^T + V$ where $V = (\text{pxp})$ is a diagonal matrix with $v_{ii} = \text{var}(v_i)$. The $\text{var}(x_i) = c_i + \text{var}(v_i)$ with $c_i = \sum_{r=1}^k \lambda_{ir}^2$, and as a diagonal element of $\Lambda\Lambda^T$ the term c_i remains unchanged for any orthogonal transformation of the loading matrix Λ , and c_i is the communality of the variable x_i .

To determine the rotation, we denote the matrix of squared factor loadings by

$F = (f_{ir}) = (\lambda_{ir}^2)$. The orthogonal solution is given by $V = \frac{1}{k} \sum_{r=1}^k \sigma_r^2$ where

$\sigma_r^2 = \frac{1}{p} \sum_{i=1}^p (f_{ir} - \bar{f}_r)^2$, $f_{ir} = \lambda_{ir}^2$ and $\bar{f}_r = \frac{1}{p} \sum_{i=1}^p f_{ir}$. σ_r^2 is the variance of the squared

loadings. Maximizing V gives $x_v = x_Q - \frac{1}{p} \sum_{r=1}^k d_r^2$ where $d_r = p\bar{f}_r = \sum_{i=1}^p \lambda_{ir}^2$. In this

instance the criterion becomes maximal if x_Q takes on its maximum and if $\sum d_r^2$ takes on its minimum.

As a simple example, assuming a correlation matrix with equal frequencies, the explained component of the factor component equals the “outer product” of a column of “factor loadings”. The outer product from a column of correlations is the square matrix formed by letting entry ij equal the product of entries i and j in the column. Assume the following correlation matrix:

r =

1.00	0.72	0.63	0.54
0.72	1.00	0.56	0.48
0.63	0.56	1.00	0.42
0.54	0.48	0.42	1.00

This matrix of correlations has the property of having one factor variable whose correlations with the 4 observed variables are 0.9, 0.8, 0.7, and 0.6.

This can be seen by the formula for partial correlation between two variables a and b partialing out a third variable g (*latent factor*):

$$r_{ab.g} = (r_{ab} - r_{ag} r_{bg}) / \sqrt{(1 - r_{ag}^2)(1 - r_{bg}^2)}$$

From the above formula $r_{ab.g} = 0$ if and only if $r_{ab} = r_{ag} r_{bg}$. The required property for a latent factor g is that the partial correlation between any two observed variables, partialing out g , is zero.

In this instance there is a set of correlations of the observed variables with g , such that the product of any two of those correlations equals the correlation between the two observed variables. The correlation matrix above has this property as, any off-diagonal entry r_{ij} is the product of the i th and j th entries in the row 0.9, 0.8, 0.7, 0.6. For instance, the entry in row 1 and column 3 is 0.9×0.7 or 0.63. Thus the correlation matrix exactly fits the hypothesis of a single common factor.

The example above shows the common portion of the variability explained by the latent factor. To determine the unique component and therefore the degree to which the data fits the factors, an analysis of the residual correlation matrix is used. If the correlations from the residual correlation matrix are sufficiently high to consider that they are not zero in the population (considered 40% to be sufficiently high for an exploratory analysis⁷) then the number of factors is not considered appropriate.

A number of assumptions underlie the data for an appropriate factor analysis. These are:

- 1) Observed measurements should be interval or ratio based
- 2) Observed variables should be normally distributed. Variables with marked skewness or kurtosis should be transformed where possible.
- 3) Observations should be based on a random sample.

As previously indicated, the final determination of the number of factors is somewhat subjective; however, the generally accepted methodology for determining the number of factors is based on a process as follows:

- 1) Eigenvalue of one or greater. In factor analysis each observed variable contributes one unit of variance to the total variance in the dataset. Any component that has an eigenvalue of greater than one is accounting for

more variance than would be accounted for by one observed variable and is considered meaningful and retained.

- 2) **Scree test.** The scree is a plot of the eigenvalues and is used to assist in identifying any obvious breaks between components or where the plot is characterised by a bend or “elbow”. The components that are shown before an obvious break or bend in the curve (that is, on the side of the curve depicting the highest eigenvalues) are retained and those after the break or bend are not considered meaningful. Figure 2 depicts the scree plot in the present study. There is a clear difference between the eigenvalues for 1 and 2 factors than there is between those for 10 and 11 factors (or higher factor numbers). This difference between eigenvalues indicates the amount of additional variance explained by including an additional factor. If the difference is small then there may not be anything gained by including the additional factor. Likewise, a bend in the curve can be visualised between 8 and 9 factors; however, it might not be useful to select a solution with such a large number of factors if some only provide a small gain. The value of that gain is also determined by the next criterion: interpretability.
- 3) **Interpretability.** This is perhaps the most important criterion as it ensures the observable variables loading onto any factor make logical sense. There are 4 components to this test:
 1. At least 3 observable variables loading on each factor.⁶ The greater the number of variables used to determine the factor the more satisfactory the result. This is fairly easily considered from the previous example where the satisfaction of an individual’s lifestyle is determined by a number of components. If this was reduced to only 1 or 2 questions the suitability of the latent variable would be questioned. A larger number of observed variables loading onto a latent variable ensure more of the ‘true’ components of that variable are included.
 2. Variables loading on a factor share a common meaning. An example for this dataset is the questions on whether the individual is restless, sad, depressed or nervous. These can

all be referred to as the mental health of an individual and as such share a common meaning.

3. Variables loading on different factors are measuring different concepts.
4. Rotated factor pattern displays a “simple structure”. This refers to the variables loading on one factor have relatively high factor loadings while near zero loadings on the other factors. If the observed variables loads on more than one factor after rotation then the variable may not be sufficiently unique in what it is measuring.

Data Assessment and Cleaning

The original dataset contained a number of different scales and both categorical and continuous data. Based on the skewed results in the continuous variables (non-professional and professional supervised hours of driving, time on open and major roads) these were transformed into ordinal variables.

An assessment of the skewness and kurtosis of each variable was undertaken to determine if further transformations were required. The following variables had higher than anticipated skewness or kurtosis: driving without supervision, professional supervised driving at night, non-professional supervised driving on gravel roads, marijuana habit, marijuana 4 times per week, alcohol 4 times per week and driving without a seatbelt. Due to the number of respondents indicating that they did not undertake these behaviours, there was no practical solution to transform these variables. They were left in their original form, noting the possibility of affecting the final solution if they did ultimately load onto one of the factors. This had the potential for these variables to not load onto the expected latent variables due to the different variability and potential lower correlations observed between these and the other observed variables used in the known scales.

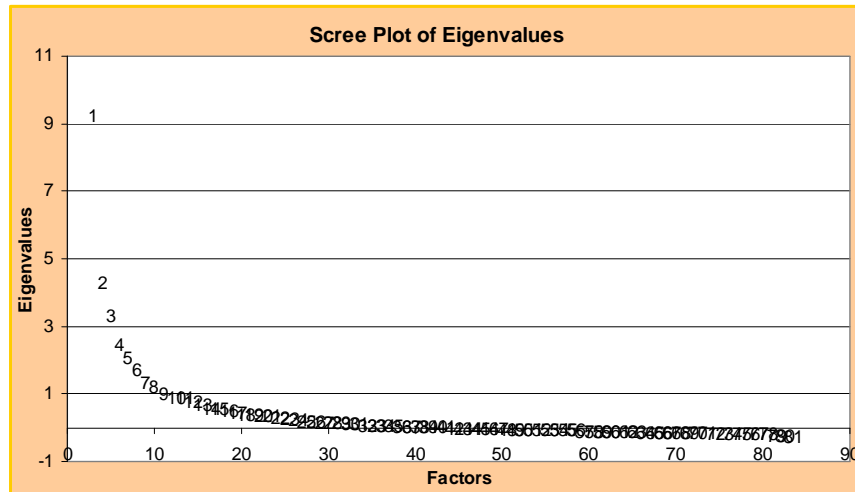
Apart from converting the true/false questions into a numerical response, a number of the sensation seeking variables had to be reversed to account for the manner in

which questions were presented. Most of the questions implied a true or more frequent response relating to potentially higher risk behaviour; however, a few questions resulted in a true response reflecting perceived safer or lower risk behaviour. The outcomes from these variables were reversed to avoid negative loadings in the factor analysis. It should be noted that reversing the order of a variable did not change the potential or absolute value for that variable to load onto a factor, only the arithmetic sign based on the correlation between the variable and the factor.

The linear transformation of the factor solution to assist in interpretation can either be undertaken assuming an orthogonal rotation or an oblique rotation. An orthogonal rotation, which assumes all factors are uncorrelated with one another, is generally implemented in a principal component analysis as the factors are already 'known' and the approach is to reduce the number of observable variables onto each factor. In an exploratory factor analysis the potential for factors to be correlated is not known therefore an oblique rotation is preferable and was undertaken for this dataset.

Utilising SAS for the factor analysis, the number of factors was not pre-determined in the initial analysis to allow the data to indicate an appropriate starting point. The scree plot shown below provides an indication of the eigenvalues and size of the "breaks" to assist in determining the number of factors.

Figure 2. Initial Scree Plot of all Variables*



* All variables includes the entire list of observed variables in their transformed state

Using the initial guidance for selecting the appropriate number of factors, the initial scree plot in Figure 2 prior to the exclusion of any variables indicates a large break between 1 and 2 factors and then further smaller breaks down to 3 and 4 factors. Additionally an 8 factor solution results in eigenvalues of greater than 1.0 which results in each of these factors explaining more of the variance. A 9 factor solution, having an eigenvalue of less than 1.0 would result in less explained variance. While it is reiterated that the final solution to a factor analysis model remains subjective, the scree plot and eigenvalues give a first indication of the potential solution.

The cutoff for determining if a variable should load onto a particular factor is not clear in the literature; however, it is apparent that a tradeoff exists between having a sufficiently low cutoff to allow enough variables to load and a cutoff that is too low resulting in meaningless variables loading on the factor. The factor correlation cutoff of 0.4⁷ was used in this analysis based on the general rule of thumb that an excellent correlation coefficient is above 0.7 and a good correlation coefficient is between 0.4 and 0.7.

Following the identification of observed variables loading onto a latent variable an assessment is required of the consistency of these variables, which addresses the extent to which each observed variable correlates with each other. The Cronbach

alpha coefficient is used to assess this internal consistency. It is used to determine the lowest estimate of reliability that can be expected for the factor. The more variables included in a factor, the higher the alpha coefficient is likely to be, assuming the variables are highly correlated with one another. This is evident from Cronbach's formula:

$$r_{xx} = \left(\frac{N}{N-1} \right) \left(\frac{S^2 - \sum_i S^2}{S^2} \right)$$

Where

r_{xx} = Coefficient alpha

N = Number of variables within the factor

S^2 = Variance of the summated scale scores. This is the variance of the total sum of each individual's responses.

$\sum_i S^2$ = The sum of the variances of the individual items within the scales

In the second term of the function the sum of the variances of the individual items within the scales is subtracted from the variance of the summated scale scores before division is performed. If the sum of the variances of the individual items within the scales is small due to high correlation then the coefficient alpha will also be high.

The subjective cutoff for the factor loadings is also apparent for the Cronbach alpha reliability coefficient. The general consensus is a reliability coefficient of at least 0.7 however a cutoff of 0.6 has been suggested as acceptable for an exploratory analysis⁸. As this is an exploratory factor analysis a cutoff of 0.6 was applied to ensure sufficient information is obtained on the potential variables loading onto factors.

The process to determine the final number of factors, involved running a series of factor analyses commencing with 8 factors based on the output from the scree plot and the number of eigenvalues above 1.0. As each factor analysis was run the non-loading variables were removed. This initially resulted in a number of latent variables with too few loaded variables. Therefore the number of factors was

reduced at each step until the criteria outlined in the interpretability criteria were met. This ultimately resulted in 6 factors by removing the non-loading variables and reducing the factor numbers if insufficient variables were loading. The reliability coefficient was then calculated to ensure each factor had a coefficient of at least 0.60.

Results

The results of three separate factor models are presented below: an overall factor model and a factor model for each gender. The gender models were included to assess whether differences existed in the correlation between behaviours in the sexes and subsequently whether items loaded differently onto factors between males and females. The research literature has repeatedly found young males to be at greater risk of fatal and serious injury crashes than young females^{9,10,11}, independent of other known risk factors, and has shown, for example, that different factors predict crash injury severity for males and females¹². Therefore it is plausible that the underlying factor structures might differ between the two sexes. Additional to the questions that loaded onto the factors, the questions comprising previously used scales (eg. Kessler 10, risk taking behaviour, sensation seeking, risk perception) that did not load onto these factors are included in tables 1 to 3.

Table 1. Questionnaire Items and Corresponding Factor Loadings from the Rotated Factor Pattern Matrix (all data)

Factor Structure						Questionnaire Item
Risk Taking Behaviour While Driving (1)						
(1)	(2)	(3)	(4)	(5)	(6)	How often do you:
0.64	0.01	0.03	-0.17	0.04	0.03	Drive fast just for the thrill of it
0.63	-0.01	0.03	-0.18	0.04	0.05	Take some risks when driving because it makes driving more fun
0.67	0.03	-0.05	0.03	0.06	0.04	Drive at about 70km/h in a 60km/h speed zone
0.51	-0.10	0.05	-0.17	-0.06	0.03	Do burnouts, donuts, or skids just for the fun of it
0.59	0.06	0	0.05	0.05	-0.06	Speed up if someone is trying to pass you
0.60	0.03	0	0.07	0.03	-0.01	Follow very close behind slower drivers
0.52	0	0.07	-0.03	-0.06	-0.06	Make rude gestures at other drivers
0.55	0.02	0.08	0.05	-0.03	-0.06	Honk your horn or flash your lights in anger at other drivers
0.63	-0.06	0.05	-0.08	-0.04	0	Race or drag race for the fun of it
0.55	0.03	-0.10	0.10	-0.02	-0.04	Drive while talking on a mobile phone
0.45	0.05	0.03	-0.14	-0.07	0.01	Drive while listening to loud music
0.51	0.04	-0.11	0.11	-0.05	-0.03	Drive while using SMS on a mobile phone
						Drive without wearing a seatbelt
						Drive with 2 or more passengers

The following questions did not load onto this factor

Mental Health Status (2)

(1)	(2)	(3)	(4)	(5)	(6)	
0.11	0.50	-0.01	0.02	0.02	0.01	During the past 4 weeks:
-0.01	0.52	-0.01	0.07	0.07	0.04	How often did you feel tired out for no reason
0	0.57	-0.02	0.02	0.03	-0.04	How often did you feel nervous
-0.05	0.77	-0.01	0	-0.01	0.02	How often did you feel so nervous that nothing could calm you down
0.03	0.46	0	-0.24	0.01	0.01	How often did you feel hopeless
0.02	0.42	-0.02	-0.27	-0.01	-0.05	How often did you feel restless or fidgety
-0.02	0.78	0.01	0.03	-0.03	0.02	How often did you feel so restless you could not sit still
0.10	0.59	0	-0.02	0.03	0.02	How often did you feel depressed
-0.01	0.78	0.01	0.03	-0.04	-0.02	How often did you feel that everything was an effort
-0.03	0.76	0.01	0.01	-0.04	-0.01	How often did you feel so sad that nothing could cheer you up
						How often did you feel worthless

Perception of Risks While Driving (3)

(1)	(2)	(3)	(4)	(5)	(6)	
-0.34	0.01	0.41	-0.04	-0.07	-0.05	When you are driving, how safe do you think the following are
-0.32	0.04	0.40	0.02	-0.02	-0.10	Driving at 70km/h in a 60km/h speed zone
0.09	0	0.70	0.05	-0.01	0.01	Driving at 110km/h in a 100km/h speed zone
0.14	-0.01	0.70	0.07	0.04	0.04	Driving with a blood alcohol level just over the legal limit
0.03	-0.01	0.60	0.06	-0.01	-0.06	Driving after smoking marijuana
0.08	-0.02	0.68	0	0	0.06	Driving a poorly maintained car
-0.24	0.03	0.59	-0.04	0	0	Going through a red light
-0.13	0.01	0.65	-0.07	0.05	0.04	Driving while talking on a mobile phone
						Driving while using SMS (text messaging) on a mobile phone

The following questions did not load onto this factor

Driving with 2 or more passengers
Driving between midnight and 6am

Sensation Seeking Behaviours (4)

(1)	(2)	(3)	(4)	(5)	(6)	
-0.02	-0.06	0.02	0.50	0.01	0.02	Does the following statement describe you or does not describe you
0.03	0.04	0.01	0.63	-0.03	-0.02	I often do things on impulse
0.05	0.02	0.03	0.49	0	0.03	I like to have new and exciting experiences and sensations even if they are a little frightening
-0.06	0.02	-0.01	0.63	-0.01	-0.03	I enjoy getting into new situations where you can't predict how things will turn out
-0.01	0.04	0	0.60	0.02	-0.03	I like doing things just for the thrill of it
-0.09	-0.03	0	0.55	0.02	-0.01	I sometimes like to do things that are a little frightening
0.01	-0.04	0.02	0.45	-0.02	0.02	I sometimes do crazy things just for fun
0	-0.05	0.02	0.54	-0.01	0.05	I prefer friends who are excitingly unpredictable
						I am an impulsive person

The following questions did not load onto this factor

I tend to begin a new job without much advance planning on how I will do it
I usually think about what I am going to do before I do it
I very seldom spend much time on the details of planning ahead
Before I begin a complicated job, I make careful plans
I would like to take off on a trip with no pre-planned routes or timetables
I tend to change interests frequently
I'll try anything once
I would like the kind of life where one is on the move and traveling a lot, with lots of change and excitement
I like to explore a strange city or section of town by myself, even if it means getting lost
I often get so carried away by new and exciting things that I never think of possible complications
I like wild uninhibited parties

Non-Professional Driver Training Experience (5)

(1)	(2)	(3)	(4)	(5)	(6)	
						About how often did you drive on the following types of roads with a non-professional instructor while learning to drive:

0.04	0	-0.02	-0.04	0.57	-0.28	Total non-professional hours
0	0.02	0.01	0.03	0.58	0.19	Major roads (70-80km/h)
0.03	0.04	0.02	0.07	0.55	0.10	Open roads (90-110km/h)
-0.05	-0.01	-0.01	-0.05	0.68	-0.01	When it was raining
-0.07	-0.01	-0.02	-0.06	0.65	-0.06	When it was dark
0.05	-0.04	0.03	0.05	0.53	0.03	On gravel (unsealed) roads
-0.05	0	-0.02	-0.03	0.63	0.04	In heavy traffic

The following question did not load onto this factor [Residential roads \(60km/h or less\)](#)

Professional Driver Training Experience (6)

(1)	(2)	(3)	(4)	(5)	(6)	About how often did you drive on the following types of roads with a professional instructor while learning to drive:
-0.02	-0.04	-0.03	-0.04	-0.25	0.62	Total professional hours
-0.03	0.04	0.01	0.03	0.18	0.63	Major roads (70-80km/h)
0.01	0.05	0.04	0.05	0.18	0.56	Open roads (90-110km/h)
-0.02	-0.01	-0.01	-0.03	-0.11	0.63	When it was raining
-0.04	0.02	0.01	-0.01	-0.09	0.44	When it was dark
0.02	-0.05	0.03	0.01	0.07	0.43	On gravel (unsealed) roads
-0.02	0	-0.03	0	0.02	0.65	In heavy traffic

The following question did not load onto this factor [Residential roads \(60km/h or less\)](#)

Table 2. Questionnaire Items and Corresponding Factor Loadings from the Rotated Factor Pattern Matrix (Females)

Factor Structure Questionnaire Item

Mental Health Status (1)

(1)	(2)	(3)	(4)	(5)	(6)	During the past 4 weeks:
0.49	0.13	0.01	0.02	0	0.03	How often did you feel tired out for no reason
0.53	-0.01	0.08	0.07	-0.01	0.01	How often did you feel nervous
0.58	-0.02	0.04	0.04	-0.03	-0.04	How often did you feel so nervous that nothing could calm you down
0.77	-0.05	0	-0.01	0	0.01	How often did you feel hopeless
0.46	0.06	-0.20	0.01	-0.01	0	How often did you feel restless or fidgety
0.41	0.01	-0.25	-0.01	-0.03	-0.05	How often did you feel so restless you could not sit still
0.78	0	0.04	-0.03	0.02	0.01	How often did you feel depressed
0.63	0.09	0	0.02	0.02	0.02	How often did you feel that everything was an effort
0.78	-0.01	0.03	-0.04	0.01	-0.02	How often did you feel so sad that nothing could cheer you up
0.78	-0.05	0.02	-0.05	0	-0.02	How often did you feel worthless

Risk Taking Behaviour While Driving (2)

(1)	(2)	(3)	(4)	(5)	(6)	How often do you:
0.04	0.56	-0.20	0.05	0.03	0.03	Drive fast just for the thrill of it
0.02	0.52	-0.20	0.04	0.01	0.05	Take some risks when driving because it makes driving more fun
0.02	0.64	0	0.03	-0.01	0.03	Drive at about 70km/h in a 60km/h speed zone
0.04	0.55	0.02	0.05	0	-0.04	Speed up if someone is trying to pass you
0.02	0.58	0.03	0.02	0.01	-0.01	Follow very close behind slower drivers
-0.01	0.46	-0.02	-0.05	0.05	-0.02	Make rude gestures at other drivers
0	0.51	0.03	-0.02	0.06	-0.03	Honk your horn or flash your lights in anger at other drivers
-0.01	0.46	-0.10	-0.05	0.03	0.02	Race or drag race for the fun of it
-0.03	0.64	0.09	-0.01	-0.06	-0.03	Drive while talking on a mobile phone
0.03	0.47	-0.16	-0.07	0.02	0.03	Drive while listening to loud music
-0.03	0.61	0.09	-0.05	-0.09	-0.01	Drive while using SMS on a mobile phone

The following questions did not load onto this factor [Do burnouts, donuts, or skids just for the fun of it](#)
[Drive without wearing a seatbelt](#)
[Drive with 2 or more passengers](#)

Sensation Seeking Behaviours (3)

(1)	(2)	(3)	(4)	(5)	(6)	
-0.03	-0.03	0.49	0.01	0.03	0.01	Does the following statement describe you or does not describe you
						I often do things on impulse
0.04	0.03	0.63	-0.03	0	-0.01	I like to have new and exciting experiences and sensations even if they are a little frightening
0.03	0.05	0.49	0	0.02	0.01	I enjoy getting into new situations where you can't predict how things will turn out
0.03	-0.04	0.63	-0.01	-0.01	-0.02	I like doing things just for the thrill of it
0.02	0.01	0.59	0.02	-0.01	0	I sometimes like to do things that are a little frightening
-0.01	-0.06	0.56	0.03	0	-0.01	I sometimes do crazy things just for fun
-0.02	0	0.47	-0.02	0.01	0.01	I prefer friends who are excitingly unpredictable
-0.01	-0.03	0.54	0	0	0.03	I am an impulsive person
0.01	-0.04	0.40	0	0.02	0.03	I'll try anything once
						I tend to begin a new job without much advance planning on how I will do it
						I usually think about what I am going to do before I do it
						I very seldom spend much time on the details of planning ahead
						Before I begin a complicated job, I make careful plans
						I would like to take off on a trip with no pre-planned routes or timetables
						I tend to change interests frequently
						I would like the kind of life where one is on the move and traveling a lot, with lots of change and excitement
						I like to explore a strange city or section of town by myself, even if it means getting lost
						I often get so carried away by new and exciting things that I never think of possible complications
						I like wild uninhibited parties

The following questions did not load onto this factor

Non-Professional Driver Training Experience (4)

(1)	(2)	(3)	(4)	(5)	(6)	
-0.01	0.06	-0.03	0.58	0	-0.27	About how often did you drive on the following types of roads with a non-professional instructor while learning to drive:
0.03	-0.01	0.03	0.59	0	0.17	Total non-professional hours
0.03	0.03	0.06	0.56	0.01	0.09	Major roads (70-80km/h)
0	-0.05	-0.06	0.68	-0.01	0	Open roads (90-110km/h)
0	-0.07	-0.06	0.64	-0.01	-0.04	When it was raining
-0.04	0.04	0.06	0.52	0.03	0.03	When it was dark
0	-0.04	-0.02	0.64	-0.03	0.07	On gravel (unsealed) roads
						In heavy traffic
						Residential roads (60km/h or less)

The following question did not load onto this factor

Risk Perception While Driving (5)

(1)	(2)	(3)	(4)	(5)	(6)	
-0.01	0.07	0.04	-0.02	0.73	0.01	When you are driving, how safe do you think the following are
-0.01	0.13	0.06	0.02	0.75	0.02	Driving with a blood alcohol level just over the legal limit
-0.02	0.02	0.05	-0.01	0.61	-0.06	Driving after smoking marijuana
-0.02	0.05	0	-0.02	0.69	0.03	Driving a poorly maintained car
0.03	-0.28	-0.05	-0.02	0.56	-0.01	Going through a red light
0.03	-0.22	-0.08	0.04	0.63	0.02	Driving while talking on a mobile phone
						Driving while using SMS (text messaging) on a mobile phone
						Driving at 70km/h in a 60km/h speed zone
						Driving at 110km/h in a 100km/h speed zone
						Driving with 2 or more passengers
						Driving between midnight and 6am

The following questions did not load onto this factor

Professional Driver Training Experience (6)

(1)	(2)	(3)	(4)	(5)	(6)	
-0.02	-0.03	-0.03	-0.26	-0.03	0.62	About how often did you drive on the following types of roads with a professional instructor while learning to drive:
0.03	0	0.03	0.17	-0.02	0.63	Total professional hours
						Major roads (70-80km/h)

0.03	0.03	0.05	0.19	0.03	0.56	Open roads (90-110km/h)
-0.01	-0.02	-0.03	-0.12	0	0.62	When it was raining
0.01	0	-0.03	-0.12	0.03	0.43	When it was dark
-0.05	0.02	0.02	0.09	0.02	0.42	On gravel (unsealed) roads
-0.02	-0.01	0	0.03	-0.03	0.66	In heavy traffic

The following question did not load onto this factor [Residential roads \(60km/h or less\)](#)

Table 3. Questionnaire Items and Corresponding Factor Loadings from the Rotated Factor Pattern Matrix (Males)

Factor Structure	Questionnaire Item
------------------	--------------------

Risk Taking Behaviour While Driving (1)

(1)	(2)	(3)	(4)	(5)	(6)	How often do you:
0.65	0.04	0.02	0.03	0.01	-0.17	Drive fast just for the thrill of it
0.65	0.04	0.02	0.04	0.03	-0.17	Take some risks when driving because it makes driving more fun
0.65	0	-0.07	0.08	0.06	-0.01	Drive at about 70km/h in a 60km/h speed zone
0.56	-0.05	0.05	-0.09	0	-0.15	Do burnouts, donuts, or skids just for the fun of it
0.63	0.06	-0.01	0.05	-0.05	0.09	Speed up if someone is trying to pass you
0.63	0	-0.02	0.03	0.01	0.09	Follow very close behind slower drivers
0.57	-0.01	0.07	-0.06	-0.08	0	Make rude gestures at other drivers
0.61	0.01	0.08	-0.04	-0.06	0.09	Honk your horn or flash your lights in anger at other drivers
0.67	-0.02	0.04	-0.03	-0.03	-0.06	Race or drag race for the fun of it
0.53	0.01	-0.14	-0.01	-0.01	0.1	Drive while talking on a mobile phone
0.46	0.01	0.05	-0.06	0.01	-0.14	Drive while listening to loud music
0.51	0	-0.14	-0.04	0.01	0.1	Drive while using SMS on a mobile phone

The following questions did not load onto this factor [Drive without wearing a seatbelt](#)
[Drive with 2 or more passengers](#)

Mental Health Status (2)

(1)	(2)	(3)	(4)	(5)	(6)	During the past 4 weeks:
0.14	0.45	-0.02	0.03	0	0	How often did you feel tired out for no reason
0	0.51	0	0.06	0.05	0.08	How often did you feel nervous
0.03	0.57	0	0.01	-0.06	0.03	How often did you feel so nervous that nothing could calm you down
-0.04	0.77	-0.01	-0.01	0.03	-0.01	How often did you feel hopeless
-0.04	0.79	0	-0.03	0.01	-0.02	How often did you feel depressed
0.1	0.54	-0.01	0.03	0.01	-0.02	How often did you feel that everything was an effort
-0.02	0.79	0	-0.04	-0.03	-0.01	How often did you feel so sad that nothing could cheer you up
-0.03	0.76	0.01	-0.04	-0.02	-0.03	How often did you feel worthless

The following questions did not load onto this factor [How often did you feel restless or fidgety](#)
[How often did you feel so restless you could not sit still](#)

Risk Perception While Driving (3)

(1)	(2)	(3)	(4)	(5)	(6)	When you are driving, how safe do you think the following are
-0.27	0	0.46	-0.06	-0.04	0.01	Driving at 70km/h in a 60km/h speed zone
-0.26	0.03	0.43	-0.02	-0.1	0.04	Driving at 110km/h in a 100km/h speed zone
0.09	-0.02	0.67	-0.02	0.02	0.04	Driving with a blood alcohol level just over the legal limit
0.12	-0.03	0.66	0.04	0.04	0.06	Driving after smoking marijuana
0.03	-0.02	0.58	-0.01	-0.05	0.04	Driving a poorly maintained car
0.08	-0.02	0.67	0.01	0.07	-0.02	Going through a red light
-0.22	0.04	0.6	0	-0.01	-0.04	Driving while talking on a mobile phone
-0.11	0.02	0.66	0.05	0.04	-0.06	Driving while using SMS (text messaging) on a mobile phone

The following questions did not load onto this factor [Driving with 2 or more passengers](#)
[Driving between midnight and 6am](#)

Professional Driver Training Experience (4)

(1)	(2)	(3)	(4)	(5)	(6)	About how often did you drive on the following types of roads with
-----	-----	-----	-----	-----	-----	--

0.02	0.01	-0.03	0.57	-0.28	-0.05	a professional instructor while learning to drive:
0.01	0.03	0.02	0.56	0.2	0.05	Total professional hours
0.04	0.04	0.03	0.53	0.13	0.1	Major roads (70-80km/h)
-0.04	-0.01	-0.01	0.69	-0.03	-0.04	Open roads (90-110km/h)
-0.08	-0.02	-0.03	0.66	-0.1	-0.07	When it was raining
0.05	-0.04	0.04	0.54	0.04	0.06	When it was dark
-0.06	-0.01	-0.02	0.63	0.02	-0.07	On gravel (unsealed) roads
						In heavy traffic

The following question did not load onto this factor Residential roads (60km/h or less)

Non-Professional Driver Training Experience (5)

(1)	(2)	(3)	(4)	(5)	(6)	About how often did you drive on the following types of roads with a non-professional instructor while learning to drive:
-0.02	-0.04	-0.03	-0.25	0.61	-0.04	Total non-professional hours
-0.03	0.04	0.02	0.19	0.63	0.02	Major roads (70-80km/h)
0.02	0.03	0.05	0.18	0.57	0.05	Open roads (90-110km/h)
-0.02	0	-0.02	-0.09	0.64	-0.03	When it was raining
-0.05	0.01	0	-0.07	0.46	-0.01	When it was dark
0.02	-0.04	0.04	0.04	0.44	0	On gravel (unsealed) roads
-0.02	0	-0.03	0	0.64	-0.02	In heavy traffic

The following question did not load onto this factor Residential roads (60km/h or less)

Sensation Seeking Behaviours (6)

(1)	(2)	(3)	(4)	(5)	(6)	Does the following statement describe you or does not describe you
0.06	0.02	0.02	-0.01	-0.02	0.65	I like to have new and exciting experiences and sensations even if they are a little frightening
0.01	0	0.03	-0.01	0.05	0.45	I enjoy getting into new situations where you can't predict how things will turn out
-0.06	0	-0.01	-0.01	-0.02	0.64	I like doing things just for the thrill of it
0.02	0.02	0.02	0.02	-0.03	0.64	I sometimes like to do things that are a little frightening
-0.14	-0.03	-0.01	0	0	0.52	I sometimes do crazy things just for fun

The following questions did not load onto this factor

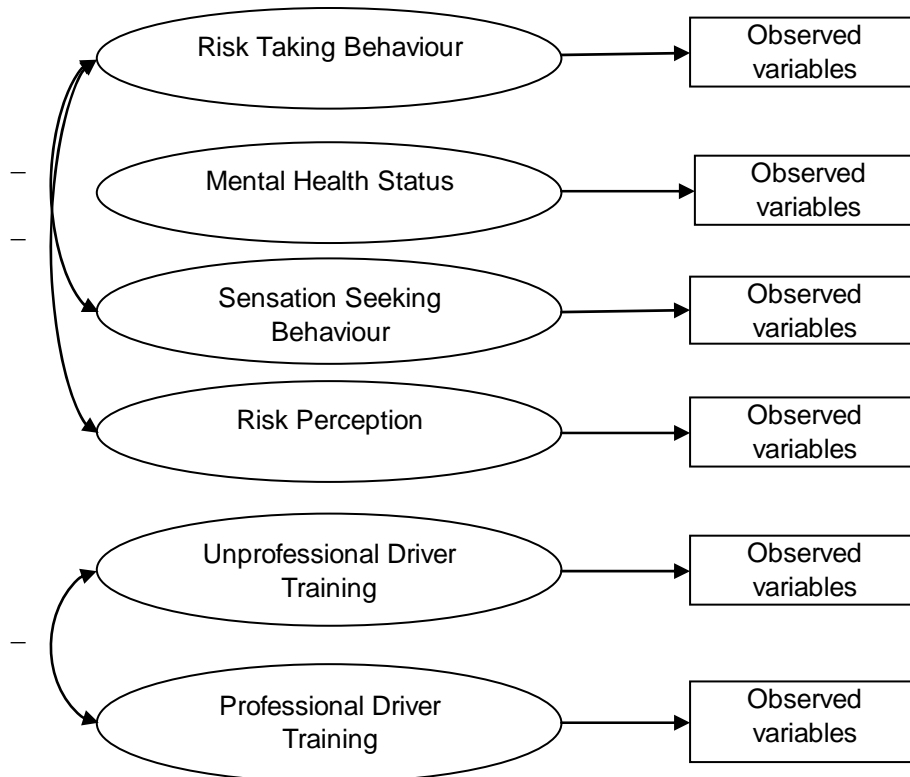
I tend to begin a new job without much advance planning on how I will do it
 I usually think about what I am going to do before I do it
 I often do things on impulse
 I very seldom spend much time on the details of planning ahead
 Before I begin a complicated job, I make careful plans
 I would like to take off on a trip with no pre-planned routes or timetables
 I tend to change interests frequently
 I'll try anything once
 I would like the kind of life where one is on the move and traveling a lot, with lots of change and excitement
 I like to explore a strange city or section of town by myself, even if it means getting lost
 I prefer friends who are excitingly unpredictable
 I often get so carried away by new and exciting things that I never think of possible complications
 I am an impulsive person
 I like wild uninhibited parties

Table 4. Cronbach Reliability Coefficient Alpha Scores

Latent Variables	Overall	Females	Males
Risk Taking Behaviour While Driving	0.86	0.83	0.87
Mental Health Status	0.86	0.87	0.86
Sensation Seeking Behaviours	0.79	0.79	0.75
Risk Perception While Driving	0.83	0.84	0.83
Professional Driver Training	0.62	0.62	0.64
Non-Professional Driver Training	0.70	0.72	0.67

The factor structure of correlations detailed in the results tables can be displayed in the path model depicted in Figure 3. A path model is a visual representation of the observed variables and how they load onto the latent variable. It also includes the nature of any correlations between the latent variables when an oblique rotation is undertaken in the factor analysis. The observed variables are those listed in each of the 6 latent factors:

Figure 3. Path model for the DRIVE dataset.



The factor loadings in the results (Tables 1, 2 and 3) show the simple structure loadings of each question and how it loads onto each latent variable. It also shows the impact of that question on each of the other latent variables. In an uncorrelated solution each of the loadings for questions on other latent variables should be very low. In a correlated solution these loadings will be either positive or negative and slightly higher than zero. The factor structure of correlations indicates a negative correlation between, risk taking behaviour and risk perception, and non-professional and professional supervised driver training. While these are not consistently strong, there are sufficient negative correlations to support this path model. Some statistically significant negative correlation loadings on the factor structure are:

- Professional supervised driving hours with non-professional supervised driving hours.
- Perceptions of driving 110km/h in a 100km/h zone, driving 70km/h in a 60km/h zone and driving from midnight until 6am with risk taking behaviours.

Risk Taking Behaviour While Driving

The questions loading onto the Risk Taking Behaviour factor essentially correspond to the original questions grouped within the risk taking behaviour questionnaire. However, the questions on the extent to whether an individual regularly wears a seatbelt and drives with 2 or more passengers did not load in any of the factor models. The extent to which an individual performs burnouts in their car did load on the male factor model however did not load on the female factor model. The reliability alpha score for this factor in each of the 3 models was consistently high at 83%-87%.

Mental Health Status

The questions in this factor correspond to the Kessler 10 index for mental health. All questions loaded onto the mental health status factor in the overall and male models however *restless* and *very restless* did not load on the female model. The reliability index in each of the 3 models was consistently high at 86%-87%.

Sensation Seeking Behaviours

The questions related to this factor were based on the Zuckerman Sensation Seeking Scale. Of the 19 original questions only 8 loaded onto this factor in the overall model and 9 loaded onto the male factor model. The variable related to *trying anything once* was the additional variable in the male model. The female factor model reduced further to only 5 loaded variables. The 4 variables that did not load onto the female model compared to the male model were, *impulsiveness*, *impulsive person*, *unpredictable friends*, and *try anything once*. The reliability scores for this factor were somewhat lower but still in the acceptable range at 75%-79%.

Risk Perception While Driving

Of the 10 original questions relating to risk perception, the same 8 of these loaded onto the overall and male factor models and a reduced set of 6 onto the female factor model. The 2 questions that did not load onto any model were related to the individual's perception of how safe it was to drive with 2 or more passengers and driving between midnight and 6am. The additional 2 questions that did not load onto the female factor model related to the perception of how safe it was to drive 70km/h in a 60km/h zone and driving 110km/h in a 100km/h zone. The reliability scores for this factor were 83%-84%.

Professional Driver Training

The professional driver training factor relates to different types of professional driving experience over different road types and driving conditions while under a professional driving instructor. The question relating to the length of time spent driving on residential roads while under professional driving instruction did not load on any of the factor models otherwise each model had the same questions loading on this factor. The reliability scores for this factor were more moderate: between 64%-66%.

Non-Professional Driver Training

The non-professional driver training factor relates to different types of driving experience over different road types and driving conditions while under a non-

professional driving instructor such as a parent or relative. Similar to the professional driver training the question relating to the length of time spent driving on residential roads while under non-professional driving instruction did not load on any of the factor models, otherwise each model had the same items loading on this factor. The reliability scores for this factor were between 73%-77%.

Discussion

This factor analysis is one small sub-study of the wider group of analyses undertaken on the DRIVE study data and the outcome of this study and consideration of the significance and direction of the final variables in this model need to be viewed in the context of the broader study. The factors derived from this analysis provided support that the observed variables loaded onto known scales although some differences between males and females occurred. Differences were observed between variables loading onto known scales between males and females in the risk perception, risk taking behaviours and sensation seeking factors. The scales that had a larger range of responses (rather than dichotomous true/false responses) resulted in more of the original variables loading onto the known factors. Additionally, there is scope to introduce an additional factor in the sensation seeking scale between unplanned and planned sensation seeking behaviours.

A large portion of the questions used in this sub-study were based on previously validated and widely used scales. This is specifically the case for latent variables such as the Kessler 10 mental health scale, and the Zuckerman Sensation Seeking Scale. Part of this sub-study is to consider the extent to which the results of the questions in this dataset load onto the same latent variables. The following discussion outlines the expected and unanticipated differences observed in the factor loadings.

The extent to which any variable will load onto a latent variable is subject to the variability that can be explained by the factor. In the case of the risk taking behaviour factor, it is intuitive that not wearing a seatbelt while driving is a risky behaviour; however, since wearing a seatbelt is such a common practice in

Australia, there is very little variability in the responses from this question. This resulted in skewed responses and similar correlations with a number of factors. Therefore the question did not load onto the risk taking behaviour factor. This questions whether the original scale requires altering if a behaviour such as wearing a seatbelt becomes too consistent. This behaviour may be somewhat particular to Australians and would vary widely in other countries, eg. United States, therefore it may be appropriate to exclude this item in the Australian context given the high rate of wearing seatbelts¹³.

A slightly different situation occurred in the risk perception factor where the variability on the questions of how safe driving between midnight and 6am and driving with 2 or more passengers was materially different to the skewed data on the other questions in this factor. The other risk perception questions loading onto this factor had skewed responses where the majority of outcomes were in the never safe or rarely safe categories. This was different in the driving at midnight and passenger questions where the data was skewed towards always being safe or mostly safe.

The number of observed variables loading onto factors determined in this analysis varies slightly between the overall model and the gender specific models. The male model had one observable variable (undertaking burnouts/skids just for the fun of it) that did not load on the female model. There are driving behaviours that may be considered more male oriented, such as performing burnouts and hence have a higher correlation with other variables in the risk taking behaviours factor for males than for females. This would explain why the question loaded onto the male factor model, but not the female factor model for which the variability of responses was greatly reduced. In the risk perception factor the two observable variables that did not load onto the female model, compared to the male model, were the perception of how safe it was to drive 70km/h in a 60km/h zone and driving 110km/h in a 100km/h zone. There is scope for future studies to consider the differences in these risk perceptions between males and females.

Within the series of questions related to sensation seeking, prior to determining the final factor model there were two separate factors developing in the female factor model. The additional factor of variables that were ultimately excluded could be considered as a group based on planning and preparedness. These were *job planning, think before doing things, doing things on impulse, seldom planning ahead, planning before a complicated job, and being an impulsive person*. While these variables were ultimately dropped, there appears scope to include more questions related to planning and preparedness features, which may result in an additional viable factor. Interestingly, the specific questions on planning do not preclude an individual to undertake sensation seeking behaviours but do suggest that the correlation within females for the planning and preparedness questions is higher than the remaining sensation seeking behaviours. For example, sky diving would be considered a sensation seeking behaviour, however there would be a large amount of planning undertaken before participating in this sport. There seems scope to segregate these questions into behaviours that might be considered organised or planned sensation seeking behaviours and a separate unplanned or more reckless sensation seeking behaviours.

There is no universal agreement on the cut-off for the Cronbach alpha reliability score though a score above 60%-70% is fairly regularly reported⁸. Due to this slight inconsistency I have left the factors in the model that exceed a reliability score of 60%. The professional driving instruction for all models had the lowest reliability scores and this was largely due to the lower correlations of driving experience on open and major roads to this factor. The main reason for the lower reliability scores is the slight skewness of this data, which did not correlate to the other driver training variables as well. It should be noted that even removing the open and major road variables only increased the reliability coefficient marginally due to the generally lower correlations of the remaining professional driving variables with the professional driving factor.

One of the underlying assumptions for exploratory factor analysis is that each variable should be normally distributed. This becomes very difficult to achieve on an observed variable such as the number of hours driven under professional

instruction. A transformation of these variables into a categorical variable still resulted in slightly skewed data due to the high proportion of drivers with little or no professional driving instruction therefore some violation of this assumption was unavoidable.

The results of the factor structure of correlations provide an interesting outcome. Risk taking behaviour factor is negatively correlated with the risk perception factor. Additionally, the professional driving training is negatively correlated with non-professional supervised training. The negative correlation between risk taking and risk perception is somewhat intuitive.

There remains a limitation on the use of different scales within the factor analysis, particularly the sensation seeking questions which relied on a true/false response. This reduces the range of responses and may result in people responding in a manner that does not completely reflect their behaviours. For example an individual who occasionally “likes doing things just for the thrill of it” may respond either true or false depending on the extent to which this is actually true. A larger choice of options may provide a greater range of responses and more accurate correlations.

Conclusion

This sub-study presents an assessment of the factor loadings from the output of a number of respondent survey questions from the DRIVE dataset both overall and by gender. Specifically, factor loadings were explored among a number of pre-determined questions used in known psychological scales as well as additional variables that may load onto these or separate factors.

From the analysis undertaken, the original questions and scales previously designed or developed by DRIVE Study investigators to assess risk taking, mental health and risk perception factors are largely consistent with the factor loadings observed from the DRIVE sample dataset. Of the other variables there appears to be important differences in the factor loadings between males and females on the sensation seeking factor which raises questions over the use of one simple scale

for both males and females to assess sensation seeking. Part of the reason for this arises from the response scale used, that is, only true/false options, which do not accommodate different levels of sensation seeking behaviour, but rather an all or nothing response. This is one area that requires further exploration and a possible change in the response scale.

Consideration should also be given to implementing a survey in which all items incorporate a consistent response scale. For example, some items in the survey allowed for 5 levels of response, however others were true/false (2 levels of response); therefore the potential variability within the true/false questions was reduced compared to the 5 point scale. This can result in a loss of information when, for example, the respondent may undertake the behaviour on an occasional basis and therefore a more correct reflection of responses could be achieved from additional response levels. The variability incorporated from a small scale potentially inhibits some relevant variables loading onto a given factor.

One limitation of this modeling approach includes the requirement of each observed variable having a normal distribution. A number of the observed variables had skewed observations therefore transformation were not easily obtainable and some violations were unavoidable.

More complex factor analysis approaches have been reported in the literature to determine factor models with outcome variables in the exponential family¹⁴ or with non-normal latent variables. Given the mixed scale of the observed variables (ordinal, nominal, binary and continuous) future exploration into this approach should be considered.

Project B

Front Sheet: Project B

Project Title

Predictors of Time to First Crash for Young Drivers: Survival Analysis.

Location and Dates

School of Public Health, University of Sydney

September 2007 – December 2008

Context

The George Institute is undertaking a number of studies based on the DRIVE dataset to ultimately assist in a greater understanding of the behaviour of young drivers and their propensity to crash. This analysis was indicated to be of interest to The George Institute to obtain a better understanding of the behaviours that influence the time until a young driver is involved in their first vehicle crash following provisional licensure.

Student Contribution

The initial decision for this project was at the suggestion of The George Institute. The initial collation of data and transformation into previously used scales was undertaken by members of The George Institute prior to me receiving the information. Regular student liaison with the project supervisors took place throughout the project. Background research on prior related studies was undertaken to establish accepted methods and previously determined variables that should be included in the sub-study. As outlined previously, the SAS and Stata code to perform the statistical analyses was developed by me apart from the construction of a number of predefined measures (eg. Kessler 10 index, SEIFA indexes) that were supplied in the original dataset. The major component was the analysis and interpretation of results to ensure the output was relevant.

Statistical Issues

The statistical issues encountered in this project were the proficiency in understanding the Cox Proportional Hazards model for the dataset provided. I had previously used survival analysis in the BCA course; however, the application to a real world dataset raised its own nuances and challenges in interpreting data and justifying this to interested parties.

Student Declaration

I declare that this project is my own work, with guidance provided by my content matter supervisor, Dr Alexandra Martiniuk and statistical supervisor, Dr Stephane Heritier, and is not the result of collaboration with others and that I have not previously submitted it for academic credit.



David Warr

19 May 2009

Supervisor Declaration

I declare that David Warr has worked alone on this project under our supervision. This work has not previously been submitted for publication or academic credit.



Alexandra Martiniuk, senior research fellow

19 May 2009



Stephane Heritier, senior lecturer

19 May 2009

Project Description

Background

Previous driver related studies have indicated elevated rates of crashes in young drivers (16-19 years) compared to drivers over the age of 20 years.¹⁵

As a result of these elevated crash rates in young drivers the DRIVE Study was initiated to investigate the effect of numerous factors on the risk of vehicle crashes and related injuries on drivers between the ages of 17 to 24 who have recently obtained their driver's licence.

Following research of the above topic it appears minimal published analysis of time until first crash in young drivers has been undertaken around the world. While most of the focus has been on factors that affect whether a young driver will crash, little is known about the time until the first crash. One such study¹⁶ was undertaken during 1997 to 1998 and considered the difference between rural and urban drivers during the 12 months after obtaining a provisional licence. This was assessed against driver training prior to obtaining a learner's permit and risk taking behaviour. Young drivers who displayed confidence/adventurousness while driving were at twice the risk of crashing early (hazard ratio 2.04) compared to young drivers who displayed low levels of confidence/adventurousness. The ability to mitigate some of these factors may reduce the mortality rate in young driver crashes and reduce the frequency of these driver crashes.

Aims of Project

Studies on crashes in young drivers usually target the number of crashes or offenses. One of the originalities of this work is the focus on time to crash bearing in mind that such an undertaking has its challenges.

1) The first objective of this work is to find a reasonable way to measure "time to crash", whether we use a) the number of days until the first police crash is reported or b) the number of driving hours until the crash occurs. A discussion of the

limitations of the choice of a particular outcome is particularly relevant for DRIVE and future studies in this area.

2) The second objective is to undertake the analysis of the DRIVE data using survival analysis, for time to first crash as recorded by police in drivers aged 17 to 24 years, after obtaining their provisional licence. These are crashes where police were in attendance following a reported crash and made a subsequent record of the incident. Specifically, we would like:

- I. To examine whether there is a difference in time until first police reported vehicle crash between the following independent variables:
 - i. Gender
 - ii. Age (17 to 24 years) at time of obtaining provisional licence
 - iii. Driver training and experience (types, average hours)
 - iv. Risk taking behaviour
 - v. Types of driver training
 - vi. Sensation-seeking score
 - vii. Kessler's psychological distress score
 - viii. Risk perception score
 - ix. Urban/rural status
 - x. Country of birth
 - xi. Number of driving offences
 - xii. Socio-economic status (SES)
- II. Validate the model using diagnostic techniques and see what variables are significantly associated with the outcome
- III. Interpret the findings (hazard ratios, and confidence intervals etc).
- IV. Consider limitations of the modeling approach and data collection that may be improved in future studies with similar aims

Overview of Study Design

Data Management

Original Data

Self reported risk factors on 20,822 young drivers between the ages of 17 and 24 were collected via the on-line DRIVE Cohort Study¹ during 2003 and 2004. Hypothesized risk factors considered to contribute to a young driver's propensity to crash have been formulated into a series of self reported responses to numerous questions relating to the individual's opinions, life experience, driver training and experience.

I received the data as a SAS dataset with all the original responses from the participants together with data collected on the participants based on police reported crashes. Specifically this related to the date of each crash, and the number of people injured directly related to each crash. A number of additional variables coded by members of The George Institute were included in the dataset, based on previously used psychological scales and rankings. Most of the variables ultimately used in the analysis were manipulated into these previously used scales and rankings.

Data Manipulation

Time to Crash Variable

The nature of the time to crash variable was modeled on two time horizons:

- a) The number of days between obtaining a provisional licence and first police reported vehicle crash, and
- b) The number of driving hours based on reported weekly driving hours multiplied by the number of weeks between obtaining a provisional licence and first police reported vehicle crash.

Both methods were implemented in the survival model to assess relevance. There was an expectation that the number of days may produce inconsistent results as each driver will have different access to a vehicle and different driving needs. As a consequence the total driving time between two drivers over a 6 month period may

be considerably different. The use of number of driving hours on the other hand will allow for the adjustment of total driving time between p-plate and vehicle crash to potentially make for more consistent comparisons.

The majority of the variables were initially coded into categorical variables due to the skewed nature of some of the outcomes, such as the range of reported hours under driving instruction. The re-categorisations ensured similar numbers of individuals and reported crashes into each category. The form of the independent variables that were initially used in this study and the manner in which they were manipulated from their original source is detailed as follows:

Gender

Gender remained in its original form (male / female).

Age

Based on age at the date each individual obtained their provisional licence. The range of ages was between 17 and 24 years old.

Driver Training and Experience

Based on the self reported number of hours of driver training, this was reported separately, based on the total hours spent under professional supervision (ie. licenced driving instructor) and total hours learning to drive with a non-professional supervisor (eg. parent, relative, or friend). Due to the range of self reported driving hours, these two variables were then re-categorised from continuous into discrete using the following categories:

Professional supervision		Non-professional supervision	
Continuous	Discrete	Continuous	Discrete
0 hours	Category 0	0 to 39 hours	Category 0
>0 to 4 hours	Category 1	40 to 49 hours	Category 1
>4 to 8 hours	Category 2	50 to 59 hours	Category 2
> 8 hours	Category 3	60 to 69 hours	Category 3
		> 69 hours	Category 4

The number of hours driving a motor vehicle prior to obtaining a learner's permit was also collected and categorised as a dichotomous variable for either never

driving prior to obtaining a learners permit or having driven at least once prior to obtain the permit.

Supervision on L Plate

This variable identified whether an individual drove a vehicle while on their learners permit without supervision. This variable was categorised as 0 if all driving was undertaken with supervision and 1 if some driving while on a learners permit was undertaken without supervision.

Time on L Plate

Time on learner's permit is a continuous variable based on the number of years the learner's permit was held. This variable was also re-categorised into an ordinal variable as follows:

Continuous	Discrete
< 1 year	Category 1
1 to 1.5 years	Category 2
> 1.5 years	Category 2

Number of Attempts at Provisional Licence

This variable is based on the number of tests taken to pass the provisional drivers test.

Country of Birth

This variable is based on the driver's country of birth. Due to the low numbers of drivers reported in a number of country categories, this variable was re-categorised as follows:

Region	Re-categorised
Australia	Category 1
NZ / UK	Category 2
Other Europe	Category 3
Asia	Category 4
Other	Category 5

Remoteness of Residence

Remote area is based on the driver's remoteness of residence at the time of obtaining their provision licence, as indicated by their postcode. Following

guidelines issued by the Australia Standard Geographic Classification, the residential postcodes of drivers were grouped into 3 levels – urban (metropolitan cities), regional (country towns and surrounds) and rural (including remote) areas - which indicate the approximate distances to public services.²⁸

*AUDIT-C*¹⁷

AUDIT-C is an industry standard screen used to help identify patients that may have a dependence on alcohol and potentially to assess whether an individual is at risk of an alcohol related disorder. The scale ranges from 0 to 12 and is based on three questions relating to alcohol use. The higher the reported number the greater the potential risk of alcohol dependence.

The scale used in this sub-study was:

Scale	Re-categorised
0 to 6	Category 1
7 to 12	Category 2

*Kessler 10*¹⁸

The Kessler 10 is based on 10 self reported questions relating to an individual's psychological state originally developed in 1992. Questions relate to an individual's nervousness, agitation and level of depression.

Kessler 10 scores range from 10 to 50, with the higher scores indicating a greater potential risk of psychological distress over the prior month.

The scale used in this sub-study was:

Scale	Re-categorised
10 to 15	Category 0-mild
16 to 21	Category 1-low
22 to 29	Category 2-moderate
30 to 50	Category 3-severe

*SEIFA Indexes*¹⁹

The SEIFA Indexes are a series of four socio-economic indexes derived from the 2001 Australian Bureau of Statistics Census data. The indexes are rank order variables and seek to identify a difference in socio-economic conditions based on geography. The four indexes are:

Disadvantage, Advantage/Disadvantage, Economic Resources, Education and Occupation.

For each of the indexes the variables were separated into equally numbered quartiles.

Number of Police Reported Offences

This variable is based on the actual number of offences recorded by police between obtaining a provisional licence and the first police reported crash. In this variable it is possible to record more than one offence at the same time (eg. speeding and driving through a red light). Due to the range of offence numbers and the small number of individuals that recorded as many as 23 traffic offences, this variable was re-categorised as:

Offence Number	Re-categorised
0 offences	Category 0
1 offence	Category 1
2 offences	Category 2
3 offences	Category 3
4 or more offences	Category 4

Sensation Seeking Score

This variable is based on the Zuckerman Sensation Seeking Scale which is a series of 19 true/false questions relating to the individual's likelihood of undertaking behaviours that may be considered less safe or planned. Areas such as impulsiveness and unpredictability are covered in this variable.

The range of outcomes in this variable is zero to 19, with the higher range indicating more sensation seeking behaviours. The categories were based on obtaining approximately equal sized groups.

Sensation Score	Re-categorised
0-4	Category 0
5-8	Category 1
9-19	Category 2

Risk Perception Score

This variable is based on a series of 10 questions relating to an individual's perception of how safe they believe particular behaviours are while driving. The

behaviours include driving 70km/h in a 60km/h zone, driving with a blood alcohol limit slightly over the legal limit, driving while talking on a mobile phone and driving through a red light. Each question includes a response scale from 1 to 4, for “always safe” to “rarely safe”. This was recorded at the time of survey (post obtaining provisional licence) and could potentially lead to slightly bias results as the time taken between obtaining a provision licence and undertaking the survey will be different for each individual. This may result in an individual’s perception of certain driving behaviours altering over this period.

The range of outcomes for this variable is from zero to 30, with the higher range indicating perceptions that these behaviours are safer. Similar to the sensation seeking variable the cutoffs for the three categories were based on obtaining similar sized groups. The variable was re-categorised into the following ranges:

Risk Perception Score	Re-categorised
0-5	Category 0
6-8	Category 1
9-30	Category 2

Risk Taking Behaviours

This variable is based on a series of 14 questions relating to the individual’s actual driving experience and ranking the occurrence of these behaviours on a scale of 1 to 4 from “never” to “very often”. Behaviours, such as driving 70km/h in a 60km/h zone, drag racing, driving while talking on a mobile phone, making rude gestures at other drivers are included in this variable. This was recorded at the time of survey (post obtaining provisional licence) and could potentially lead to slightly bias results based on the individual’s driving experience prior to undertaking this survey.

The range of possible outcomes for this variable is from zero to 56, with the higher range indicating an individual who undertakes more risk taking behaviours. The categories were determined based on approximately equal frequency counts. The variable was re-categorised into the following ranges:

Risk Taking Score	Re-categorised
0-8	Category 0
9-14	Category 1
15-56	Category 2

Modeling Approach

The majority of the potential predictors had to be re-categorised due to the skewed nature of the responses to the variables and the small number of reported crashes in the dataset. While the original dataset consisted of 20,255 observations, only 1,499 individuals recorded a crash during the observation period. A number of variables were based on derived scores (eg. risk taking score from 0 – 56 based on 14 separate questions), which reduced the number of reported crashes for each level within the variable. These variables were grouped to allow sufficient observations in each category and to make interpretation of the final result easier to explain.

Kaplan-Meier Curves

The initial analysis involved producing Kaplan-Meier curves for each variable and assessing whether there was a discernible difference between the curves. The Kaplan-Meier survivorship function at time t can be formulated as follows:

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i}$$
, where n is the total number of independent observations and the number of individuals at risk of crashing at time $t_{(i)}$ is denoted as n_i and the observed number of crashes at time $t_{(i)}$ is denoted as d_i . It provides a non-parametric estimate of the survival curve estimate (i.e. one minus the cumulative probability of having a crash). The method can account for right censoring present in the data, meaning each individual was followed during the observation period until their first police reported crash or the observation period finished.

Cox Proportional Hazards Model

A standard way to analyse survival data is to use a Cox Proportional Hazard model. The Cox model does not assume a fully parametric distribution for failure time, but relies instead on a semi-parametric formulation for the instantaneous risk of failure or hazard at a given time. Specifically, the hazard function²⁰ is defined as:

$h(t, x, \beta) = h_0(t)e^{x\beta}$ where x represents the covariates in the model. $h_0(t)$ models

how the hazard function changes with time, while $e^{x\beta}$ specifies the effect of covariates. This formulation allows for an easy assessment of the effect of a particular predictor x (e.g. gender) on survival through the hazard ratio, i.e. when x changes from x_0 to x_1 the hazard is multiplied by $HR(x_1, x_0) = e^{\beta(x_1 - x_0)}$, a ratio that does not depend on t .

For this sub-study the main objective is the comparison of crash times between individuals with different behaviours or characteristics and providing further insight on which characteristics are related to crashing early. In this context the hazard function is extremely useful as we can assess the time until first crash for an individual who, for instance, obtained their provisional licence at 18 years old relative to someone who obtained their licence at 17 years of age while adjusting for other potential predictors such as gender, risk taking behaviors etc. Generally speaking, the use of a Cox model allows an easy comparison of the differences between sub-groups and can accommodate a large number of combinations of variables and interactions.

Stratified Proportional Cox Hazard Model

As the proportional hazard assumption underlying the Cox is relatively stringent, a stratified Cox model can be considered in the analysis. The stratified model addresses the issue of different rates of change in a survival model within categories of the same variable. The mathematics of the stratified model is otherwise the same as described above however the stratified model separates the outcome between the predetermined categories within the variable. This can be used to relax the strict proportional hazard assumption of the Cox model, in particular when the baseline hazard rate does not appear to be the same across categories. In this study the age variable was used in this context.

Modeling Strategy

The modeling strategy was undertaken in a number of stages. This was initially carried-out through an exploratory analysis using the Kaplan-Meier and Nelson-Aalen cumulative hazard curves. The inclusion of previously well known scales

such as the Kessler 10 mental health index and AUDIT-C scale for alcohol use were also included. The exploratory analysis provided some insight into the appropriate form and significance of each variable via the log-rank test.

A univariate analysis using the Cox proportional hazards model was also performed to obtain unadjusted hazard ratios and 95% confidence intervals for all potential predictors. Discussion was held at this point with my project supervisors and other staff at The George Institute on the appropriateness of these results to ensure they were intuitive and potentially supported previous literature. Considerable checking was undertaken to ensure the variables had been input and interpreted correctly.

The multivariate analysis was undertaken using a manual elimination process (backward procedure). Each variable with global significance in the univariate analysis was included in the multivariate Cox proportional hazards model. Manual elimination occurred in the multivariate model to eliminate all categories of a particular predictor when the corresponding likelihood ratio (LRT) test was not significant and stopped when remaining independent variables were all significant (assessed by the global test involving all categories of a particular predictor). The overall process was summarised in a table displaying the change in minus twice the LRT statistic, the degree of freedom involved and the p-value.

Once the covariates were identified the proportional hazard (PH) assumption was systematically checked. Age categories clearly violated this assumption (p-value < 0.01) and as no clear pattern emerges from the residual plots the model was further stratified by age categories to overcome the problem.

Model Validation

Model diagnostics were undertaken for the PH model for each of the selected variables and the overall model.

Schoenfeld residuals were calculated for each observation and then used to assess the proportional hazards assumption. This involves fitting a smooth function

of time to the residuals and test whether there is a relationship (non-zero slope). A non-zero slope suggests the categories within a variable are not consistent over time and would therefore violate the general assumption for a Cox PH model.

Martingale residuals and Cox-Snell residuals were also calculated to make an assessment of the overall goodness of fit for the final model. A 45° linear slope in the plot of the Cox-Snell residuals versus the Nelson-Allen cumulative hazard function usually indicates a reasonable fit for the overall model. The result of the test for proportional hazards using the Schoenfeld residuals for overall goodness of fit are shown in the results section

Specific problems linked to outcome data

The use of hours to first crash as an outcome generated unexpected problems. The distribution of average weekly reported driving hours is shown in Table 5 and displays a wide range of values. Clearly some drivers reported values that are not possible or unrealistic. There were 574 individuals (excluded from the analysis) who reported average weekly driving hours above 50 or had not reported an average driving time. Average driving time above 50 hours seems rather unlikely. For example, a truck driver may work an average 40 hours, plus additional driving on the weekend may approach 50 hours of driving in an average week. Values above $7 \times 24 = 168$ hours are simply impossible but we did observe some of them in the data, making us question the validity of this data. Additional to this is the 395 missing observations for average weekly driving times and the 444 individuals who recorded zero average weekly driving times. It is possible that some of the individuals who reported zero average weekly hours may be accurate however it seems improbable that all of these are correct as 30 of these individuals recorded a crash. This also equated to 56 recorded crashes omitted from this analysis. This raises suspicion over the reliability of the number of driving hours until crash as a valid outcome but we did not make any further attempt to correct it any further.

Table 5. Distribution of Reported Average Weekly Driving Hours

Hours	0	1	2	3	4	5	10	15	20	50	>50	Miss -ing	Total
Frequency	444	1,319	2,149	2,105	1,607	2,676	5,582	1,756	1,106	937	179	395	20,255
Crashes	30	57	111	106	100	183	469	178	114	95	27	29	1,499

In the hours to first crash analysis 1,114 subjects were therefore excluded. These were the observations with missing, zero and greater than 50 hours a week for average weekly driving times as well as 96 other observations with missing values recorded in the statistically significant variables. As a result 98 individuals who recorded a crash were excluded from this analysis. The mean average weekly driving hours of the modeled observations was 8.8 hours (females 8.3 hours, males 9.3 hours)

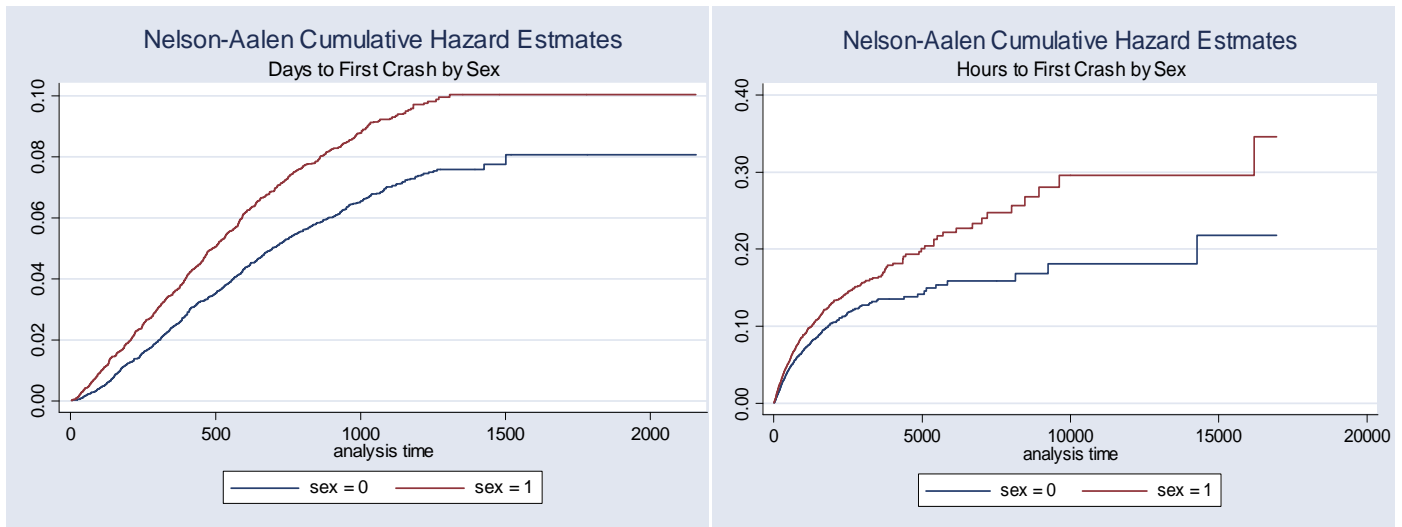
Results

The process of undertaking a univariate analysis of each variable was completed to provide an indication of the potential variables that showed some statistically significant differences in relation to time (days and hours) until first crash. The results of the univariate analysis are shown in Table 6.

Univariate Analysis

Kaplan-Meier survivorship estimator curves and Nelson-Aalen cumulative hazard curves were plotted against each categorical variable as an initial exploratory analysis to obtain a feel for the data in both the days to crash and hours to crash analysis. Additionally a series of univariate Cox models were calculated for each variable as an additional exploratory analysis to assess whether the variable was indicating a level of significance to early crash. In this exploratory exercise the analysis is unadjusted for any other effect; therefore other variables may be confounding these results; however this is acknowledged during this initial analysis. The Nelson-Aalen cumulative hazard curves for both models separated by gender are displayed in Figure 4.

Figure 4. Nelson-Aalen cumulative hazard curves of time until first police report crash (female = 0, male = 1)



The cumulative hazard curves shown above indicate a likely difference between the likelihood of males crashing than females. Similar differences in cumulative hazard curves were observed in the age and remoteness of residence variables.

The Kaplan-Meier and Nelson-Aalen estimators are usually supplemented by a log-rank test to assess whether survival curves are significantly different. Such a test was incorporated into the exploratory analysis to provide some additional guidance for the significant variables. This was evident in the assessment of significance for gender (hours: $\chi^2_{(1)}$ 22.29 p-value <0.01; days: $\chi^2_{(1)}$ 31.60 p-value <0.01). The results of the univariate analysis are displayed in Table 6 for independent variables defined above. A number of the potential predictors were shown to be statistically significant in their categorical form: *gender*, *age*, *remoteness of residence*, *risk taking score*, *risk perception score*, *sensation seeking score*, *time on learner's permit*, *non-professional driving instruction*, *driven before obtaining learner's permit*, *driven without supervision on learner's permit*, *previous crash* and *number of traffic offences*.

Table 6. Univariate analysis of each independent variable using Cox Proportional Hazards Model

Covariates		Days to First Crash		Hours to First Crash	
		Hazard Ratio (95% CI)	P-value (global test)	Hazard Ratio (95% CI)	P-value (global test)
Gender	Female	1 (referent)		1 (referent)	
	Male	1.34 (1.21 – 1.48)	<0.01	1.27 (1.14 – 1.41)	<0.01
	Global test		(<0.01)		(<0.01)
Remote area	Urban	1 (referent)		1 (referent)	
	Inner Regional	0.79 (0.70 – 0.91)	0.01	0.79 (0.69 – 0.91)	<0.01
	Rural	0.61 (0.45 – 0.82)	0.01	0.64 (0.46 – 0.88)	<0.01
	Global test		(<0.01)		(<0.01)
Country of Birth	Australia	1 (referent)		1 (referent)	
	NZ / UK	0.81 (0.53 – 1.23)	0.32	0.71 (0.45 – 1.11)	0.14
	Other Europe	0.77 (0.19 – 3.06)	0.71	0.95 (0.24 – 3.79)	0.94
	Asia	0.49 (0.36 – 0.67)	<0.01	0.60 (0.44 – 0.83)	<0.01
	Other	0.80 (0.65 – 0.99)	0.04	0.87 (0.70 – 1.08)	0.21
	Global test		(<0.01)		(0.01)
AUDIT C Score	0 – 6	1 (referent)		1 (referent)	
	7 - 12	1.34 (1.17 – 1.54)	<0.01	1.11 (0.96 – 1.27)	0.17
	Global test		(<0.01)		(0.17)
Kessler 10 Score	10 – 15	1 (referent)		1 (referent)	
	16 – 21	0.90 (0.79 – 1.02)	0.11	0.92 (0.81 – 1.05)	0.23
	22 – 29	1.00 (0.87 – 1.15)	0.99	1.02 (0.88 – 1.17)	0.83
	30 – 50	0.93 (0.75 – 1.15)	0.49	0.90 (0.72 – 1.12)	0.33
	Global test		(0.31)		(0.40)
Risk Taking Score	0 – 8	1 (referent)		1 (referent)	
	9 – 14	1.26 (1.10 – 1.45)	<0.01	1.09 (0.95 – 1.26)	0.22
	15 - 56	1.77 (1.55 – 2.01)	<0.01	1.30 (1.14 – 1.48)	<0.01
	Global test		(<0.01)		(<0.01)
Risk Perception Score	0 – 5	1 (referent)		1 (referent)	
	6 – 8	1.05 (0.91 – 1.21)	0.50	1.06 (0.92 – 1.23)	0.41
	9 - 30	1.41 (1.24 – 1.60)	<0.01	1.32 (1.16 – 1.51)	<0.01
	Global test		(<0.01)		(<0.01)
Sensation Seeking Score	0 – 4	1 (referent)		1 (referent)	
	5 – 8	1.17 (1.01 – 1.34)	0.03	1.12 (0.97 – 1.29)	0.11
	9 - 19	1.60 (1.40 – 1.82)	<0.01	1.35 (1.18 – 1.54)	<0.01
	Global test		(<0.01)		(<0.01)
SEIFA Disadvantage Rank	Top Quartile	1 (referent)		1 (referent)	
	Second Quartile	0.96 (0.83 – 1.11)	0.58	0.88 (0.76 – 1.02)	0.08
	Third Quartile	0.90 (0.78 – 1.04)	0.18	0.77 (0.66 – 0.90)	0.01
	Bottom Quartile	1.00 (0.87 – 1.15)	0.98	0.87 (0.76 – 1.01)	0.07
	Global test		(0.47)		(0.01)
SEIFA Advantage / Disadvantage Rank	Top Quartile	1 (referent)		1 (referent)	
	Second Quartile	1.10 (0.95 – 1.27)	0.19	0.98 (0.85 – 1.13)	0.79
	Third Quartile	0.98 (0.85 – 1.14)	0.84	0.83 (0.71 – 0.97)	0.2
	Bottom Quartile	1.02 (0.88 – 1.18)	0.81	0.90 (0.77 – 1.05)	0.18

	Global test		(0.44)		(0.06)
SEIFA Economic Resource Rank	Top Quartile	1 (referent)		1 (referent)	
	Second Quartile	1.03 (0.89 – 1.19)	0.71	0.93 (0.80 – 1.08)	0.33
	Third Quartile	1.02 (0.88 – 1.17)	0.80	0.87 (0.75 – 1.01)	0.07
	Bottom Quartile	0.87 (0.75 – 1.00)	0.06	0.78 (0.67 – 0.90)	<0.01
	Global test		(0.07)		(0.01)
SEIFA Education/Occupation Rank	Top Quartile	1 (referent)		1 (referent)	
	Second Quartile	1.01 (0.87 – 1.17)	0.86	0.91 (0.78 – 1.05)	0.20
	Third Quartile	1.07 (0.93 – 1.24)	0.34	0.91 (0.78 – 1.05)	0.20
	Bottom Quartile	1.10 (0.95 – 1.27)	0.20	0.94 (0.81 – 1.09)	0.38
	Global test		(0.52)		(0.54)
Number of Provisional Test Attempts	0	1 (referent)		1 (referent)	
	1	1.09 (0.96 – 1.23)	0.17	1.09 (0.97 – 1.24)	0.15
	2 or more	1.16 (0.99 – 1.36)	0.07	1.11 (0.94 – 1.31)	0.20
	Global test		(0.12)		(0.22)
Time on Learner's Permit	< 1 year	1 (referent)		1 (referent)	
	1 – 1.5 years	0.90 (0.80 – 1.01)	0.08	0.98 (0.87 – 1.10)	0.72
	> 1.5 years	0.77 (0.68 – 0.88)	<0.01	0.86 (0.75 – 0.98)	0.03
	Global test		(<0.01)		(0.07)
Professional Driving Instruction	0 hours	1 (referent)		1 (referent)	
	1 – 4 hours	0.98 (0.84 – 1.15)	0.82	1.01 (0.86 – 1.18)	0.95
	5 – 8 hours	1.13 (0.96 – 1.33)	0.14	1.16 (0.98 – 1.36)	0.09
	> 8 hours	0.96 (0.82 – 1.12)	0.61	0.96 (0.82 – 1.12)	0.61
	Global test		(0.13)		(0.08)
Non-professional Driving Instruction	0 – 39 hours	1 (referent)		1 (referent)	
	40 – 49 hours	1.08 (0.90 – 1.29)	0.43	1.11 (0.92 – 1.34)	0.26
	50 – 59 hours	1.09 (0.93 – 1.28)	0.27	1.15 (0.98 – 1.36)	0.09
	60 – 69 hours	1.20 (1.00 – 1.43)	0.04	1.24 (1.04 – 1.48)	0.02
	> 69 hours	1.23 (1.05 – 1.44)	0.01	1.14 (0.96 – 1.34)	0.13
	Global test		(0.08)		(0.21)
Age Group	17 years	1 (referent)		1 (referent)	
	18 years	0.83 (0.73 – 0.94)	<0.01	0.79 (0.69 – 0.89)	<0.01
	19 years	0.72 (0.60 – 0.86)	<0.01	0.69 (0.58 – 0.83)	<0.01
	20 years	0.84 (0.66 – 1.05)	0.13	0.83 (0.66 – 1.05)	0.12
	21 years	0.61 (0.43 – 0.85)	<0.01	0.59 (0.42 – 0.83)	<0.01
	22 years	0.58 (0.38 – 0.89)	0.01	0.59 (0.39 – 0.91)	0.02
	23 years	0.43 (0.24 – 0.76)	<0.01	0.37 (0.20 – 0.68)	<0.01
	24 years	0.46 (0.25 – 0.86)	0.02	0.46 (0.25 – 0.86)	0.02
	Global test		(<0.01)		(<0.01)
Driven Before Learner's Licence	No	1 (referent)		1 (referent)	
	Yes	1.27 (1.13 – 1.42)	<0.01	1.17 (1.04 – 1.32)	<0.01
	Global test		(<0.01)		(0.01)
Driven Without Supervision	No	1 (referent)		1 (referent)	
	Yes	1.23 (1.10 – 1.38)	<0.01	1.10 (0.98 – 1.24)	0.09
	Global test		(<0.01)		(0.09)
Prior Crash	No	1 (referent)		1 (referent)	
	Yes	13.18 (1.86 – 93.66)	0.01	4.94 (0.69 – 35.30)	0.11
	Global test		(0.07)		(0.21)

	0	1 (referent)		1 (referent)	
Number of Traffic Offences	1	4.21 (3.67 – 4.82)	<0.01	3.70 (3.21 – 4.26)	<0.01
	2	5.39 (4.60 – 6.31)	<0.01	4.27 (3.63 – 5.03)	<0.01
	3	7.84 (6.53 – 9.41)	<0.01	5.79 (4.79 – 6.99)	<0.01
	4 or more	7.44 (6.17 – 8.98)	<0.01	5.00 (4.10 – 6.10)	<0.01
	Global test		(<0.01)		(<0.01)

Highlighted rows indicate variables with statistically significant effects at a 5% level of significance

Although the conclusions at this stage are still tentative as the analysis is unadjusted we do observe similarities between the days to crash analysis and the hours to crash analysis. The difference in gender (days: HR 1.34, CI 1.21-1.48; hours: HR 1.27, CI 1.14-1.41) and remoteness of residence (days: inner regional HR 0.79, CI 0.70-0.91, rural HR 0.61, CI 0.45-0.82; hours: inner regional HR 0.79, CI 0.69-0.91, rural HR 0.64, CI 0.46-0.88) were statistically significant for both outcomes. They indicate that males have a 27% to 34% higher probability of crashing compared to females and an individual living in an urban area has a 21% higher probability of crashing compared to an individual living in an inner regional and 36% to 39% higher probability of crashing than an individual living in a rural area. The age variable indicated that an individual who obtained their provisional licence between the ages of 18 to 24 had 17% to 63% less chance of crashing compared to an individual who was 17 years old at the time of obtaining their provisional licence.

Despite its limitations, the preliminary analysis has helped identify important predictors and discover that linearity on the log-hazard scale was not necessarily met for variables like age or non-professional driving instruction (5 categories) and raises some questions over the form of these variables in the multivariate model.

Multivariate Analysis

A multivariate analysis was carried out next to account for potential confounding,

Days to First Crash

In the multivariate analysis utilizing the days until first police reported crash the initial variables incorporated into the model were based on the univariate analysis.

All variables in the univariate analysis with a p-value less than 0.05 for the global test were included in the full model.

The sensation seeking and risk perception variables were excluded from the multivariate analysis as these were perceived to be measuring the same underlying response as risk taking behaviour. Additionally, risk perception and risk taking related specifically to driving, therefore risk perception was considered a “softer” variable. As these were previously considered less predictive of crashes²¹ than specific driving behaviours they were excluded.

The determination of the full model was undertaken via a manual elimination procedure as explained earlier, the detail being given in table 7.

The final model included only the following variables: *gender, remoteness of residence, country of birth, risk taking behaviours, driven before learners permit and age categories used as strata*. A preliminary analysis fitting age categories as covariates in the model indicated a violation of the proportional hazards assumptions (test for PH assumptions: age 18 p-value <0.01, age 19 p-value <0.01, age 20 p-value <0.01, global test p-value <0.01). No indication was found in the martingale residuals on how better to accommodate the age effect than stratifying.

Some of the categories within the country of birth variable showed insignificance. As we were interested in the overall strength of association of each variable rather than the effect of one category compared to the referent category the indicator variables were left in the model in their current form. The hazard ratios, confidence intervals and p-values of the final model are shown in Table 8. The fit of this model is displayed in Table 9.

Table 7. Likelihood Ratio Test from Variable Elimination

Model Variables	Removed Variable	2*Log Likelihood	Degrees of freedom	p-value
Gender, Remote Area, Country of Birth, Audit Cutoff, Risk Taking Score, SEIFA Economic Resource Rank, Time on L Plates, Non-professional Driving Instruction, Age, Driven Before L Plate, Driven without Supervision		26,891.8		
Gender, Remote Area, Country of Birth, Audit Cutoff, Risk Taking Score, SEIFA Economic Resource Rank, Time on L Plates, Non-professional Driving Instruction, Age, Driven Before L Plate	Driven without Supervision	26,892.2	1	0.53
Gender, Remote Area, Country of Birth, Risk Taking Score, SEIFA Economic Resource Rank, Time on L Plates, Non-professional Driving Instruction, Age, Driven Before L Plate	Audit Cutoff	26,894.2	1	0.16
Gender, Remote Area, Country of Birth, Risk Taking Score, SEIFA Economic Resource Rank, Non-professional Driving Instruction, Age, Driven Before L Plate	Time on L Plates	26,895.8	2	0.45
Gender, Remote Area, Country of Birth, Risk Taking Score, Non-professional Driving Instruction, Age, Driven Before L Plate	SEIFA Economic Resource Rank	26,897.8	3	0.57
Gender, Remote Area, Country of Birth, Risk Taking Score, Age, Driven Before L Plate	Non-professional Driving Instruction	26,902.8	4	0.29

Table 8. Final Model Days to First Crash – Cox Proportional Hazard Model

Covariate	Hazard Ratio	P-value
Gender		
Female	1 (referent)	
Male	1.17 (1.05 – 1.30)	0.01
Driven Before Learner's Permit		
Not driven before L-plate	1 (referent)	
Driven before L-plate	1.14 (1.01 – 1.28)	0.04
Remoteness of Residence		
Urban	1 (referent)	
Inner Regional	0.74 (0.64 – 0.85)	<0.01
Rural	0.55 (0.40 – 0.76)	<0.01
Country of Birth		
Australia	1 (referent)	
NZ / United Kingdom	0.72 (0.46 – 1.14)	0.16
Other Europe	0.79 (0.20 – 3.18)	0.75
Asia	0.53 (0.38 – 0.72)	<0.01
Other	0.75 (0.59 – 0.94)	0.01
Risk Taking Behaviours		
Category – 1 (low)	1 (referent)	
Category – 2 (medium)	1.21 (1.05 – 1.39)	0.01
Category – 3 (high)	1.57 (1.37 – 1.80)	<0.01
Stratified by age		

Validation

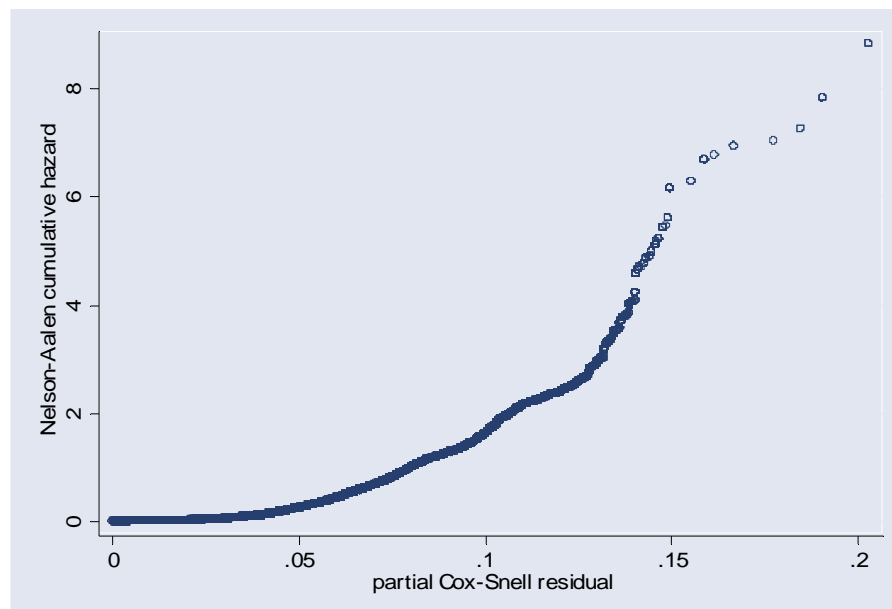
The tests for the PH assumption for all individual covariates in the final model are shown in Table 9 and no longer indicate any violation.

Table 9. Final Model - Test of Proportional Hazards Assumption

Variable	Rho	χ^2	df	Prob> χ^2
Driven Before L Plate	-0.03	1.60	1	0.21
Remoteness of residence – inner regional	<0.01	0.00	1	0.99
Remoteness of residence – rural	-0.02	0.50	1	0.48
Country of Birth – NZ / UK	0.01	0.24	1	0.63
Country of Birth – Other Europe	0.03	1.32	1	0.25
Country of Birth – Asia	0.03	0.92	1	0.34
Country of Birth - Other	-0.01	0.25	1	0.62
Risk Taking Behaviour – medium	-0.01	0.17	1	0.68
Risk Taking Behaviour – high	<-0.01	0.01	1	0.92
Global test		6.40	10	0.78

The plot of the Cox-Snell residuals shown in figure 5 indicates that the fitted model is not completely appropriate as it substantially differs from a 45% line that usually corresponds to a good fit, however we were not able to bring further improvements using standard techniques

Figure 5. Final model – Cox Snell Residuals for Test of Goodness of Fit



Interpretation

The results of the final model using days until first crash indicated that males had a 17% higher probability (HR: 1.17, CI 1.05-1.30, p-value 0.01) of crashing compared to females. Individuals who had driven before obtaining their learners permit had a 14% (HR: 1.14, CI 1.01-1.28, p-value 0.04) increased probability of crashing compared to individuals who had not driven prior to obtaining their learners permit. An individual living in an inner regional or rural area had a 26% (HR: 0.74, CI 0.64-0.85, p-value <0.01) to 45% lower probability (HR: 0.55, CI 0.40-0.76, p-value <0.01) of crashing compared to an individual living in an urban area.

The model indicated that an individual who was born in Asia had a 47% lower probability (HR: 0.53, CI 0.38-0.72, p-value <0.01) of crashing than an individual born in Australia. While individuals born outside of New Zealand, United Kingdom, Europe or Asia had a 25% lower probability (HR: 0.75, CI 0.59-0.94, p-value 0.01) of crashing than individuals born in Australia.

This model indicated that an individual who undertakes risk taking behaviours while driving had an increased probability of crashing compared to individuals who undertook lower levels of risk taking behaviour while driving. This is evident from the 21% (HR: 1.21, CI 1.05-1.39, p-value 0.01) increase in medium risk takers and 57% (HR: 1.57, CI 1.37-1.80, p-value <0.01) increase in high risk takers of crashing compared to low risk takers.

The separate results of each stratum are not included in this analysis as the overall effect was of main interest rather than the specific hazard ratios under each age category. The results of this model should be viewed with some caution due to the lack of overall goodness of fit observed through the Cox-Snell goodness of fit plot. Using standard survival analysis methods did not allow for these violations to be remedied.

Hours to First Crash

The modeling of the number of driving hours to first crash was done using the same modeling process described previously for the days to first crash.

The determination of the full model was undertaken using the same selection process as for the days to first crash analysis. Manual elimination was again carried out sequentially by means of the likelihood ratio test involving all categories of a particular predictor – see Table 10 for details. The final model included only the following variables: *gender, remoteness of residence, country of birth, risk taking score and age.*

Again, the test of PH assumption for the *age* variable indicators were rejected (test for PH assumptions: age 18 p-value <0.01, age 22 p-value 0.03, global test p-value 0.13) even though the global test was above the 0.05 significance level. This problem was overcome in a similar manner as before, i.e. resorting to a stratified Cox model by age category.

The hazard ratios, confidence intervals and p-values of the final model are shown in Table 11. The fit of the model is shown in Table 12.

Table 10. Likelihood Ratio Test from Variable Elimination

Model Variables	Removed Variable	2*Log Likelihood	Degrees of freedom	p-value
Gender, Remote Area, Country of Birth, Risk Taking Score, SEIFA Economic Resource Rank, SEIFA Disadvantage Rank, Time on L Plates, Age, Driven Before L Plate		24,718.4		
Gender, Remote Area, Country of Birth, Risk Taking Score, SEIFA Economic Resource Rank, Time on L Plates, Age, Driven Before L Plate	SEIFA Disadvantage Rank	24,723.8	3	0.14
Gender, Remote Area, Country of Birth, Risk Taking Score, Time on L Plates, Age, Driven Before L Plate	SEIFA Economic Resource Rank	24,726.0	3	0.53
Gender, Remote Area, Country of Birth, Risk Taking Score, Age, Driven Before L Plate	Time on L Plates	24,729.4	2	0.18
Gender, Remote Area, Country of Birth, Risk Taking Score, Age	Driven Before L Plate	24,733.0	1	0.06

The Wald Test indicated that one variable (driven before L plate) was showing borderline significance. However, it was removed following the assessment of the LR test. Prior to its immediate removal the individual category p-value was 0.06.

Table 11. Final Model Hours to First Crash – Cox Proportional Hazard Model

Covariate	Hazard Ratio	P-value
Gender		
Female	1 (referent)	
Male	1.17 (1.05 – 1.31)	0.01
Remoteness of Residence		
Urban	1 (referent)	
Inner Regional	0.75 (0.65 – 0.86)	<0.01
Rural	0.60 (0.43 – 0.83)	<0.01
Country of Birth		
Australia	1 (referent)	
NZ / United Kingdom	0.66 (0.41 – 1.06)	0.09
Other Europe	0.99 (0.25 – 3.97)	0.99
Asia	0.65 (0.47 – 0.90)	0.01
Other	0.82 (0.45 – 1.04)	0.10
Risk Taking Behaviours		
Low – 1 (low)	1 (referent)	
Medium – 2 (medium)	1.06 (0.92 – 1.22)	0.45
High – 3 (high)	1.18 (1.03 – 1.35)	0.02
Stratified by age		

Validation

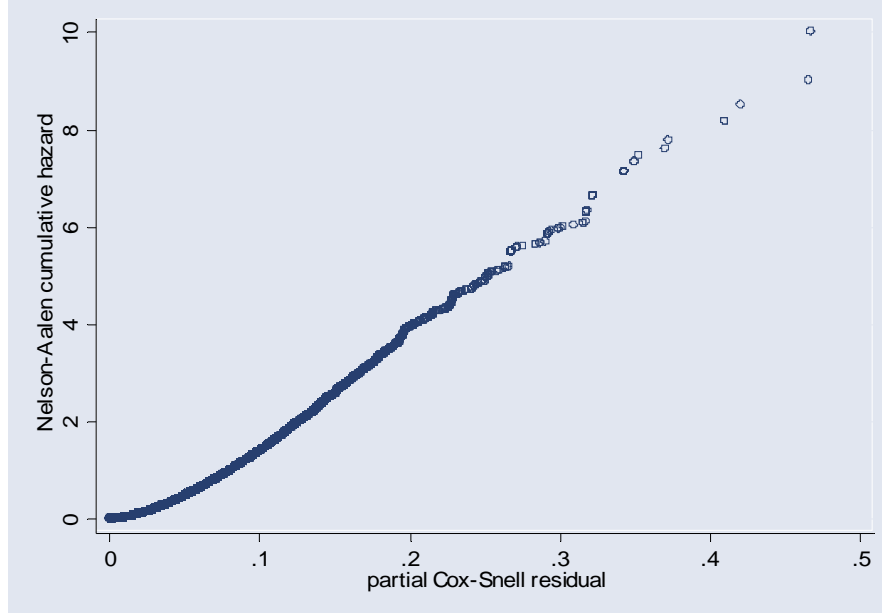
Model validation was undertaken using the Schoenfeld residuals to assess whether the independent variables were violating the proportional hazards assumption and the Cox-Snell residual plot to assess the overall goodness of fit of the final model. The Schoenfeld residuals in the final model are shown in Table 12 and indicate that each variable, and the overall model, was not violating the PH assumptions.

Table 12. Final Model - Test of Proportional Hazards Assumption

Variable	rho	χ^2	df	Prob> χ^2
Male	0.03	0.87	1	0.35
Remoteness of Residence – inner regional	-0.02	0.68	1	0.41
Remoteness of Residence – rural	0.03	0.84	1	0.36
Country of Birth – NZ / UK	-0.02	0.67	1	0.41
Country of Birth – Other Europe	0.01	0.09	1	0.76
Country of Birth - Asia	0.01	0.02	1	0.90
Country of Birth - Other	-0.01	0.12	1	0.73
Risk taking Behaviours - medium	0.02	0.50	1	0.48
Risk Taking Behaviours – high	0.04	2.44	1	0.12
Global test		6.58	9	0.68

The plot of the Cox-Snell residuals shown in figure 6 is reasonably good for this model as a 45% line is observed on that plot.

Figure 6. Cumulative hazard plot of the Cox-Snell Residuals



Interpretation

The results of the final model using hours until first crash indicated that males had a 17% higher probability (HR: 1.17, CI 1.05-1.31, p-value 0.01) of crashing compared to females. An individual living in an inner regional area had a 25% lower probability (HR: 0.75, CI 0.65-0.86, p-value <0.01) of crashing compared to an individual living in an urban area and a rural individual had a 40% lower probability (HR: 0.60, CI 0.43-0.83, p-value <0.01) of crashing compared to an individual living in an urban area. An individual who was born in Asia had a reduced probability of 35% (HR: 0.65, CI 0.47-0.90, p-value 0.01) of crashing compared to individuals born in Australia. Additionally individuals who undertook high risk taking behaviours had an 18% higher probability of crashing (HR: 1.18, CI 1.03-1.35, p-value 0.02) compared to those who undertook low risk taking behaviours. Overall we found good consistency between the two analyses in terms of selected predictors and direction of the associations. However, the strength of risk taking behavior was somehow dampened in hours to crash, while having

driven before permit seems to be a significant predictor in the final model for days until crash but not for number of driving hours until first police reported crash.

Discussion

This analysis is one small sub-study of the wider group of analyses undertaken on the DRIVE study data and therefore the outcome of this study and consideration of the significance and direction of the final variables in this model need to be viewed in the context of the broader study. The results of both the days until first crash and hours until first crash models provided consistent results. Both models result in significant age, gender, remoteness of residence, risk taking behaviours and country of birth variables. Similar hazard ratios were obtained in both models. The exception was the inclusion of the driven before learners permit in the days until first crash model although this had borderline significance in the hours to first crash model. The results of the significant variables are supported by existing literature which has shown similar factors to be significantly related to an increased risk of crash. The following discussion summarises the results in terms of data quality, the statistically significant variables, goodness of fit and limitations of the analysis.

Data Quality

Two main outcomes were considered due to the issue arising over the time interval. The date of provisional licence and the police reported crash were viewed as reliable, given they were obtained from the Roads and Traffic Authority, NSW (RTA) and Police records. However, when using the days between these dates as the time interval there was a potential to observe different driving experiences as each individual has different access to vehicles and drive for varied lengths of time over an average week.

The use of hours to first crash also has several limitations. The distribution of reported average weekly driving hours is likely to be biased due to the unreliability of the reported hours. Therefore we were obliged to exclude the data of drivers who reported more than 50 driving hours per week. Such reports are clearly unrealistic and the inclusion of observations near 50 hours per week of driving time may well be overstated. More generally, this alludes to the possibility of grossly

overstating the 'true' average weekly hours some individuals may drive in a given week. However offsetting some of these issues was that all drivers had to keep a log book of their driving hours while on their L-plates, therefore most would have a reasonable perception of the amount of time they had driven on an average week on their provisional licence. The number of driving hours assumes that the weekly driving hours remains constant over time. Given that some individuals were followed for a couple of years, this assumption is questionable.

There may not be a perfect solution to the issue of reliable time, however consideration in using the odometer reading may improve results as the time would then be measured in kilometres driven. This would be achievable for those individuals who have access to only one vehicle, however may be difficult for those people who either drive more than one vehicle or share a vehicle.

Statistically Significant Variables

The results of both final models (hours to crash and days to crash) produced very similar findings. The most significant variables were gender, remoteness of residence, country of birth, risk taking behaviours and age in both models with the same direction and similar hazard ratios. The consistency of these results provides reassurance for the significance of these variables and overall magnitude of the effect. The findings of gender and age are supported by existing literature where males are deemed more likely to be involved in a vehicle crash than females and younger drivers are more likely to crash than older drivers²².

The cause of increased probability of crashing from driving before obtaining a learners permit may be a result of over confidence caused by prior driving. For example an individual living in a rural area may have driven regularly in open spaces on a farm however this is quite different to driving on a road and within all the road rules and surrounding traffic. This finding must however be interpreted with care as having driven before obtaining a learners permit was a significant predictor for days until first crash but not for number of driving hours before first crash.

A more consistent result was observed in the remoteness of residence variable where the hazard ratios for individuals living in urban areas indicated either a 25% (hours to first crash) or 26% (days to first crash) increased probability of crashing compared to individuals living in an inner regional and 40% (hours to crash) or 45% (days to crash) compared to individuals in a rural area. This may be a result of greater congestion on urban roads resulting in more opportunities for a crash to occur. However the findings of increased rates of early crashing in people living in urban areas were slightly different to those reported in other literature¹⁶ where there was no reported statistical difference in time to crash between urban and rural drivers.

The increased probability of crashing for those individuals who undertook risk taking behaviours is supported by previous studies and showed the highest probability in crashing as the high risk takers had a 57% increase in probability compared to the lowest risk takers in the days to first crash model and 18% increase in probability in the hours to crash model. As this behaviour was directly related to habits while driving the direction and significance of this variable is intuitive.

Drivers born in Asia had lower probability of crashing and this would be interesting to follow-up in further studies to assess whether this difference is consistent with Asian drivers who remain in their home country or whether this is solely an issue for Asian born people who now reside in Australia. The time at which the individual moved to Australia may also be a factor as an individual who moved early in life may have a different probability of crashing compared to an individual who moved to Australia later in life. There is some support for the differences in probability to crash in existing literature as ethnicity has been found to be a significant factor in the risk of fatalities even after adjusting for socioeconomic status²³.

The variables; sensation seeking, risk perception and risk taking were considered to be largely measuring similar concepts in terms of propensity to take driving risks. As risk perception was based on perceptions rather than behaviours and sensation seeking included non-driving related items, risk taking behaviour was considered to

be more representative of actual behaviours while driving, which was of greatest interest. Both the sensation seeking and risk perception variables were excluded from the analysis.

One of the variables was assessed to be more of an outcome rather than an observable behaviour. This variable was the police reported offences prior to the first crash. One individual had reported 24 offences and the univariate analysis indicated that a high offence rate would greatly increase the probability of crashing. It was possible to receive multiple offences for one incident (eg, speeding and running a red light). This variable was not included in the analysis as it was seen as an outcome in itself.

The driven before learner's permit variable was removed in the manual elimination procedure for the hours to first crash model even though this variable showed borderline significance using both the Wald Test and LR test. This decision was based on a desire to ensure that each variable in the final model was truly statistically significant.

Goodness of Fit

There was not a perfect solution to the issue of goodness of fit particularly for the days until first crash model using the survival analysis techniques learned through the Master's degree program. Other techniques that are not based on the proportional hazard assumption may prove useful. One of these approaches is censored regression quantiles²⁴ that is readily available but requires the use of a specific R package. Another possible solution would be to include splines in the covariates that violated the PH assumption. This is outside the scope of this work but can be attempted in future research.

The reliance on Cox-Snell residuals to determine goodness of fit has been questioned by other authors²⁵ who found that they can be misleading in practice. Therefore the models have been left 'as is' after the elimination procedure was carried out and no indication of PH assumption was observed. Interestingly, a slight change in the stratification of the age variable (from individual age categories

to 17, 18, 19-20, 21-24 age groups) resulted in a material change in the Cox-Snell residuals plot with only minor changes in the hazard ratios and significance of each of the variables in the final models. This brings more support to the strength of the associations found in this analysis.

Two alternatives were available for the starting point of each observation and the time interval used. There is a time difference between an individual obtaining a provisional licence and then undertaking the DRIVE survey. An individual's behaviour may change during this period once the driver has spent more time driving and without a supervisory driver in the car. Initial analysis indicated that there was little difference in the outcome as the majority of participants submitted the survey within 5 to 6 months of obtaining their provisional licence and only 55 (3.6%) crashes occurred between the individual obtaining their provisional licence and undertaking the survey.

Other Limitations

The information collected in the DRIVE survey is based on individuals' recollection of their driving behaviours and experience as well as their experience in a number of other areas of their life. This highlights the potential reliability of some of the responses collected due to the difficulties with recall; for example the risk taking behaviour or sensation seeking variable. There is a possibility that the person's perception of how often they drive and undertake these perceived risky actions may be skewed. One driver may consider driving 70km/h in a 60km/h zone once a day as very frequent where as someone else may consider this infrequent. Additionally, the sensation seeking variable relates to other facets of the person's life (eg, unplanned trips, undertaking thrill seeking activities) that can also easily be skewed. The original questionnaire does not indicate what constitutes a number of the potential predictors such as an unplanned trip or thrill seeking behaviour therefore each person may view the extent to which they undertake these behaviours in different ways.

Approximately 7% of drivers in the study had a police reported crash, which means a large dataset is required to ensure sufficient crashes are recorded. This was met

in this study although there may be issues of reliability when a large portion of the data collected is based on respondent's opinions of their own behaviour. However the cost of collecting most of this data in a different format would likely be prohibitive and not feasible to continuously observe an individual's driving behaviour. The issue of reliability was addressed by using previously validated questionnaires. Additionally the large number of observations collected assists in reducing some of the potential bias.

Conclusion

This study considered a number of potential risk factors on the time until first police reported vehicle crash. Little previously published analysis on time until first crash analyses for this age group have been conducted therefore a large number of variables were considered in the exploratory analysis. While the overall approach resulted in models with non-statistically significant goodness of fit the results were intuitive and similar to the outcomes observed in existing literature. Specifically, this related to the increased probability of crashing in younger males as a result of increased levels of risk taking behaviour.

One of the main issues encountered in this analysis was the basis on which time was determined either through days or hours until first crash. However reassurance has been gained as the results consistently identified age, gender, remoteness of residence, risk taking behaviours and drivers born in Asia as predictors and displayed hazard ratios of similar magnitude. The magnitude of the hazard ratios in the risk taking behaviour variable was different with the two outcomes. Prior literature had indicated an impact of risk taking behaviours on early crashing and since this variable, as with the other self reported variables, are based on an individual's opinion, further work may need to be undertaken to provide a more consistent and accurate manner in determining these scales. As the PH assumption was not met for age and goodness of fit was questionable in the final models, more flexible methods that do not rely on the proportional hazard assumption especially for age may provide further insights on this data.

References

1. Ivers R.Q., Blows S.J., Stevenson M.R., Norton R. N., Williamson A., Eisenbruch M., Woodward M., Lam L., Palamara P., Wang J., A cohort study of 20 822 young drivers: the DRIVE study methods and population, *Injury Prevention* 2006;12:385–389
2. Özkan T., Lajunen T., Multidimensional Traffic Locus of Control Scale (T-LOC): factor structure and relationship to risky driving, *Personality and Individual Differences*, Volume 38, Issue 3, February 2005, Pages 533-545
3. Ulleberg P., Rundmo T., Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers, *Safety Science*, Volume 41, Issue 5, June 2003, Pages 427-443
4. Darlington R., Weinberg S., Walberg H., (1973). Canonical Variate Analysis and Related Techniques. *Review of Educational Research*, 453-454.
5. Hatcher L., *A Step by Step Approach to Using SAS System for Factor Analysis and Structural Equation Modeling*, 1996, SAS Institute Inc
6. Spector P., *Summated Rating Scale Construction: An Introduction*, 1992 SAGE University Paper
7. Shoukri, M.M., Edge V.L., 1996. *Statistical Methods for the Health Sciences*. CRC Press.
8. Moss S., Prosser H., Costello H., et al (1998) Reliability and validity of the PAS–ADD Checklist for detecting psychiatric disorders in adults with intellectual disability. *Journal of Intellectual Disability Research*, 42, 173 – 183
9. Ferguson S.A., Teoh E.R., McCartt A.T., Progress in teenage crash risk during the last decade. *Journal of Safety Research*, 2007. 38: 137-145.
10. Harrison J.E., Berry J.G., Serious injury due to transport accidents, Australia, 2003–04. Australian Institute for Health and Welfare AIHW cat. no. INJCAT 101. Canberra: AIHW & ATSB
11. Twisk D.A.M., Stacey C., Trends in young driver risk and countermeasures in European countries. *Journal of Safety Research*, 2007. 38: 245-257.
12. Islam S., Mannering F., Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence. *Journal of Safety Research*, 2006. 37: 267-276.
13. Cammisa M.X. Williams A.F. Ferguson S.A. 2000. Self-reported seat belt use in four countries: a telephone survey. *Journal of Crash Prevention and Injury Control*, 2 (2):103-110
14. Wedel M., Kamakura W., Factor Analysis with (Mixed) Observed and Latent Variables in the Exponential Family, *Psychometrika*, December 2001, vol.66, no. 4 pp.515-530
15. Braitman K., Kirley B., McCartt A., Chaudhary N., Crashes of Novice Teenage Drivers: Characteristics and Contributing Factors, *Journal of Safety Research* 39 (2008), pp.47-54
16. Stevenson M.R., Palamara P., Morrison D., Ryan A., Behavioural Factors as Predictors of Vehicle Crashes in Young Drivers, *Traffic Injury Prevention*, Vol. 2 Issue 4, 2001

17. Bush K., et al., The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch Intern Med*, 1998. 158: p. 1789-1795.
18. Andrews G., Slade T., 2001, Interpreting scores on the K10. *Australian and New Zealand Journal of Public Health*, 25, 494-497.
19. Trewin D., *Socio-Economic Indexes for Areas*, Australian Bureau of Statistics, 2003
20. Hosmer D., Lemeshow S., *Applied Survival Analysis*, 1999, John Wiley & sons, inc.
21. Ivers R., Senserrick T., Boufous S., Stevenson M., Chen H-y., Woodward M., Norton R., Novice drivers' risky driving behaviour, risk perception and crash risk: Findings from the DRIVE Study. *American Journal of Public Health*, in press
22. Massey D., Campbell K., *Traffic Accident Involvement Rates by Driver Age and Gender*, *Accident Analysis and Prevention*, Vol. 27 No. 1, 1995
23. Braver E., Braver E.R., Race, Hispanic origin, and socioeconomic status in relation to motor vehicle occupant death rates and risk factors among adults. *Accident Analysis & Prevention* 2003; 35:295-309
24. Koenker R., Hallock K., *Quantile Regression*, *Journal of Economic Perspectives*, Vol. 15 No. 4, 2001
25. Collett D., *Modelling Survival Data in Medical Research*, Chapman & Hall, 2nd ed, 2003
26. Gorsuch R., *Factor Analysis* 1983 Lawrence Erlbaum & Associates
27. Zuckermann M., Kuhlman D.M., Joireman J., et al. A comparison of the three_structural models for personality: the big three, the big five, and the alternative five. *J Pers Soc Psychol* 1993;65:757-68.
28. Australian Bureau of Statistics, *ASGC Remoteness Classification: Purpose and Use*, Census Paper (03/01), 2003
29. Knusel L., *Factor Analysis: Chisquare as Rotation Criterion*, University of Munich, Technical Report No. 040, 2008
30. Lawley D.N., Maxwell A.E., *Factor Analysis as a Statistical Method*. Second edition. 1971 Butterworths London.
31. Cureton E.E., Mulaik S.A., The Weighted Varimax Rotation and the Promax Rotation, University of Tennessee and Georgia Institute of Technology, *Psychometrika*, vol.40 no.2, June 1975

Appendix

SAS Code for Project A: Factor Analysis

Condensed initial factor analysis using orthogonal solution and 0.40 flag

```
proc factor data=dr.fa_data_3 simple method=prin priors=smc scree rotate=varimax round
flag=0.40;
  var sex1 age prof_hours unprof_hours prof_residential prof_major prof_open
  prof_raining prof_dark prof_gravel prof_traffic unprof_residential unprof_major
  unprof_open unprof_raining unprof_dark unprof_gravel unprof_traffic job_planning1
  think_before1 impulse1 seldom_plan_ahead1 exciting_experiences1
  complicated_job1 not_planned_trip1 new_situations1 thrill1 change_interests1
  frightening_things1 anything_once1 travelling1 crazy_things1 getting_lost1
  unpredictable_friends1 ca_new_things1 impulsive_person1 wild_parties1 thrill_driving
  risky_driving driving_70_in_60 burnouts someone_passing slower_drivers
  rude_gestures horn race no_seatbelt mobile_phone loud_music driving_passengers
  driving_sms tired_no_reason feel_nervous very_nervous hopeless restless
  very_restless depressed everything_effort very_sad worthless no_supervision
  driven_motorcycle1 alcohol_4_weeks marijuana_habbit marijuana_4_weeks
  licensing_rate general_rate self_harm1 auditc1 safe_70_in_60_1 safe_110_in_100_1
  safe_drink_1 safe_marijuana_1 safe_unmaintained_car_1 safe_redlight_1
  safe_mobile_phone_1 safe_passengers_1 safe_midnight_1 safe_sms_1;
run;
```

Calculate Cronbach Coefficient Alpha for each factor

```
proc corr data=dr.fa_data_3 alpha nomiss;
  var thrill_driving risky_driving driving_70_in_60 burnouts someone_passing
  slower_drivers rude_gestures horn race mobile_phone loud_music driving_sms;
run;
proc corr data=dr.fa_data_3 alpha nomiss;
  var tired_no_reason feel_nervous very_nervous hopeless restless very_restless
  depressed everything_effort very_sad worthless;
run;
proc corr data=dr.fa_data_3 alpha nomiss;
  var impulse1 exciting_experiences1 new_situations1 thrill1 frightening_things1
  crazy_things1 unpredictable_friends1 impulsive_person1 wild_parties1;
run;
proc corr data=dr.fa_data_3 alpha nomiss;
  var safe_70_in_60_1 safe_110_in_100_1 safe_drink_1 safe_marijuana_1
  safe_unmaintained_car_1 safe_redlight_1 safe_mobile_phone_1 safe_sms_1;
run;
proc corr data=dr.fa_data_3 alpha nomiss;
  var prof_hours prof_raining prof_dark prof_gravel prof_traffic;
run;
proc corr data=dr.fa_data_3 alpha nomiss;
  var unprof_hours unprof_raining unprof_dark unprof_gravel unprof_traffic;
run;
```

STATA Code for Project B: Survival Analysis

Generate crash variable

```
gen accdate1 = date(accdate, "dmy")
gen acc = 1 if accdate1 >= 0
replace acc = 0 if accdate == ""
```

Generate days and hours to crash variables for provisional date and survey date

```
gen regdate = date(dreg, "dmy", 2040)
gen transdate = date(transaction_date, "mdy", 2040)
gen timetocrash_reg = accdate1 - regdate
gen timetocrash_trans = accdate1 - transdate
format regdate %d
format accdate1 %d
format transdate %d
gen hourstocrash_reg = timetocrash_reg * avg_driving / 7
gen hourstocrash_trans = timetocrash_trans * avg_driving / 7
gen lastdate = d(31may2006)
gen totaldays = lastdate - transdate
gen hourstocrash_trans = (avg_driving / 7) * totaldays if hourstocrash_trans == .
stset hourstocrash_trans, id(l_no) failure(acc == 1) exit(acc==1 time d(31may2006))
stset timetocrash_trans, id(l_no) failure(acc == 1)
```

initial Multivariate model on significant univariate variables for days to crash

```
i: stcox i.sex i.age i.remote_area i.cob i.risk_taking_scr i.audit_c i.eco_qu i.l_period
i.unp_hours i.driven_b_l i.no_super
```

Final model for days to crash, proportional hazards assumption and martingale residuals

```
xi: stcox i.sex i.driven_b_l i.remote_area i.cob i.risk_taking_scr , strata(age) noshow nolog
schoenfeld(sch*) scaledsch(sca*)
stphtest, detail
xi: stcox i.sex i.driven_b_l i.remote_area i.cob i.risk_taking_scr , strata(age) mgale(mg)
predict coxsn, csnell
stset coxsn, failure(timetocrash_trans)
sts gen H=na
tway (scatter H coxsn)
```