PREDICTING TRAFFIC CRASH INVOLVEMENT USING INDIVIDUAL DRIVING HABITS, DRIVING RECORD, AND TERRITORIAL RISK INDICES

July 2017

Authors: Michael A. Gebers,
Jeff Moulton
Research and Development Branch
Licensing Operations Division

RSS-17-254

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
July 2017	Final Report	

4. TITLE AND SUBTITLE
Predicting Traffic Crash Involvement Using Individual Driving Habits, Driving Record, and Territorial Risk Indices

5a. CONTRACT NUMBER

5b. GRANT NUMBER TR1016

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michael A. Gebers, Jeff Moulton

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
California Department of Motor Vehicles
Research and Development Branch
P.O. Box 932382
Sacramento, CA 94232-3820

8. PERFORMING ORGANIZATION REPORT NUMBER
CAL-DMV-RSS-17-254

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
California Office of Traffic Safety
2208 Kausen Drive # 300, Elk Grove, CA 95758

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES email: research@dmv.ca.gov

14. ABSTRACT

This study surveyed a sample of California drivers to determine their habits and opinions on selected traffic issues. The study also assessed the importance of exposure and territorial risk indices as predictors of traffic crashes beyond that of driver record factors. The information provided in this report is intended to assist traffic safety administrators and lawmakers in improving services and in developing more effective driver safety programs.

15. SUBJECT TERMS Crash-conviction relationships, Crash factors, Crash proneness, Driver control, Driver improvement, Driving record characteristics, Longitudinal analysis, Multiple regression, Multivariate analysis, Negligent-operator point system, Probability model fitting, Questionnaire survey techniques

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Douglas P. Rickard
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPONE NUMBER (*Include area code*)
Unclassified	Unclassified	Unclassified	None	73	916-657-5768

PREFACE

This project is part of the California Traffic Safety Program. Funding for this program was provided by a grant from the California Office of Traffic Safety, through the National Highway Traffic Safety Administration. The report was prepared by the Research and Development Branch of the California Department of Motor Vehicles. The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the State of California or the National Highway Traffic Safety Administration.

ACKNOWLEDGMENTS

The study was conducted under the general direction of David DeYoung, Research Chief (retired), and the supervision of Robert Hagge, Research Chief (retired). The authors extend thanks to Bayliss Camp, Ph.D., Chief, Research and Development Branch for his reviews and edits to the final draft of this report. Mark Talan of the Office of Traffic Safety was very helpful in guiding the project's grant process.

Appreciation goes to Doug Rickard, Associate Governmental Program Analyst, who supervised the printing and mailing operations associated with the survey and helped to produce the final draft of the report, and to Douglas Luong, Staff Services Analyst, for the help in typing and proofing the many drafts of this report.

The authors would like to extend thanks and appreciation to the following employees of the California Department of Motor Vehicles for their valuable contributions to this project: Florence Lagrose, Diyand Heu, Sangita Patel, Nyia Lo, Elanda Wise, Joe Amaro, Juan Bonillas, Doug Mitchell, David Iwate, Norma Campisi, Chue Veng, Tyrone Phillips, Liew Saetern, Darryl Curry, and Rico Hamm.

EXECUTIVE SUMMARY

Background

To profile the California driving population, the California Department of Motor Vehicles (DMV) periodically extracts data on a 1% random sample of licensed drivers from its driver licensing files. However, certain data of interest are not available in the database and can be best obtained by use of a survey questionnaire. These data include driving exposure indices (e.g., mileage, driving territory, and avoidance of certain driving conditions), personal habits related to the association between alcohol and alcohol impaired driving (e.g., frequency and amount of alcohol consumption and attitudes about the risk associated with driving while impaired), and indicators of socioeconomic status (e.g., annual income and occupation class). Questions measuring a respondent's attitude about certain driving and safety issues of current interest are best captured in surveys of public opinions on these matters.

DMV's Research and Development Branch has conducted two prior surveys that targeted representative samples of drivers from the general driving population (Frincke & Ratz, 1984; Peck & Kuan, 1982). As the behavior and attitudes measured by such questionnaires can be expected to change in just a few years, with new, younger drivers entering the population and older drivers dropping out, it was deemed necessary to conduct the updated survey of the general driving population sampled in this study. Behavior and attitudes may shift for other reasons as well (e.g., social movement activity such as Mothers Against Drunk Driving, technology changes, changes in laws and enforcement regimes, immigration, and other demographic factors).

Information obtained from the survey allowed driving exposure and habit response items in the survey to be correlated with driver record variables and territorial risk indices. The analyses performed in this study involved the construction of multivariate profiles of crash-free and crash-involved drivers in an attempt to better understand the correlates of crash risk and, thereby, to aid in the development and application of interventions for drivers with high crash-risk propensities.

Methods

Subjects

Two groups of drivers were selected for this study.

The first group, referred to as the non-survey sample, consisted of a random sample of 268,480 licensed California drivers. These drivers are representative of validly licensed drivers in general and were used for general statistical modeling purposes.

The second group, referred to as the survey sample, consisted of 21,323 drivers who were randomly selected from the general driving population and mailed the California Driver Survey (see below) in 2010. The survey sample was used to develop regression equations based partly on information not available from the driver record (e.g., mileage, driving habits, drinking habits, and socio-economic factors).

Survey Materials

The California Driver Survey questionnaire consists of 22 items covering the following topical areas, which have been demonstrated in earlier studies to be correlated with traffic crash involvement:

- Exposure – amount of driving,
- Exposure – type of driving,
- Vehicle type driven,
- Distracted, drowsy, and aggressive driving,
- Socio-economic status, and
- Alcohol and drug usage.

Two cover letters were constructed. One was used as the initial contact letter, and the other, a slightly different one, was sent as a follow-up mailing to individuals who did not respond to the first letter.

Procedure

The initial (wave 1) cover letter and survey were mailed on September 3, 2010 to all drivers in the survey sample. On November 8, 2010, drivers not responding to the first wave were mailed a second (wave 2) cover letter and survey. April 1, 2011 was selected as the cutoff date for the survey returns. After this date, all drivers in the survey sample were categorized as respondents (30.71%), returned unclaimed (19.74%), or non-respondents (49.55%).

Analyses

A series of Poisson multiple regression equations were constructed to determine which set or subset of predictors (i.e., driver record variables, territorial crash and conviction indices, and driving exposure and habits survey variables) contributed to the prediction of 17-month total traffic crash counts (during the 17-month period of January 1st 2009 through May 31st 2010 for all drivers).

Missing data from survey respondents were modeled by way of the multiple imputation techniques available in SAS software.

Results

The study was designed to answer four questions. These questions and their answers are summarized below.

1. *What driver record variables and territorial rate indices predict total crash involvement?*

   Results from the non-survey sample regression models indicated that increasing total crash frequency was associated with the following:

   - Being male,
   - Being young,
   - Holding a commercial license,
   - Increased prior total citation frequency,
   - Increased prior total crash frequency, and
   - Residing in a higher territorial crash rate index area.

2. *Does the driving habits and exposure information obtained from the driver survey add any unique contribution to the prediction of total crash involvement beyond that provided by the driver record variables and territorial rate indices?*

   After having established the association between crashes and the driver record and territorial predictors in the larger non-survey sample, the next logical extension was to examine the unique contributions of the driving habits and exposure variables obtained from the survey in predicting the total crash criterion. Several equations were constructed with each differing in the subsets of driver record, territorial rate indices, and driving habits and exposure predictors that were eligible for entry in the model.

   The parameter estimates from the various equations indicated that increased crash frequency was associated with the following self-reported driving habits and exposure factors:

   - High weekly mileage,
   - Driving aggressively,
   - Driving while distracted,
   - 20-28 drinking days per month,
   - Driving after using alcohol or other illegal drugs, within the past 12 months,
   - Lower educational level, and
   - Lower income level.

3. *What is the relative contribution of the various subsets of predictor variables?*

   The following observations are offered based on the findings from the regression models for the non-survey and survey samples:

   - Each of the three sets of predictors (i.e., person-centered driver record variables, territorial indices, and person-centered survey variables) made unique contributions to crash prediction.
   - Person-centered driver record variables were superior to territorial indices in predicting crashes.
   - Person-centered driver record variables were roughly similar to the survey variables in their ability to predict counts of total traffic crashes, with the person-centered driver record variables having slightly better prediction.

4. *What is the utility of applying the regression equations to predict individual crash involvement for individual drivers with selected profiles?*

   To illustrate the accuracy of the regression equations in predicting crash risk of individual drivers, 2x2 cross-classification tables were constructed displaying the classification of individuals in the survey sample based on each person's predicted and actual crash-involvement frequency.

   It was demonstrated that a modest gain in individual prediction was obtained by adding the statistically significant habits and exposure variables to the equation containing the driver record and territorial crash index predictors. That is, the true-positive rate increased slightly, from 21.35% to 25.19%, when adding the habits and exposure predictors. However, it was also demonstrated that in all models, there was low accuracy in predicting which specific individuals were crash-involved, as evidenced by the misclassification of the majority of crash-involved drivers.

Recommendations

The following recommendations are offered based on the study findings.

1. Although the focus of the present study was on statistical modeling of total crash counts, the survey produced ample information for additional analyses, such as further investigation of the relationships among other driver record and driving habits indices**.** Specifically, it is recommended that a subsequent report containing a series of description-based contingency tables be produced. Such tables would examine the bivariate relationships between pairs of variables such as exposure and education, exposure and occupational status, and exposure and crash group. Such an effort (currently being planned by the California DMV) should result in additional and more complex profiles of crash-free and crash-involved drivers and perhaps to a better understanding of the correlates of crash risk.

2. The present study assessed only the associations between prior driver record, territorial indices, and driving habits and exposure variables and the total crash criterion. The availability of these data invites use of other criteria of interest. Specifically, it is recommended that future surveys be planned and conducted to model (1) total traffic citations and (2) crashes in which the driver was deemed by the reporting officer as had-been-drinking and obviously impaired.

3. Historically, driver record data are commonly aggregated into multi-year (e.g., two-year or three-year) predictor and criterion periods for use in regression models. A different modeling strategy, regarding the use of a survey sample larger than the one used in the current effort, would be to treat separate yearly counts of driver record entries (e.g., crashes) as repeated measures in the regression models. That is, fixed-effects and/or random effects regression methods could be applied to these data by treating the annual counts as panel data consisting of measurements of predictor and response variables at two or more points in time for many individuals. Panel data have two major attractions: (1) the ability to control for unobserved variables and (2) the development of models that make it possible to determine the direction of any suspected causality among variables associated with one another.

4. The fact that self-reported aggressive driving was significantly associated with crash frequency in each model containing survey variables should encourage the Department to conduct its planned empirical study of aggressive driving. This study would analyze the historical driving records of a large representative sample of California drivers to determine

what patterns and combinations of driving behaviors thought to be aggressive in nature would be good predictors of having a future crash risk greater than that posed by *prima facie* negligent operators in California**.** Establishing that chronic aggressive driving tends to lead to high future crash risk would provide justification for administering intervention actions (such as license suspension) against these drivers earlier than would otherwise occur under the Department's existing post licensing control system. Increasing the severity of sanctions against high-risk, aggressive drivers is also supported explicitly in the California Strategic Highway Safety Plan. Such a study is currently being conducted by the California Department of Motor Vehicles (Wu, in press).

5. Given the statistically significant relationship presented in this study between crash involvement and self-reported distracted driving, the Department continued its efforts to evaluate the relationship between cell phone use while driving and traffic crash involvement as reported in Limrick, Lambert, and Chapman (2014). Specifically, this finding lead to further research (funded by an Office of Traffic Safety Grant) establishing that distracted driving violations in combination with negligent operator treatment points identified higher risk drivers for potential licensing actions than negligent operator points alone (Lambert, Fox, & Camp, 2017). This finding substantiated the consistent and reliable association between distracted driving and traffic crash risk displayed in the results and tabulations from the present study.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

TABLE OF CONTENTS (continued)

LIST OF TABLES (continued)

INTRODUCTION

Background

The responsibility for reducing traffic crashes and their related fatalities, injuries, and costs resides with many diverse groups. Overall, the ultimate goal is to reduce crash rates, and each group sets out to achieve this goal with different mechanisms, methodologies, and measurable objectives. Safety performance must be measured using data from more than one source (i.e., more than just a count of traffic convictions and crashes obtained from a state's driver licensing files). One mechanism for gathering data that has become common is the use of public surveys.

Over the past several decades, the practice of surveying has evolved along with the development of supporting technologies. As a result, public surveys are becoming more widely used to assist traffic safety administrators in setting public policy. Public surveys are commonly used not only to determine general attitudes towards traffic safety but also to measure a driver's experience with law enforcement, self-reported driving behavior, and perception of being detected and cited for traffic law infractions. These surveys can be short-term research projects or ongoing efforts to track long-term behavioral trends.

As an example of such survey efforts, the AAA Foundation for Traffic Safety (2016) published a report entitled 2015 Traffic Safety Culture Index. This report presented the findings from a telephone interview survey of 2,545 drivers from around the United States. The goal of the survey was to measure driver knowledge, attitudes, behaviors, and experiences relevant to traffic safety.

Another example is a project funded by Transport Canada/MADD Canada that consisted of a series of surveys and focus groups involving Canadian drivers (EKOS Research Associates Inc., 2007). The primary objective of this project was to measure the concerns, knowledge, attitudes, and behaviors of Canadian drivers on impaired driving issues. The researchers collected information to determine where awareness needs to be heightened and knowledge needs to be increased. Their report provides detailed information about public views on a variety of impaired driving issues.

The California Department of Motor Vehicles (DMV) conducts various traffic safety research studies involving the development and evaluation of programs aimed at increasing driver

competency and reducing crash risk. These studies rely heavily on driver record information stored on the Department's automated Driver Record Master file. This database provides extensive information on driver license variables, prior traffic violations, crash involvements, demographic indicators, and other such information. However, certain data of high research value are not available in this database. These unavailable data include exposure and driving habits measures (e.g., mileage, driving territory, and avoidance of certain driving conditions), personal habits related to alcohol and impaired driving (e.g., frequency and amount of alcohol consumption and attitudes about the risk associated with driving while impaired), and indicators of socioeconomic status (e.g., annual income and occupation class). This type of information can be inexpensively obtained through the use of mailed questionnaires.

The Department has conducted two prior mailed surveys of representative samples of licensed drivers from the general driving population (Frincke & Ratz, 1984; Peck & Kuan, 1982). The information from these surveys is obviously outdated and not very useful for the development and evaluation of new licensing programs and countermeasures to reduce traffic crashes.

The survey in the present project obtained updated information that will enable new research projects to be conducted to support efforts to reduce traffic crashes. In this project, alcohol consumption patterns, socio-demographic characteristics, and other survey response items were correlated with future crash involvements. Multivariate profiles of crash-free and crash-involved drivers were developed to better understand the correlates of crash risk and, thereby, aid in the development and application of interventions for groups with high crash propensities.

This project also explored the confounding influence of exposure variables, such as mileage and conditions of driving, on crash risk. Specifically, it assessed the relationships between driver record (e.g., prior crashes and citations), territorial driver-record indices (e.g., territorial crash rate index), and driving exposure/habit variables (e.g., weekly miles driven and number of times driving after consuming alcohol) and the likelihood of traffic crash involvement among a random sample of drivers from the general population of licensed California drivers.

For decades, drivers living in certain neighborhoods have paid dramatically more for their automobile insurance than have people residing just across the zip code line, because of insurers' heavy reliance on geography in setting premiums. Such models relate risk indices to geographical aggregates of people and not to individuals. All drivers living in a given zip code area or territory were charged the same premium unless other factors were also considered. On intuitive grounds alone, it seems obvious that an expected loss-insurance premium model based

solely on geographical location of residence is not the most reasonable or optimal strategy, as individuals within any given territory will vary in their driving skills, attitudes, driving habits, and other variables that influence their crash propensity. Even the physical nature of the driving environment (e.g., traffic density, type of roads, and driving conditions) is not constant within any given territory or zip code area.

Under terms of Proposition 103 (passed by California voters in 1988) and regulations issued by California's Insurance Commissioner in 2006, each auto insurance company is required to submit a special application to the California Department of Insurance showing that it was complying with the rules specified by Proposition 103. Specifically, in August of 2006, the companies were given two years to phase in a new safety-record based system in which the insurance premium is based primarily on an individual's safety record. The new rules require that a driver's record, the annual miles they drive, and the number of years they have been licensed must each have greater impact on insurance premiums than zip code or other factors such as marital status, which historically were weighted heavily in setting premiums. Certainly, there will still be geographic variation in premiums, but not nearly as large as seen in the past. For these reasons, it is necessary to learn more about variables that are descriptive of individuals.

Although many correlates of individual crash liability have been found, it is incorrect to conclude that individual crash involvement can be predicted with a high degree of precision. It is also not true that the majority of crashes are caused by a small number of "crash-prone" drivers. Several large-scale studies have shown that the majority of crashes in any time period involve drivers with average or good prior driving records (e.g., Gebers, 1998, 1999, 2003; Gebers & Peck, 2003a, 2003b). There is a large amount of luck or chance in determining crash involvement because of the complex chain of interactive events that determine a given crash occurrence. A very negligent driver may not become crash-involved for long periods through pure luck or through the defensive driving of attentive drivers, whereas a safe driver may have the misfortune of being victimized by some other driver's carelessness. All of these factors operate against being able to accurately predict the crash-involvement frequencies for individual drivers.

Individual Versus Group Risk Prediction

The distinction between individual and group risk predictions is important to keep in mind when evaluating the efficacy of any crash-prediction system. Although accurate individual risk prediction is a highly desirable goal, it is not always a critical one. The actuarial sciences inevitably involve a very large number of risk factors, and the actuary must establish a premium structure and funding pool sufficient to offset the net dollar amount of claims made over any fixed interval of time. If one has established, for example, that persons whose blood pressure exceeds a certain threshold have a two-fold greater than average probability of dying before, say, 60 years of age, all members of this blood pressure group might be charged a higher life insurance premium, ideally one that is proportionate to that group's higher average early mortality risk. In doing so, it should be recognized that many individuals in the high blood pressure group will actually live longer than average and end up paying more than their "fair share." Conversely, many persons with normal blood pressure die early and pay less than their "fair share." A large number of such miss-assessments is a consequence of the fact that blood pressure ratings, despite being one of the single best indicators of life expectancy, still only predict a small percentage of the variance in the death rate of the individuals comprising any population.

In regard to crash prediction, the distinction between group and individual predictions is illustrated by the following two tables derived from a random sample of California drivers obtained from the Department's California Driver Record Study Database, which consists of records for a 1% random sample of licensed California drivers.

Table 1

Rate of Total Crash Involvements in the Subsequent 3 Years by Number of Total Crashes in the Prior 3 Years

Prior total crashes	Number of drivers	Mean subsequent crash rate	Times-as-many subsequent crashes	% subsequent crash-free drivers
0	172,115	0.136	1.00	87.83
1	26,208	0.208	1.53	82.18
2	3,341	0.296	2.18	76.71
3+	478	0.452	3.32	67.78

Note. Pearson correlation between prior total crashes and subsequent total crashes = .082 ($p < .0001$). The "times-as-many" ratio represents the relative increase in each group's subsequent crash rate compared to the zero-group's subsequent crash rate.

Table 2

Rate of Total Crash Involvements in the Subsequent 3 Years by Number of Total Citations in the Prior 3 Years

Prior total citations	Number of drivers	Mean subsequent crash rate	Times-as-many subsequent crashes	% subsequent crash-free drivers
0	140,035	0.125	1.00	88.81
1	40,484	0.178	1.42	84.47
2	13,182	0.227	1.82	80.55
3	4,854	0.273	2.18	77.54
4	1,970	0.288	2.30	76.19
5	844	0.345	2.76	72.16
6+	773	0.376	3.01	69.34

Note. Pearson correlation between prior total citations and subsequent total crashes = .105 ($p < .0001$). The "times-as-many" ratio represents the relative increase in each group's subsequent crash rate compared to the zero-group's subsequent crash rate.

The tables show an obvious trend toward increased crash involvements as a function of a driver's prior citation and crash frequencies. However, the majority of drivers are crash-free at all prior record levels. This implies that any graduated premium structure based on prior record would necessarily penalize a vast number of drivers who would not be involved in a crash during the period of time for which the premium is charged. When the data, however, are examined on a group basis, that is, in terms of the number of crashes per 100 drivers in each category, drivers with poor records are found to have many more crashes than drivers who are free of convictions or crashes. Therefore, from an actuarial viewpoint, these data would clearly support charging drivers with bad records higher premiums because the expected number and valuations of crash claims filed by them is much higher.

Multi-Variable Prediction

Because crash risk is a complex stochastic function of many variables, strategies for optimally estimating and predicting individual risk must be multidimensional in nature. There are a variety of techniques for doing this, and one of the most powerful and frequently used is multiple regression (Gebers, 1999; Gebers & Peck, 2003a). When used to model crashes, multiple regression analysis produces an equation giving the most accurate possible prediction of the crash-involvement rate or probability for each driver, using an optimum linear composite of the various predictor variables (e.g., age, gender, prior driving record, and mileage). Although

multiple regression assumes a model that is linear and additive in its parameters, nonlinear and interactive relationships between the independent variables and the crash criterion can be evaluated by incorporating additional parameters (e.g., polynomials and interactions) into the model.

As one of its primary tools, the multiple regression equation can be used to predict whether a given driver will be crash-involved in a specified subsequent time period. The accuracy of such prediction is often presented in a four-fold classification schematic. This is illustrated in Table 3.

Table 3

Crash Prediction Accuracy

	Predicted state	
Actual state	Crash-involved	Crash-free
Crash-involved	a = true positive	b = false negative
Crash-free	c = false positive	d = true negative

With perfect prediction, all drivers would be in cells a or d, and no drivers would be in cells b or c. Drivers in cell c are termed false positives. These drivers are predicted to be crash-involved but are actually crash-free. Drivers in cell b are termed false negatives. These drivers are predicted to be crash-free but are actually crash-involved. It is desirable to minimize the proportion of drivers in cells b or c and to make fewer errors than would be made in classifying drivers without the prediction equation. To be of any value, the equation must result in more classification accuracy than would be expected by chance alone and be sufficient to offset the cost of its application.

For the present study, inferential and descriptive statistical techniques were applied to address the following key issues:

(1) What is the relative importance of driver exposure, attitude, habit variables, territorial indices, driver demographics, and prior driver record in predicting future total crash involvement?

(2) How accurately can future driver crash involvement be predicted from an optimum combination of predictor variables?

The following sections of this report present the methodology, results of the statistical analyses, and a discussion of the implications of the study's findings.

METHODS

This section describes the project's methods. Some methodological details are reserved for the Results section because they are more understandable within the context of the study findings.

Subjects

The driver record data for drivers in the survey and non-survey samples that were analyzed in this study were extracted from the California Department of Motor Vehicles Driver Record Master file in April of 2011. These data are an extension of the California Driver Record Study, which consists of a systematic 1% random sample of California drivers whose driver licenses end in terminal digits (TD) 01. The sampling design and data collection methods for the California Driver Record Study are described in detail by Gebers and Peck (2003a). Specifically, the following two groups of subjects were selected for this study.

Non-Survey Sample

This group consisted of all of the 268,480 TD 01 drivers who were alive on the 2011 extract date and whose licenses had not been expired for more than 12 months. These drivers are representative of validly-licensed drivers in general. The analyses involving this sample used driver record variables and territorial indices only (see below). Because of its larger size, the non-survey sample was used for general modeling purposes. The directions and magnitudes of the parameter estimates from the sample's driver record equations (presented in the Results section) were compared to those for similar parameter estimates from the smaller survey sample's driver record equations (presented in the Results section) to assess the stability of the survey sample's estimates.

Survey Sample

This group consisted of 21,323 randomly selected TD 01 drivers who were mailed the California Driver Survey (see below) in 2010.[1] The survey sample was used to develop regression equations (presented in the Results section) based partly on information not available from the

[1]SAS PROC SURVEY SELECT (SAS Institute Inc. 2009) was used to select this sample from the larger sample of TD 01 drivers.

driver record (e.g., mileage, driving habits, drinking habits, socio-economic factors). Two mailing waves were used to maximize the response rate. Out of the 21,323 drivers sent surveys, 6,548 (30.71%) responded, 4,209 (19.74%) were returned by the post office as undeliverable, and 10,566 (49.55%) were non-respondents (i.e., received but failed to return the survey). Removing the undeliverable survey forms from the total number of surveys raises the response rate to 38.26%. Since the survey sample was purposively stratified to over represent crash- and multiple-crash-involved drivers, all survey responses were subsequently normalized by appropriate population weights (see the discussion of weighting in the Regression Model Development section below).

Materials

The California Driver Survey is presented in Appendix A. The survey has 22 items. The survey's content was derived from an extensive review of driver survey literature and the experiences of previous traffic safety studies (e.g., AAA Foundation for Traffic Safety, 2008; Blomberg, Peck, Moskowitz, Burns, & Fiorentino, 2005; Frincke & Ratz, 1984; Gebers, 2001; Gruenewald & Nephew, 1994; Hennessey, 1995; Kelsey & Janke, 2005; Peck, Gebers, Voas, & Romano, 2008; Peck & Kuan, 1983).

The survey items cover the following topical areas, which have been demonstrated in earlier studies to be correlated with crash involvement:[2]

(1) Driver opinion: Item 1,
(2) Exposure – amount of driving: Items 2, 3, 4, 6, 7, 12, and 14,
(3) Exposure – type of driving: Items 8, 9, 10, and 11,
(4) Vehicle type driven: Item 5,
(5) Distracted, drowsy, and aggressive driving: Item 13,
(6) Socio-economic status: Items 15, 16, 17, and 18,
(7) Alcohol usage: Items 19, 20, 21, and
(8) Drug usage: Item 22.

[2] The interested reader is referred to Blomberg, Peck, Moskowitz, Burns, and Florentino (2005), Gebers (2001), and Peck and Kuan (1983) for parameter estimates and relative risk indices of crash involvement associated with similarly worded survey items.

Two cover letters were constructed (Appendix B and Appendix C). One was used for the first mailing (wave 1), and the other, slightly modified, was used for the follow-up mailing to individuals who did not respond to the first letter (wave 2). An identifying number was recorded on each survey and cover letter, allowing each driver in the sample to be linked to their response status and to their driver record for subsequent analysis.

Appendix D contains the descriptive response distribution for each item on the California Driver Survey.

Procedure

The wave 1 cover letter and survey were mailed on September 3, 2010 to all drivers in the survey sample. On November 8, 2010, drivers not responding to the first wave were mailed a second cover letter and survey. April 1, 2011 was selected as the cutoff date for survey returns. After this date, all drivers in the survey sample were categorized as respondents (those drivers completing and returning the survey), undeliverable (those drivers whose letters were returned unclaimed from the U.S. Post Office), or non-respondents (those drivers who supposedly received the survey but didn't return it). Results in this report are based on all non-duplicate surveys returned from the two mailings as of April 1, 2011.

Microsoft Excel was used to track and record survey response status. Microsoft Access was used for keying the survey responses. Following their keying, respondents' data were matched and merged to their driver records and appropriate territorial driving-record indices for use in the subsequent analyses described in the next section.

Analyses

This section presents an overview of the statistical analyses and describes the sequential steps used in the parameter estimation process.

Variables

The following is a brief description of the variables used in the analyses.

1. Criterion measure: Total crashes reported by law enforcement agencies and/or involved-drivers occurring during the 17 months from January 1, 2009 through May 31, 2010**.** The

original intent was to use total crashes occurring between January 1, 2010 and December 31, 2010 as the criterion measure because it was closer in time to the survey mailings and responses in late 2010**.** However, an internal analysis conducted by the authors found a substantial lag of over 10 months between crash-involvement date and the date the crash was recorded on the Department's Driver Record Master file for police reported crashes occurring during 2010.[3] Given this substantial lag in the California Highway Patrol's (CHP's) reporting of crashes to the Department, there were concerns that any modeling of crashes occurring in 2010 for data extracted in April 2011 on both the non-survey and survey samples could produce unstable and biased regression parameter estimates and perhaps even result in models that fail to converge**.** The authors believe that this modified criterion period still maintains a substantial temporal relationship between the driving record and surveyed driving habits**.** That is, the events recorded on the driving record and the self-reported measures from the survey are close enough in time to obtain meaningful prediction of the total crash criterion**.** Additionally, prior work by Peck and Gebers (1992) demonstrated that more reliable regression parameters and test statistics are obtained by increasing the length of the continuous criterion period (in this case, from 12 months to 17 months).

2. Predictor variables: The independent variables used in the analyses are listed below:

a) Prior total crashes during 2006-08 that were reported by law enforcement agencies and/or involved parties and later extracted from the Department's Driver Record Master file.

b) Prior total citations (convictions, failure-to-appear violations, and traffic violator school citation dismissals) occurring during 2006-08 and extracted from the Department's Driver Record Master file.

c) Miscellaneous biographical and licensing variables (driver age, gender, class of license, etc.) extracted from the Department's Driver Record Master file.

d) Survey variables (self-reported mileage, drinking/driving practices, socio-economic variables, and other factors not available from the driver record files)**.** As with any survey, there are respondents who for some reason did not answer one or more survey items**.** Rather than entirely eliminating such respondents or non-answered items from the statistical analyses, a strategy of multiple imputation of missing data was employed on the 6,548 survey respondents through the use of SAS PROC MI (SAS Institute Inc., 2009)**.** This approach has become widely used for handling all kinds of missing data in a wide variety of statistical models (Allison, 2002, 2009).

[3] Crashes reported only by police account for over 50% of all crashes reported to the California Department of Motor Vehicles. Historically, a 3-month lag between crash date and crash update date for police reported crashes has been reported. The increase to a 10-month lag has been attributed to staffing shortages at CHP and to furloughs of non-sworn CHP personnel.

e) Territorial variables or indices reflecting the incidences of traffic crashes and convictions and other driver record entries in the area of residence [zip code] of each driver occurring during 2006-08.

The territorial indices described above were derived from the results of a file-pass program that is run annually against the Driver Record Master file at the beginning of each calendar year. Each index represents how the mean value for a given zip code compares to the statewide statistical mean. To determine how well these zip code indices predict an individual's crash likelihood, two territorial indices were calculated, and each driver in the sample received the values indicated for his/her territory of residence. For example, all drivers living in zip code 95820 would have received the two index values (one that is a composite of exposure, licensing actions, crashes, and convictions and one that represents just crashes) computed for that zip code. The indices are the same as those recommended by Peck and Kuan (1983) and consisted of the following:

1. The territorial composite index: This index is a composite of the eight driver-record variables listed below**.** The index was previously used by the insurance industry in developing assigned-risk territorial areas.

    a) Driving exposure (defined as the average number of years drivers have been licensed in a zip code)

    b) License suspensions/revocations

    c) Personal injury crashes

    d) Total crashes

    e) Major (2-point) convictions

    f) Minor (1-point) convictions

    g) Non-moving (0-point) convictions

    h) Failure-to-appear violations

2. The territorial total crash rate index.

Regression Model Development

For highly skewed count data such as traffic crashes, appropriate regression modeling techniques include Poisson, Poisson with correction for overdispersion, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial. In the present study, each of these techniques was assessed by running a series of preliminary regression models and reviewing diagnostic statistics associated with model fit and related assumptions. Based on the findings for these models, it was decided that Poisson regression with correction for overdispersion is the most appropriate modeling technique for both the survey and non-survey samples.[4] The interested reader is referred to Boyer, Dionne, and Vanasse (1990), Davis (1990), Famoye and Singh (2006), Grogger (1990), Lee, Wang, Scott, Yau, and McLachlan (2006), Lord and Mannering (2010), and Lord, Washington, and Ivan (2005) for detailed discussions of these regression techniques.

PROC GENMOD was used to produce the Poisson regression models for the non-survey sample (SAS Institute Inc., 2009). Both PROC GENMOD and PROC MIANALYZE were used to produce the Poisson regression models for the survey sample (SAS Institute Inc., 2009). PROC MIANALYZE combines the results of the analyses of missing data imputations (described above) and generates estimated parameters and statistical tests from the PROC GENMOD output.

The regression models presented in the next section address the following four questions:

1. *What driver record variables and territorial rate indices predict total crash involvement?* This question was addressed by using the non-survey sample for general modeling purposes**.** Several regression models were developed to assess how the prediction of crashes would be affected by including or excluding various subsets of the driver record predictors and territorial rate indices**.** Driver record predictors are defined as variables that are available on the driver record**.** These consist of age, gender, class of license, prior 3-year total convictions, and prior 3-year total crashes**.** As discussed above, the territorial indices consist of the territorial composite index and the territorial crash rate index**.** For these analyses, a

[4] The Poisson model corrected for overdispersion (variance of the criterion measure being greater than the mean of the criterion measure) uses a dispersion parameter in the equation. The inclusion of the dispersion parameter does not introduce a new probability distribution, but rather just gives a correction term for testing the parameter estimates under the Poisson model. The Poisson models are fit in the usual way, and the parameter estimates are not affected by the correction term. The estimated covariance matrix is inflated by this correction factor, thereby producing more accurate standard errors for significance testing. This method leads to non-biased results if overdispersion is modest (Cox, 1983). In the present data, overdispersion was negligible.

predictor is considered statistically significant if the associated *p*-value is .10 or lower, meaning that one would expect to find an association as large or larger than obtained in no more than 10 out of 100 random samples if the predictor had no real association with the crash criterion.

2. *Does the driving habits and exposure information obtained from the driver survey add any unique contribution to the prediction of total crash involvement beyond that provided by the driver record variables and territorial rate indices?* To address this question, several regression equations were developed from the survey sample using subsets of person-centered driver record variables, territorial rate indices, and habits and exposure variables obtained from the survey. The available predictor pool consisted of well over 100 variables. In this kind of situation where there are many explanatory variables in the maximum model, an "all possible variables" model is often impractical**.** It is even possible that such a model will be mis-specified and be no better than an intercept only model that predicts the mean criterion value for all individuals upon which the model was constructed. Since the goal of the analyses designed to address question two was to obtain a regression model with only variables providing unique predictability of the total crash criterion (and hence eliminating variables that are statistically unreliable and/or increase prediction error), it was decided to apply a statistical selection criterion to these data.

   Several statistical selection criteria are available. These include forward selection, stepwise selection, backward selection, etc**.** The interested reader is referred to a classic text such as Pedhazur (1973) for a discussion of the various selection methods available for multiple regression analyses and the strengths and weaknesses of each method.

   The one employed for the preliminary models constructed for the present study was the forward selection criterion. A forward selection procedure starts with an "empty" model with no explanatory variables and adds variables one at a time until there is no further statistically significant improvement to the model by adding another variable. To be consistent with the non-survey sample regression models, a variable had to meet an entry criterion of *p* less than or equal to .10 in the forward selection regression method. When assessing the regression models presented below, it is important to consider the fact that just because a particular variable or category does not enter the equation does not necessarily indicate that they are not predictive of crashes. Such occurrences indicate that the non-selected predictors do not uniquely contribute to predicting the total crash criterion after the

other variables, which might be highly correlated to variables not yet entered, have already entered the model**.**

Before addressing the next study question, it is appropriate at this point to briefly discuss the weighting methodologies utilized in the survey sample equations.[5] Two types of weights were used in the present study.

One was a design weight. As stated above, the survey sample was intentionally oversampled on crash-involved drivers in order to obtain a representative sample of these drivers and for possible long-term follow-up of them for future research efforts. The design weight was calculated as the inverse of the sampling fraction for crash-free and crash-involved drivers. The design weight was used to adjust for the oversampling of the crash-involved drivers so that drivers with two crashes represented a proportion of 0.0315 or 3.15%. This same group of drivers represents a proportion of approximately 0.0089 or 0.89% in the general population of all drivers. Therefore, the design weight (i.e., the weight assigned for each observation) for this group of drivers was approximately 0.28254 or [1/(0.0315/0.0089)].

The second weight was used to correct for, at least in part, non-response bias. Izrael, Hoaglin, and Battaglia (2004) state that it is often the case in survey research that a survey sample may cover segments of the target population in proportions that do not match the proportions of those in the population itself. The authors note that the differences may arise, for example, from sampling fluctuations, from non-response, or because the sample design was not able to cover the entire target population. Under such scenarios, the association between the sample and the population can be improved by adjusting the sampling weights of the cases in the sample so that the marginal totals of the adjusted weights on specified characteristics agree (for the most part) with the corresponding totals for the population (referred to as the controls). This procedure is known as sample balancing (also referred to as "raking").

In the present study, the concern was on the selectivity of individuals deciding to respond, or not respond, to the survey and to apply a correction to reduce, at least in part, associated bias. Analysis of survey response status indicated that respondents did differ on a number of person-centered biographical and driver record characteristics. As summarized in Table 4, respondents were more likely to be female and older, less likely to have a commercial license

[5] For a general presentation of weighting techniques, the reader is referred to Cochran (1977).

and a prior suspension or revocation action taken against their driver license, exhibit a lower rate of traffic citations (minors and majors), and exhibit a slightly higher prior crash rate.

The correction consisted of applying a balancing weight value to each survey respondent such that the weighted distribution of the respondents is similar to the unweighted distribution of the controls**.** Specifically, this methodology involved the calculation of a weight modeled from a logistic regression equation (not displayed) developed from the survey sample combined with the larger general driving population sample. The weight was computed as the reciprocal of the predicted probability of a survey response**.** The two weights (i.e., the design weight and the balancing weight) were multiplied together to create a total weight used for analyses addressing question two.

Table 4

Distribution of Biographical and Driver Record Variables for the General Driving Population and Survey Samples

Variable	General driving population (*N* = 268,480)	General driving population survey sample (*N* = 21,323)	General driving population survey sample respondents (*N* = 6,548)
% male	52.55	53.16	46.96
Mean age	46.13	46.03	50.55
% with commercial license	3.00	2.93	2.82
Prior 3-year % under S/R action	8.10	8.24	5.89
Prior 3-year moving, safety related citations per 100 drivers	40.82	40.53	35.39
Prior 3-year major citations per 100 drivers	1.04	1.10	0.65
Prior 3-year total crashes per 100 drivers	11.43	11.38	12.41

3. *What is the relative contribution of the various subsets of predictor variables?* The Akaike Information Criterion (AIC) estimates calculated from the various Poisson regression

equations developed for questions 1 and 2 above were used to assess each model's relative usefulness in predicting crashes. The AIC is calculated as AIC = -2 Log Likelihood + 2(S), where S is the total number of predictors in the model. The AIC is used for comparing models within the same sample. It penalizes for the number of predictors in the model. The best fitting equation within each of the non-survey and survey samples is defined as the one with the smallest AIC value.

4. *What is the utility of applying the regression equations to predict individual crash involvement for individual drivers with selected profiles?* A two-step strategy was used to answer this question. The first step involved estimating the crash rates for several hypothetical groups of drivers in the non-survey sample with differing driver record and territorial risk characteristics and then assessing the variation in crash expectancies as a function of these differing characteristics. The second step involved generating four-fold contingency tables (discussed earlier in this report) for the survey sample and showing the relationship between each individual's predicted and actual crash-involvement statuses. Each table differed on the subset of driver record, territorial rate indices, and habit/exposure variables that were used to construct the equations. This technique allowed for a direct assessment of the accuracy of the regression equations in predicting individual crash expectancy.

RESULTS

Question 1 – What driver record variables and territorial rate indices predict total crash involvements?

Tables 5 through 10 present the results of the Poisson multiple regression analyses for the non-survey sample. As stated above, an alpha level of .10 was used to assess the level of statistical significance of each predictor in each equation, and the AIC was used to assess the predictive accuracy of each of the various subsets of predictors forming the equations. The best-fitting equation (the one with the smallest AIC value) for these data (Table 6) has an AIC of 91,902. In this model, 3-year total citation frequency is the most significant predictor (as evidenced by the largest Chi-Square value in the table), followed by the territorial total crash rate index and 3-year prior total crash frequency. Three other variables are also statistically significant predictors of 17-month total crash frequency. The directions (signs) of the parameter estimates indicate that increased total crash frequency is associated with the following:

- Being male,
- Being young,
- Holding a commercial license,
- Increased prior total citation frequency,
- Increased prior total crash frequency, and
- Residing in a higher territorial crash rate index area.

Table 5

Poisson Regression Predicting 17-Month Total Crashes from Significant Driver Record Variables and Two Territorial Risk Indices for the Non-Survey Sample

Predictor	Parameter estimate	Wald χ^2	*p*	90% confidence interval	
Intercept	-3.8322	2,610.07	<.0001	-3.9558,	-3.7090
Gender	-0.0344	3.13	.0771	-0.0664,	-0.0024
Age	-0.0104	234.90	<.0001	-0.0114,	-0.0093
Commercial license	0.4298	90.82	<.0001	0.3548,	0.5032
Prior 3-year total citations	0.1991	781.02	<.0001	0.1873,	0.2107
Prior 3-year total crashes	0.3154	266.66	<.0001	0.2814,	0.3491
Territorial composite index	0.0105	0.06	.8056	-0.0596,	0.0802
Territorial total crash index	0.9486	270.66	<.0001	0.8532,	1.0443

Note. N = 268,480. Likelihood ratio $\chi 2$ versus intercept only = 3,818 ($p < .10$). AIC = 91,905. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise.

Table 6

Poisson Regression Predicting 17-Month Total Crashes from Significant Driver Record Variables and Significant Territorial Total Crash Index for the Non-Survey Sample

Predictor	Parameter estimate	Wald χ^2	*p*	90% confidence interval	
Intercept	-3.8239	3,525.29	<.0001	-3.9344,	-3.7138
Gender	-0.0343	3.11	.0776	-0.0633,	-0.0023
Age	-0.0104	234.87	<.0001	-0.0114,	-0.0093
Commercial license	0.4307	91.72	<.0001	0.3559,	0.5039
Prior 3-year total citations	0.1993	789.74	<.0001	0.1875,	0.2108
Prior 3-year total crashes	0.3154	271.55	<.0001	0.2814,	0.3491
Territorial total crash index	0.9509	374.94	<.0001	0.8567,	1.0453

Note. *N* = 268,480. Likelihood ratio χ^2 versus intercept only = 3,820. AIC = 91,902. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise.

Table 7

Poisson Regression Predicting 17-Month Total Crashes from Significant Age, Gender, License Class, and Two Significant Territorial Risk Indices for the Non-Survey Sample

Predictor	Parameter estimate	Wald χ^2	*p*	90% confidence interval	
Intercept	-3.7054	2,398.99	<.0001	-3.8298,	-3.5810
Gender	-0.0330	2.90	.0888	-0.0650,	-0.0027
Age	-0.0145	539.17	<.0001	-0.0155,	-0.0135
Commercial license	0.4896	115.81	<.0001	0.4148,	0.5645
Territorial composite index	0.1011	5.58	.0182	0.0307,	0.1714
Territorial total crash index	1.0330	309.86	<.0001	0.9364,	1.1295

Note. N = 268,480. Likelihood ratio χ^2 versus intercept only = 3,430. AIC = 92,282. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise.

Table 8

Poisson Regression Predicting 17-Month Total Crashes from Significant Driver Record Variables for the Non-Survey Sample

Predictor	Parameter estimate	Wald χ2	*p*	90% confidence interval	
Intercept	-2.8545	8,235.40	<.00001	-2.7929,	-2.9163
Gender	-0.0397	4.16	.0415	-0.0778,	-0.0015
Age	-0.0104	273.73	<.0001	-0.0117,	-0.0092
Commercial license	0.4154	85.20	<.0001	0.3261,	0.5025
Prior 3-year total citations	0.2045	829.73	<.0001	0.1904,	0.2183
Prior 3-year total crashes	0.3345	265.24	<.0001	0.2940,	0.3745

Note. N = 268,480. Likelihood ratio χ^2 versus intercept only = 3,710. AIC = 92,008. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise.

Table 9

Poisson Regression Predicting 17-Month Total Crashes from Significant Two Prior Driver Record Variables for the Non-Survey Sample

Predictor	Parameter estimate	Wald χ^2	*p*	90% confidence interval
Intercept	-3.3399	92,611.60	<.0001	-3.3615, -3.3185
Prior 3-year total citations	0.2328	1,242.19	<.0001	0.2197, 0.2456
Prior 3-year total crashes	0.3576	305.76	<.0001	0.3172, 0.3973

Note. N = 268,480. Likelihood ratio χ^2 versus intercept only = 3,640. AIC = 92,180.

Table 10

Poisson Regression Predicting 17-Month Total Crashes from Significant Two Territorial Risk Indices for the Non-Survey Sample

Predictor	Parameter estimate	Wald χ^2	*p*	90% confidence interval
Intercept	-4.3997	4,046.92	<.0001	-4.5114, -4.2840
Territorial composite index	0.1945	21.05	<.0001	0.1247, 0.2642
Territorial total crash index	1.0259	304.37	<.0001	0.9293, 1.1226

Note. N = 268,480. Likelihood ratio χ^2 versus intercept only = 3.390. AIC = 92,460.

Tables 5 through 10 display equations constructed on the basis of including or excluding various subsets of predictors for the non-survey sample. The equations show how the choice of predictors affects the structure of the models and the AIC estimate. The AIC values associated with these equations and the survey equations presented in the next section are summarized in Table 14. It is apparent from these tables that total predictive utility (indicated by the AIC) is improved by including the two prior driver record variables (Tables 5 and 6), and that deleting these two predictors (Table 7) results in a greater loss of predictive power (greater increase in the AIC) than does deleting the territorial risk indices (Table 8). This will be discussed in greater detail in a subsequent section of this report.

Question 2 – Does the driving habits and exposure information obtained from the driver survey add any unique contribution to the prediction of total crash involvement beyond that provided by the driver record variables and territorial rate indices?

Having established in the above models the associations between crashes and the driver record and territorial predictors in the larger non-survey sample, the next logical step is to examine the unique contribution of the driving habits and exposure variables obtained from the survey in predicting total crash involvement.

Tables 11-13 summarize the results of the regression equations for predicting 17-month total crashes for the survey sample**.** As before, the equations presented in the tables include different predictors, which were selected using an entry criterion of $p = .10$ in a forward selection regression method**.** Specifically, in Table 11, the driver record variables, the two territorial rate indices, and all survey variables were candidates for inclusion in the equation**.** The equation in Table 12 used the driver record variables and the two territorial indices as potential predictors**.** In Table 13, all survey variables were candidates for possible inclusion in the regression equation.

The best fitting equation is Table 11, which contained the statistically significant driver record variables, territorial indices, and survey responses. This equation produced an AIC of 3,260 with 22 significant ($p \leq .10$) predictors.[6] As was the case with the non-survey sample equations, the most significant predictor was prior 3-year total citation frequency**.** Inspection of several of the parameter estimates from the various equations displayed in the tables show that increased crash frequency is associated with the following:[7]

- Increased prior total citation frequency,
- Being male,
- Being young,
- Increased prior total crash frequency,
- Residing in higher territorial crash rate index areas,

[6] Excluding the intercept.

[7] Using forward selection (the final step showing all entered covariates adjusted for each other) resulted in the final regression parameter estimates for certain indicator variables reflecting orthogonal comparisons. For example, education in Table 11 compares the difference in the average (log) crash rate of 9th to 12th grade to the combined average of the other education groups (e.g., high school+some college, no degree+associate degree+master or higher). Orthogonal comparisons can have greater power than the combined tests of the total dummy variable since the omnibus test tests the average of all possible comparisons, while the orthogonal tests focus on the specific comparisons (e.g., 9th to 12th versus all others).

- Holding a commercial license,
- High weekly mileage,
- Driving aggressively,
- Driving while distracted,
- 20-28 drinking days per month,
- Driving after using alcohol or other illegal drugs during the past 12 months,
- Lower educational level, and
- Lower income level.

Table 11

Poisson Regression Model Predicting 17-Month Total Crashes from All Significant Driver Record, Territorial Risk Indices, and Survey Variables for the Survey Sample

Predictor	Parameter estimate	*t*	*p*	90% confidence interval
Intercept	-4.1609	-11.73	<.0001	-4.7443, -3.5775
Gender	-0.0217	-3.94	<.0001	-0.0774, -0.0044
Age	-0.0115	-3.75	<.0001	-0.0110, -0.0098
Commercial license	0.7343	3.12	.0018	0.4227, 1.0460
Prior 3-year total citations	0.1371	7.59	<.0001	0.0649, 0.2092
Prior 3-year total crashes	0.6081	3.88	<.0001	0.4764, 0.7398
Territorial total crash index	1.1463	3.98	<.0001	0.6726, 1.6199
Driving 8-11 years total	-0.3452	-2.23	.0254	-0.5992, -0.0911
Driving 12-15 years total	-0.3092	-2.04	.0416	-0.5588, -0.0596
Drive 201-300 miles to and from work	0.3796	2.33	.0199	0.1115, 0.6477
Drive on residential streets least often	0.2467	1.74	.0816	0.0137, 0.4798
Avoid driving in bad weather	0.2510	2.73	.0063	0.0999, 0.4021
Reading or sending a text while driving in the past month	0.8170	3.38	.0007	0.4196, 1.2145
Drive aggressively in the past month	0.3638	1.95	.0517	0.0562, 0.6714
Wear headphones while driving in the past month	1.2697	3.05	.0023	0.5840, 1.9554
Watch a video while driving in the past month	1.6113	3.72	.0002	0.8993, 2.3233
Using a GPS while driving in the past month	0.2333	2.26	.0236	0.0638, 0.4029
Income ($25,000-$34,999)	0.3036	2.25	.0254	0.0808, 0.5265
Education (9^{th} to 12^{th} grade, no diploma)	0.3420	2.41	.0162	0.1083, 0.5758
Employed part-time	0.2818	2.22	.0264	0.0730, 0.4907
Beer alcoholic beverage most often drank	0.2499	1.83	.0679	0.0248, 0.4750
20-28 drinking days per month	0.4765	2.41	.0165	0.1504, 0.8027
In the past 12 months, sometimes drove after using marijuana or other illegal drugs	0.8952	2.46	.0140	0.2962, 1.4941

Note. N = 6,548. Likelihood ratio χ^2 versus intercept only = 137. AIC = 3,260. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise. Survey variables coded 1 if risk factor is present; 0 otherwise.

Table 12

Poisson Regression Model Predicting 17-Month Total Crashes from All Significant Driver Record and Territorial Risk Indices for the Survey Sample

Predictor	Parameter estimate	Wald χ^2	*p*	90% Confidence interval	
Intercept	-3.7314	122.73	<.0001	-0.2855,	-3.1774
Gender	-0.0336	3.87	.0500	-0.0784,	-0.0013
Age	-0.0056	3.92	.0477	-0.0103,	-0.0009
Commercial license	0.4693	12.40	.0004	0.2566,	0.9819
Prior 3-year total citations	0.1891	60.66	<.0001	0.0943,	0.2379
Prior 3-year total crashes	0.3522	14.47	<.0001	0.1909,	0.6538
Territorial total crash index	1.2363	18.06	<.0001	0.7578,	1.7148

Note. N = 6,548. Likelihood ratio χ^2 versus intercept only = 39. AIC = 3,325. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise.

Table 13

Poisson Regression Model Predicting 17-Month Total Crashes from All Significant Survey Variables for the Survey Sample

Predictor	Parameter estimate	*t*	*p*	90% Confidence interval	
Intercept	-2.9787	-22.63	<.0001	-2.7621,	-2.7621
Driving fewer than 4 days per week	-0.4780	-1.86	.0629	-0.9001,	-0.0553
Driving 21 or more hours per week	0.4634	3.19	.0015	0.2241,	0.7028
Driving over 150 miles per week	0.3384	2.32	.0204	0.0984,	0.5783
Driving 8-11 years	-0.4389	-2.96	.0031	-0.6830,	-0.1948
Driving 12-15 years	-0.3681	-2.48	.0133	-0.6127,	-0.1236
Drive on residential streets least often	0.4687	2.69	.0071	0.1824,	0.7549
Drive on rural roads least often	0.2490	1.95	.0513	0.0389,	0.4592
Avoid driving in bad weather	0.2641	2.82	.0048	0.1100,	0.4182
Drive heavy commercial vehicle most often	0.5937	2.02	.0436	0.1098,	1.0776
Reading or sending a text while driving in the past month	0.7658	3.04	.0023	0.3519,	1.1798
Drive aggressively in the past month	0.3929	2.04	.0411	0.0765,	0.7093
Wear headphones while driving in the past month	1.2640	2.97	.0030	0.5630,	1.9650
Watch a video while driving in the past month	1.4405	3.20	.0014	0.7006,	2.1804
Using a GPS while driving in the past month	0.2327	2.23	.0258	0.0610,	0.4044
Income ($25,000-$34,999)	0.3392	2.45	.0152	0.1102,	0.5682
Education (9th to 12th grade, no diploma)	0.3327	2.30	.0216	0.0946,	0.5709
Employed part-time	0.2971	2.34	.0194	0.0880,	0.5063
Beer alcoholic beverage most often drank	0.2569	1.90	.0582	0.0339,	0.4799
20-28 drinking days per month	0.5328	2.50	.0137	0.1798,	0.8858
In the past 12 months, sometimes drove after using marijuana or other illegal drugs	0.9030	2.38	.0174	0.2783,	1.5278

Note. N = 6,548. Likelihood ratio χ^2 versus intercept only = 29. AIC = 3,348. Gender coded 1 if female; 0 otherwise. Commercial license coded 1 if commercial driver; 0 otherwise. Survey variables coded 1 if risk factor is present; 0 otherwise.

Question 3 – What is the relative contribution of the various subsets of predictor variables?

Although the relative importance of individual predictors can be assessed by examining predictor-specific parameters (e.g., the confidence interval and the standardized regression coefficient), an accurate assessment of a variable's unique predictive ability and contribution requires an alternative approach. When predictors are correlated with each other, as they are in the present data, the unique contribution of a predictor (or subset of predictors) can only be

accurately assessed by deleting the predictor (or subset of predictors) from the equation, running the new equation, and calculating the reduction in overall predictability between the two. As discussed above, the AIC is used in the present study to assess the efficiency and accuracy of a model in predicting total crash frequency. When comparing models that include different subsets of predictors, the model with the lowest AIC is considered the "best" one for the data at hand. The increase in the AIC resulting from deleting a predictor (or set of predictors) from the model represents that predictor's (or predictor set's) unique contribution to the overall prediction achieved by the equation.

Table 14 displays the AIC values from the non-survey and survey samples' regression equations presented in the prior tables. The following points can be made from these values:

- Each of the three predictor sets (i.e., person-centered driver record variables, territorial rate indices, and person-centered survey variables) makes a unique contribution to crash prediction.
- Comparing F with G and H with I indicates that person-centered driving record variables are better than territorial indices as crash predictors, although the difference is not large.
- Similarly, comparing B with C indicates that person-centered driving record predictors are better than the survey predictors in their ability to predict counts of total traffic crashes, although the difference is not large.

Table 14

The Contribution of Various Subsets of Multiple Poisson Regression Predictors to the 17-Month Total Crash Criterion as Measured by the Akaike Information Criterion for the Non-Survey and Survey Samples

Variables in the Regression Equation	Non-survey sample (N = 268,480) AIC	Survey sample (N = 6,548) AIC
A. All significant driver record, survey, and territorial risk indices (Table 11)	-	3,260
B. Equation A with all survey variables excluded (Table 12)	-	3,325
C. Equation A with all driver record variables excluded (Table 13)	-	3,348
D. All driver record and territorial risk indices included (Table 5)	91,905	-
E. All significant driver record and territorial crash index included (Table 6)	91,902	-
F. Equation E (all significant driver record and territorial risk indices predictors) with the two prior driving record variables excluded (Table 7)	92,282	-
G. Equation E (all significant driver record predictors) with the two territorial risk indices excluded (Table 8)	92,008	-
H. Only significant prior total crash and total convictions predictors included (Table 9)	92,180	-
I. Only two significant territorial risk indices included (Table 10)	92,460	-

Question 4 – What is the utility of applying the regression equations to predict individual crash involvement for individual drivers with selected profiles?

To illustrate how predicted crash involvement varies as a function of driver profile, the "best" non-survey sample equation (shown in Table 6) was applied to four hypothetical groups of California drivers having the characteristics displayed in Table 15.

Table 15 represents one example of crash prediction for drivers classified within a 2x2 matrix of driver characteristics. Drivers in category A1 reside in a low-crash zip code area and have a clean prior driver record, while those in B1 reside in a low-crash zip code but have a bad prior record. Similarly, drivers in A2 reside in a high-crash zip code but have a clean prior record, while those in B2 reside in a high-crash zip code and have a bad prior record. In this example, the low- and high-crash zip codes (for San Ysidro and Hayward, respectively) are selected to represent very low and very high crash risk territories. The group crash expectancies per 1,000 drivers computed from the equation are displayed in Table 16.

Table 15

Characteristics of Four Hypothetical Driver Groups Used to Generate Total Crash Prediction

	Driver group			
Variable	(A_1) Low crash zip code, clean prior record	(B_1) Low crash zip code, bad prior record	(A_2) High crash zip code, clean prior record	(B_2) High crash zip code, bad prior record
Territorial crash rate index	0.455[1]	0.455[1]	1.503[2]	1.503[2]
Prior 3-year total citations	0	4	0	0
Prior 3-year total crashes	0	2	0	2

Note. All other variables in the regression model are fixed at their mean through the use of centering.

[1]The San Ysidro zip code used for these calculations is 92173.

[2]The Hayward zip code used for these calculations is 94557.

These predicted crash frequencies indicate that higher risk scores on each factor (territory and prior driving record) predict increased crash expectancies. However, the two factors are not equal in this regard. For these hypothetical profiles, having a high-risk prior driver record predicts more subsequent crash involvements than does residing in a high-risk zip code, predicting 31 more crashes per 1,000 drivers (66 vs. 35). This can be seen by comparing the low crash rate zip code, high prior driver record group (87 predicted crashes) to the high crash rate

zip code, low prior record group (56 predicted crashes). This scenario demonstrates that clean-record drivers residing in the highest risk areas represent a lower crash risk than do bad-record drivers who reside in lower risk areas. These results reflect the relative predictive importance of the driver record variables in the regression equations (compare Equations H and I in Table 14).

The results in Table 16 demonstrate how the regression equations modeled in this project can be used to calculate crash expectancies for highly contrasted groups. The findings have value in that they provide an actuarial basis for setting insurance premium rates. Most people would probably agree that drivers in group A_1 should not be charged the same amount for auto insurance as drivers in group B_2. However, it is important to note that while group crash rates can be predicted with a reasonable degree of accuracy, the same cannot be said for predicting which individual drivers in the groups will be involved in future crashes. A majority of drivers do not have values this extreme (most having no crashes) and, therefore, are more difficult to differentiate statistically. Additionally, the results in the example are averages (crashes per 1,000 drivers) and therefore do not apply to every individual driver.

Table 16

Expected Number of Total Crashes per 1,000 Drivers as a Function of Selected Territory and Prior Driver Record Variables

	Low-risk driver	High-risk driver	Net difference
Low-risk territory	21	87	66
High-risk territory	56	106	49
Net difference	35	19	-

To illustrate the degree of accuracy in predicting individual crash expectancy using the regression equations with differing subsets of predictors, four-fold contingency tables were generated showing the relationship between each individual's predicted and actual crash-involvement frequency.[8] The principles underlying this classification technique and interpretation of results emanating from it were discussed earlier. The survey sample was used to generate the prediction tables for this demonstration.

[8] It should be noted that use of a simple dichotomy (0 versus 1 or more) results in a slight under-estimate of the utility of the prediction model because multiple crash involved drivers are treated in the analysis as if they are single-crash drivers. Although the regression equations are based on the full range of crash counts, this dichotomous tabular presentation was adopted for simplicity.

Two such tables were generated and are presented here for demonstration purposes. Table 17 was constructed from the equation containing the statistically significant driver record variables and territorial crash index (Table 12). Table 18 was constructed from the equation containing the statistically significant set of driver record variables, the territorial crash index, and the habit/exposure variables (Table 11). For comparative purposes, both tables used a prediction cutoff score that equalized the marginal distributions. Use of such a cutoff score has the property of producing equal numbers of false negative and false positive predictions.[9]

Table 17 shows a statistically significant association ($\chi 2 = 138.84$, $p < 0.001$) between predicted and actual crash involvement. Persons predicted to have crashes are approximately 3.15 times more likely to actually have had crashes than are those predicted to be crash-free ($1.70 \div 7.94 = 21.41\%$ versus $6.25 \div 92.06 = 6.79\%$). However, the equation failed to correctly predict the majority of crash-involved drivers, as evidenced by the low true-positive rate of 21.35%. Although the false-negative rate ($6.25 \div 92.06 = 6.79\%$) appears low, this percentage of misclassification represents the majority of the 7.94% of the sample who were truly crash-involved.

The phi coefficient and odds ratio, shown at the bottom of each table, are commonly used indices for quantifying the degree of association in contingency tables. The phi coefficient represents the Pearson correlation between the actual and predicted crash-status categories. The odds ratio refers to the relative odds of being crash-involved as a function of predicted crash category.

In Table 17, the odds of predicted crash-involved drivers actually having a crash as opposed to not actually having a crash are $(1.70\% \div 6.25\%) = 0.272$. The same odds for the predicted crash-free group are $(6.25\% \div 85.81\%) = 0.073$. The ratio of these two odds (i.e., the odds ratio) is 3.73. The fact that this ratio is greater than 1.0 indicates that the odds of actually having a crash did vary as a function of the predicted score for this sample. The fact that the odds ratio and phi coefficient in Table 17 are both of modest size indicates that the prediction equation has low accuracy in predicting which individual drivers will be crash-involved. This is evident from the

[9] This approach has the advantage of giving equal weight to the two types of errors and maximizing the expected value of the phi coefficient. However, there are situations in which one type of prediction error is more important than the other, and different prediction-cutoff scores can be used. If one wishes to minimize the proportion of crash-free drivers erroneously predicted to have crashes (false positives), one can do so by increasing the cutoff score used to predict whether a driver is crashed-involved. If one wishes to minimize the proportion of crash-involved drivers who are erroneously predicted to be crash-free (false negative error), the cutoff score can be decreased. Unfortunately, the two errors are reciprocally related. That is, if the cutoff score is lowered to reduce false negatives, the proportion of false positives will be increased. The interested reader is referred to Gebers and Peck (2003a) for an application of this technique.

high false-positive rate (6.25 ÷ 7.94 = 78.72%) and the fact that the equation misclassifies the majority of crash-involved drivers.

Table 17

17-Month Actual Versus Predicted Crash-Involvement Status for the Poisson Regression Model Using Significant Driver Record and Territorial Crash Index Predictors for the Survey Sample

	Predicted crash status		
Actual crash status	Crash-involved	Crash-free	Total
Crash-involved	111 (1.70%)	409 (6.25%)	520 (7.94%)
Crash-free	409 (6.25%)	5,619 (85.81%)	6,028 (92.06%)
Total	520 (7.94%)	6,028 (92.06%)	6,548 (100.00%)
Percent correctly classified	21.35	93.21	

Note. $N = 6{,}548$. A cutpoint of 0.1542387 was used to equalize the marginals. The χ^2 is 138.84 ($p < .0001$). The phi coefficient is 0.1456. The odds ratio is 3.73.

Table 18 illustrates the modest gain in individual prediction obtained by adding the statistically significant habit/exposure variables to the equation containing the driver record and territorial crash index predictors. The true-positive rate increases slightly, to 25.19%, while the false negative rate decreases slightly, to 6.45%. However, with the use of the cutpoints resulting in equal marginal values, the more important indicators of the gain in individual prediction are the Chi-square values, the phi-coefficients, and the odds ratios. The reader will note that these indices increase significantly ($p \leq .10$) with the inclusion of the habit/exposure variables to the equation (138.84 to 229.94, 0.1456 to 0.1874, and 3.73 to 4.88, respectively). As was the case with the entries in Table 17, the continuing high false-positive rate in Table 18 (5.94 ÷ 7.94 = 74.81%) and misclassification of the majority of crash-involved drivers indicates a low accuracy in predicting which individuals will be crash-involved.

Table 18

17-Month Actual Versus Predicted Crash-Involvement Status for the Poisson Regression Model Using Significant Driver Record, Territorial Crash Index, and Habit/Exposure Predictors for the Survey Sample

	Predicted crash status		
Actual crash status	Crash-involved	Crash-free	Total
Crash-involved	131 (2.00%)	389 (5.94%)	520 (7.94%)
Crash-free	389 (5.94%)	5,639 (86.12%)	6,028 (92.06%)
Total	520 (7.94%)	6,028 (92.06%)	6,548 (100.00%)
Percent correctly classified	25.19	93.55	

Note. $N = 6{,}548$. A cutpoint of 0.172742 was used to equalize the marginals. The χ^2 is 229.94 ($p < .0001$). The phi coefficient is 0.1874**.** The odds ratio is 4.88.

DISCUSSION

Study Limitations

Before discussing the study's results and offering recommendations, it is informative to consider the data and statistical limitations present in the analyses. Four such limitations are presented below.

1. In addition to the earlier stated problem associated with the 17-month total crash criterion period, it is perhaps self-evident that the crash data on which the regressions are based are limited to crashes reported to the Department. By law (at the time of the writing of this report), all traffic crashes involving fatalities or injuries, or property damage in excess of $750, must be reported to the Department. Therefore, most crashes under the $750 threshold would not be reported to the Department, though they may still be reported on claims to insurance companies. A substantial number of property-damage-only crashes above the reporting threshold also go unreported to the Department. The probable effect of non-reporting is to decrease the predictive accuracy or reliability of the regression equations compared to what would be obtained under complete reporting. There is also some potential for bias in the regression parameters, since non-reporting is likely to be non-random relative to the individual characteristics of the driving population.

2. In any type of regression analysis, there are critical assumptions that (1) the predictor variables were measured without error (i.e., are true scores having perfect reliability) and (2) the observations were selected randomly from the population of interest. The first assumption chiefly relates to situations where one uses the regression weights to make inferences about the cause-effect relationships underlying the variables in the equation. As stated above, this was not the purpose of the present analyses. Therefore, there is no real concern over this assumption other than recognizing that measurement error often attenuates regression coefficients and can result in an underestimation of the relationship. The second assumption (random sampling) was met by the data in the case of the non-survey sample analyses. However, this assumption was violated by the survey analyses, due to potential self-selection bias as a result of non-response to the survey. The primary impact of self-selection bias is to render the significance level of tests of hypotheses ambiguous, since it is no longer clearly evident as to what population is being generalized. That is, the responses may no longer be representative of the survey sample and, therefore, may not generalize to

the general driving population from which the survey sample was randomly selected. However, the regression results are still valid descriptions of the relationships among variables for the respondent sample. As described in the Methods section of this report, weights were employed as a mechanism to adjust for non-response. Such a strategy can in no way guarantee that the results are representative of the target population consisting of licensed California drivers. However, there exists justification for some degree of confidence in the weighting strategy utilized in the present study as the direction and magnitude of the parameter estimates in driver record and territorial risk indices equations for the non-survey and survey samples are similar (Tables 6 and 12, respectively).

3. The survey response rate was 31% (increasing to 38% when removing undeliverable survey forms from the total number of surveys in the calculation), which resulted in only 6,548 cases being available for the analyses. Although this is consistent with response rates reported in the literature for mail surveys, (e.g., Kohut, Keeter, Doherty, Dimock, & Christian, 2012; Schoeni, Stafford, McGonangle, & Andreski, 2013; Steeh, Kirgis, Cannon, & DeWitt, 2001), it is only half the 60% response rate obtained from the 1983 California driver survey (Frincke & Ratz, 1984) and far below the 72% response rate from the 1975 California driver survey (Peck & Kuan, 1982).[10] On the basis of a similar sampling strategy to these earlier California driver surveys, it was anticipated that the current survey would achieve a response rate somewhere between 50% and 60%. With a larger sample, the plan was to evaluate selected interactive or moderating relationships between miles driven, drinking habits, age, gender, and prior driving record after adjusting for all other measured driving habits and exposure-related variables. Several of these interactive relationships were tested in a series of preliminary regression models; however, none of them met the level of statistical significance ($p \leq .10$) used for the study. Post-hoc power analyses indicated that the power associated with these interactive relationships were at or well below .60.[11]

4. Obviously, the survey items represent self-reported data. The primary advantage of the self-reported survey items used in the present study is that they query drivers on information (e.g., weekly mileage, use of cell phones, and alcohol and drug usage) that is not available on the driving record. However, self-reported data also have specific disadvantages due to the way

[10] The 1983 driver survey used a two-wave mailing approach. The 1975 driver survey used a three-wave mailing approach.

[11] Statistical power is the probability that a statistical test will produce a significant result if there is a real or reliable relationship between variables (or sets of variables) and the criterion measure. A value of .80 or above is generally regarded as acceptable power.

that individuals tend to behave. For example, self-reported answers may be exaggerated, respondents may be too embarrassed to reveal private details such as alcohol usage related to driving practices, and respondents tend to reply in a manner they think is more socially desirable or will be viewed more favorably by a governmental entity such as the Department of Motor Vehicles. To the extent possible, the procedures used in the present study attempted to reduce such response distortions. The wave one and wave two contact letters emphasized that participation was voluntary, responses were confidential, and no licensing actions would be taken. In addition, no personally identifying information (e.g., name or driver license number) appeared on the survey form.

Conclusion

This study provides evidence that prior driver record, driver habits and exposure variables, and territory are significantly related to the likelihood of subsequent crash involvement among survey respondents. Specifically, the results indicate that prediction of traffic crashes is enhanced with the inclusion of driving habits and exposure information and territorial risk indices in the regression models.

Several findings from the equations presented in this paper are worthy of note and are discussed below.

The equations for the non-survey sample demonstrated that both prior driving record and territory are significant predictors of subsequent crash involvement. Of the two sets of predictors, the prior driver record variables are better predictors of crash frequency than are the territorial variables. Prior total citation frequency is a better predictor of subsequent crashes than is prior total crash frequency.

Although this study was not explicitly designed to address automobile insurance rating issues, it is interesting that while the territorial crash risk index achieved statistical significance in the non-survey and survey regression equations containing prior driver record and/or habits and exposure variables, the territorial composite index did not. The territorial composite index, as discussed earlier in this report, was previously used by the insurance industry in generating assigned area risk ratings. This lack of statistical significance, however, was not unexpected, since the majority of the regression equations contained two predictors functionally related to the composite index, namely the territorial crash rate index and prior total traffic citation frequency. The territorial crash rate index is less correlated with prior citation frequency and has a

somewhat higher association with the individual crash count criterion**.** Therefore, it did a better job predicting crash frequency than did the territorial composite index in the context of the equations evaluated in the present study.

Similarly, it was demonstrated from the use of the 2x2 classification tables that crash frequency (and loss expectancy) for individual drivers cannot be accurately predicted from any combination of factors. Therefore, any graduated insurance premium structure would result in a significant number of individuals paying more than indicated by their actual future crash losses, and many others paying less than indicated by their actual losses. This effect is an inevitable consequence of the large random component underlying the occurrence of crashes and does not necessarily mean that drivers are being improperly or incorrectly classified. However, known group-risk differentials related to driver record and territory do exist and are sufficient to justify some degree of premium variations based on these variables.

In assessing equations using driving habits and exposure items from the survey sample, the results provide clear evidence that drivers who report driving while distracted and driving aggressively have a higher risk of crash involvement. For example, the parameters obtained from the survey sample equation containing driver record, territorial crash rate index, and the driving habits and exposure variables indicated that reading or sending a text message, using a GPS device, wearing headphones, watching a video while driving, and driving aggressively were all associated with a significant increase in crash involvement even after accounting for variance associated with driver record and territory.

These findings reinforce the efforts being conducted in California as part of the state's Strategic Highway Safety Plan (SHSP) started in 2005. California's SHSP is a statewide, comprehensive, data driven effort to reduce fatalities and serious injuries on public roadways. At the time of the writing of this report, SHSP has 17 challenge areas, two of which (areas 10 and 17) attempt to reduce speeding/aggressive driving and distracted driving, respectively, by focusing on behavioral, infrastructure, and technology services (e.g., cell-phone disabling, electronic message board warnings, etc.).

The finding that the linear term for mileage is related to crash involvement is consistent with prior studies. Although the linear term to mileage failed to reach the level of statistical significance used for this study and, therefore, was not presented in the Results, there was some evidence from the preliminary models that the relationship between weekly miles driven and traffic crash involvement may be quadratic in nature. That is, there was an observed "dip" in the

crash risk curve for drivers reporting higher mileage. This nonlinear relationship between miles driven and crash risk has been reported in several studies (e.g., Janke, 1991; Massie, Green, & Campbell, 1997; Mercer, 1989). These studies also found several other variables associated with average miles driven. For example, those who drive few miles tend to accumulate a higher proportion of miles in an urban setting on local streets. Drivers reporting higher average miles tend to drive a higher portion of their miles on rural interstates or highways. Because of differences in their design and traffic density, rural interstates and highways generally have a lower risk of crash per mile than do urban roadways. Although the driving habits and exposure variables included in the present study would have statistically controlled for the influence of some of these differences, it certainly was not possible to fully account for the effects of differences in the driving environment as they relate to the prediction of traffic crashes. In any event, when using mileage as a covariate in studies on crash risk, the nonlinearity in the relationship between mileage and crash rates should be investigated whenever possible.

The traffic safety literature contains compelling evidence that alcohol and drug use are major factors in traffic crash causation. In the present study, it was demonstrated that self-reported heavy drinking (i.e., 20 – 28 drinking days per month) and use of marijuana or other illegal drugs prior to operating a motor vehicle were associated with an increase in crashes. However, it should be noted that of all the items on the survey, the items measuring alcohol and/or drug use are likely affected most by response bias. For example, problem drinkers tend to underestimate their alcohol consumption (Blomberg, Peck, Moskowitz, Burns, & Florentino, 2005). To the extent that this occurred in the present study, the relationship between crashes and alcohol/drug use could be severely attenuated and/or inaccurate.

Recommendations

The following recommendations are offered based on the findings in this study.

1. Although the present study's emphasis was on modeling total crash counts by way of multiple regression equations for the non-survey and survey samples, the survey contains ample information for additional analyses, and the survey responses could be used to further investigate the relationships among other driver record and driving habits indices**.** Specifically, it is recommended that a subsequent report containing a series of descriptive based contingency tables be produced**.** Such contingency tables could examine bivariate relationships between pairs of variables such as exposure and education, exposure and

occupational status, and exposure and crash group. The contingency tables could be extended to display multi-way relationships such as respondents within education, crash, and exposure levels. Such an effort (currently being planned by the California DMV) should result in additional and more complex profiles of crash-free and crash-involved drivers and perhaps to a better understanding of the correlates of crash risk.

2. As noted above, the delay in updating police reported crashes on the Department's Driver Record Master file necessitated a "shifting" in the criterion window from a planned 12-month period to a 17-month period. The shifting likely may have resulted in a less temporal relationship between reported driving habits and the crash criterion. When the updating of crashes is complete, it is recommended that the regression models used in the present study be replicated on the planned 12-month criterion period and that the regression parameters be examined for any change in direction or magnitude.

3. The present study assessed only the association between prior driver record, territorial indices, and driving habits and exposure variables with the total crash criterion. The existence of these data invites the use of other criteria of interest. Specifically, it is recommended that these data be used to model two additional criterion measures (1) total traffic citations (i.e., convictions, failure-to appear violations, and traffic violator school dismissals) and (2) had-been-drinking CHP reported crashes (i.e., crashes in which the driver was deemed by the reporting officer as had-been-drinking and obviously impaired).

4. Historically, driver record data are commonly aggregated into multi-year (e.g., 2-year or 3-year) predictor and criterion periods for use in regression models. A different or supplemental, and perhaps even more appropriate, modeling strategy (requiring the use of a survey sample larger than the one used in the current effort) would be to treat separate yearly counts of driver record entries (e.g., crashes) as repeated measures in the regression models. That is, fixed-effects regression methods could be applied to these data by treating the annual counts as panel data – the most common type of longitudinal data – consisting of measurements of predictor and response variables at two or more points in time for many individuals. Panel data have two major attractions: (1) the ability to control for unobserved variables and (2) the development of models that make it possible to determine which variable causes the other if they are truly causally related. This approach uses only within-individual variation to estimate the regression coefficients. Fixed-effects models can be applied to a wide variety of statistical techniques that are well suited to driver record data, such as logistic regression, modeling of count data, and survival analysis.

5. The fact that self-reported aggressive driving was significantly associated with crash frequency in each model containing survey items should encourage the Department to conduct its planned empirical study of aggressive driving**.** This study would analyze the historical driving records of a large representative sample of California drivers to determine what patterns and combinations of driving behaviors thought to be aggressive in nature would be good predictors of having a future crash risk greater than that posed by *prima facie* negligent operators in California.[12] Establishing that chronic aggressive driving is associated with high future crash risk would justify administering intervention actions, such as license suspension, earlier than would otherwise occur under the Department's existing post licensing control system**.** Increasing the severity of sanctions against high-risk, aggressive drivers is also supported explicitly in the California Strategic Highway Safety Plan. Such a study is currently being conducted by the California Department of Motor Vehicles (Wu, in press).

6. Given the statistically significant relationship presented in this study between crash involvement and self-reported distracted driving, the Department continued its efforts to evaluate the relationship between cell phone use while driving and traffic crash involvement as reported in Limrick, Lambert, and Chapman (2014). Specifically, this finding lead to further research (funded by an Office of Traffic Safety Grant) establishing that distracted driving violations in combination with negligent operator treatment points identified higher risk drivers for potential licensing actions than negligent operator points alone (Lambert, Fox, & Camp, 2017.) This finding substantiated the consistent and reliable association between distracted driving and traffic crash risk displayed in the regression weights from Table 11.

[12] A *prima facie* negligent operator is defined as a driver whose record shows four or more points in 12 months, six or more points in 24 months, or eight or more points in 36 months. The licensing actions for these drivers consist of a 12-month probationary period with a 6-month suspension component.

REFERENCES

AAA Foundation for Traffic Safety. (2008). *2009 Traffic safety culture index.* Washington, D. C: AAA Foundation for Traffic Safety.

AAA Foundation for Traffic Safety. (2016). *2015 traffic safety culture index*. Washington, D. C: AAA Foundation for Traffic Safety.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

Allison, P. D. (2009, Spring). *Course notes on Missing Data*. Personal Collection of Paul D. Allison.

Blomberg, D., Peck, R. C., Moskowitz, H., Burns M., & Fiorentino, D. (2005). *Crash risk of alcohol involved driving: A case control study*. Stamford, Connecticut: Dunlap and Associates, Inc.

Boyer, M., Dionne, G., & Vanasse, C. (1990). *Econometric models of accident distributions*. Montreal, Canada: University of Montreal, Center for Research on Transportation.

Cochran, W. G. (1977). *Sampling techniques, third edition*. New York, NY: John Wiley & Sons.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika, 70,* 269-274.

Davis, C.S. (1990). The DeKalb County, Georgia, Driver Education Demonstration Project: Analysis of its long-term effect. Washington, DC: U.S Department of Transportation, National Highway Traffic Safety Administration.

EKOS Research Associates Inc. (2007). *Impaired driving survey for Transport Canada/MADD Canada*. Ontario, Canada: Transport Canada Road Safety and Motor Vehicle Regulation Directorate Road Users Division.

Famoye, F., & Singh, K. P. (2006). Zero inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science, 4*, 117-130.

Frincke, K., & Ratz, M. (1984). *California driver survey: The habits and opinions of drivers on selected traffic safety related issues* (Report No. 92). Sacramento, CA: California Department of Motor Vehicles.

Gebers, M. A. (1998). Exploratory multivariable analyses of California driver record accident rates. *Transportation Research Record, 1635,* 72-80.

Gebers, M. A. (1999). *Strategies for estimating driver accident risk in relation to California's negligent-operator point system* (Report No. 183). Sacramento, CA: California Department of Motor Vehicles.

Gebers, M. A. (2001). *The contribution of driving exposure in the prediction of traffic accidents*. Sacramento, CA: California Department of Motor Vehicles.

Gebers, M. A. (2003). *An inventory of California driver accident risk factors* (Report No. 204). Sacramento, CA: California Department of Motor Vehicles.

Gebers, M. A., & Peck. R. C. (2003a). *Development and evaluation of a risk management strategy for reducing crash risk* (Report No. 202). Sacramento, CA: California Department of Motor Vehicles.

Gebers, M. A., & Peck, R. C. (2003b). Using traffic conviction correlates to identify high accident-risk drivers. *Accident Analysis and Prevention, 35,* 903-912.

Grogger, J. (1990). The deterrent effect of capital punishment: An analysis of daily homicide counts. *Journal of the American Statistical Association, 85,* 295-302.

Gruenewald, P., & Nephew, P. (1994). Drinking in California: Theoretical and empirical analyses of alcohol consumption patterns. *Addiction, 89,* 707-723.

Hennessy, D. F. (1995). *Vision testing of renewal applicants: Crashes predicted when compensation for impairment is inadequate* (Report No. 152). Sacramento, CA: California Department of Motor Vehicles.

Izrael, D., Hoaglin, D. C., & Battaglia, M. P. (2004). To rake or not to rake is not the question anymore with the enhanced raking macro. *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*, Paper 275.

Janke, M. K. (1991). Accident, mileage, and the exaggeration of risk. *Accident Analysis & Prevention, 23,* 183-188.

Kelsey, S. L., & Janke, M. K. (2005). *Pilot educational outreach to high-risk elderly drivers*. (Report No. 213). Sacramento, CA: California Department of Motor Vehicles.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., & Christian, L. (2012). *Assessing the representativeness of public opinion surveys*. Washington, DC: Pew Research Center.

Lambert, A., Fox, M., & Camp, B. J. (2017). *Assigning points for cell phone violations: Effects and implications* (Report No. 253). Sacramento, CA: California Department of Motor Vehicles.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. W., & McLachlan, G. J. (2006). Multi-level zero-inflated Poisson regression modeling of correlated count data with excess zeros. *Statistical Methods in Medical Research, 15,* 47-61.

Limrick, K., Lambert, L., & Chapman, E. (2014). *Cellular phone distracted driving: a review of the literature and summary of crash and driver characteristics in California* (Report No. 248). Sacramento, CA: California Department of Motor Vehicles.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A, 44*(5), 291-305.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention, 37*, 35-46.

Massie, E. L., Green, P. E., & Campbell, K. L. (1997). Crash involvement rates by driver gender and the role of average annual mileage. *Accident Analysis & Prevention, 19*, 675-685.

Mercer, G. W. (1989). Traffic accident and convictions: Group total versus rate per kilometer driven. *Society for Risk Analysis, 9*, 71-77.

Peck, R. C., & Gebers, M. A. (1992). *The California driver record study: A multiple regression analysis of driver record histories from 1969 through 1982*. Sacramento, CA: California Department of Motor Vehicles.

Peck, R. C., Gebers, M. A., Voas, R. B., & Romano, E. (2008). The relationship between blood alcohol concentration (BAC), age, and crash risk. *Journal of Safety Research, 39*, 311-319.

Peck, R. C., & Kuan, J. K. (1982). *A statistical model of individual accident risk prediction using driver record, territory and other biographical factors* (Report No. 84). Sacramento, CA: California Department of Motor Vehicles.

Peck, R. C., & Kuan, J. K. (1983). A statistical model of individual accident risk prediction using driver record, territory, and other biographical factors. *Accident Analysis and Prediction, 15*, 371-393.

Pedhazur, E. J. (1973). *Multiple regression in behavioral research: Explanation and prediction* (2nd edition). New York, NY: Holt, Rinehart, and Winston.

SAS Institute Inc. (2009). *SAS/STAT® 9.2 user's guide second edition*. Cary, NC: SAS Institute Inc.

Schoeni, R., Stafford, F., McGonangle, K., & Andreski, P. (2013). Response rates in national panel surverys. *The annals of the American Academy of Political and Social Science, 645*(1), 60-87.

Steeh, C. N., Kirgis, B., Cannon, B., & DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of twentieth century. *Journal of Official Statistics, 17*(2), 277-247.

Wu, V. (In press). *Speeding and aggressive driving risk factors and interventions*. Sacramento, CA: California Department of Motor Vehicles.

APPENDICES

Appendix A

2010 California Driver Survey

2010 CALIFORNIA DRIVER SURVEY

Please check only **one** answer box for each question unless indicated otherwise.

1. How would you rate the service you have received from DMV?

   1 ☐ very poor 2 ☐ poor 3 ☐ average 4 ☐ good 5 ☐ excellent

2. How many **days** do you normally drive each week?

   1 ☐ 1 2 ☐ 2 3 ☐ 3 4 ☐ 4 5 ☐ 5 6 ☐ 6 7 ☐ 7

   8 ☐ Check here if you do not drive in most weeks

3. How many **hours** do you normally drive each week?

   1 ☐ 1 2 ☐ 2-4 3 ☐ 5-9 4 ☐ 10-14 5 ☐ 15-20 6 ☐ 21 or more

4. How many **miles** do you normally drive each week?

   1 ☐ 0 – 9 2 ☐ 10 – 20 3 ☐ 21 – 50 4 ☐ 51 – 150 5 ☐ 151 – 250

   6 ☐ 251 – 350 7 ☐ 351 – 500 8 ☐ 501 – 1,000 9 ☐ Over 1,000

5. What type of vehicle do you drive **most** often?

   1 ☐ Car 2 ☐ Pickup truck 3 ☐ Sports utility vehicle

   4 ☐ Minivan 5 ☐ Heavy commercial truck 6 ☐ Motorcycle

   7 ☐ Other

6. How many years have you been driving?

   1 ☐ 0-3 2 ☐ 4-7 3 ☐ 8-11 4 ☐ 12-15

   5 ☐ 16-19 6 ☐ 20 or more

7. How many years have you driven **in California**?

1 ☐ 0-3 2 ☐ 4-7 3 ☐ 8-11 4 ☐ 12-15

5 ☐ 16-19 6 ☐ 20 or more

8. What type of driving do you **most** often do?

1 ☐ To and from work 2 ☐ Recreational

3 ☐ Errands (shopping, appointment) 4 ☐ On the job

5 ☐ Trips out of town 6 ☐ Other

9. On what type of roadway do you drive **most** often?

1 ☐ Residential streets 2 ☐ Rural roads 3 ☐ Freeways

4 ☐ Non-residential city streets 5 ☐ Other

10. On what type of roadway do you drive **least** often?

1 ☐ Residential streets 2 ☐ Rural roads 3 ☐ Freeways

4 ☐ Non-residential city streets 5 ☐ Other

11. In what situations do you **avoid driving** (check all that apply)?

1 ☐ None 2 ☐ At night 3 ☐ On freeways 4 ☐ In bad weather

5 ☐ In unfamiliar areas 6 ☐ During rush hour

12. How many total miles do you drive **to and from work** each week?

1 ☐ 1-50 2 ☐ 51-100 3 ☐ 101-200 4 ☐ 201-300

5 ☐ over 300 6 ☐ don't drive to work

13. During the past month, did you do any of the following **while driving**? (Check all that apply).

1 ☐ Feel drowsy 2 ☐ Use a cell phone 3 ☐ Read

4 ☐ Eat or drink 5 ☐ Groom yourself (comb your hair, apply makeup, etc.)

6 ☐ Drive aggressively 7 ☐ Been emotionally upset

8 ☐ Read or send a text message

9 ☐ Use an MP3 player, IPOD, or other personal electronic device

10 ☐ Wear headphones 11 ☐ Watch a video

12 ☐ Use a global positioning system (GPS) 13 ☐ Adjust a video player

14. How many total miles do you drive in a typical week **as part of your job**?

1 ☐ 0-50 2 ☐ 51-100 3 ☐ 101-200

4 ☐ 201-300 5 ☐ over 300 6 ☐ don't drive on the job

15. What was the **combined gross** (pre-tax) income for all members of your household in 2009?

1 ☐ Less than $25,000 2 ☐ $25,000 – $34,999 3 ☐ $35,000 – $49,999

4 ☐ $50,000 – $74,999 5 ☐ $75,000 – $99,999 6 ☐ $100,000 – $149,999

7 ☐ $150,000 – $199,999 8 ☐ $200,000 or more

16. What is your marital status?

1 ☐ Now married 2 ☐ Widowed 3 ☐ Divorced

4 ☐ Separated 5 ☐ Never married

17. What is your highest level of completed education?

1 ☐ Less that 9th grade 2 ☐ 9th to 12th grade, no diploma

3 ☐ High school graduate 4 ☐ Some college, no degree

5 ☐ Associate's degree 6 ☐ Bachelor's degree

7 ☐ Master's degree or higher 8 ☐ Other

18. What **best** describes your employment status? Check all that apply.

1 ☐ Employed full-time 2 ☐ Employed part-time 3 ☐ Self-employed

4 ☐ Student 5 ☐ Homemaker 6 ☐ Retired

7 ☐ Not employed

19. Which type of alcoholic beverage do you **most** often drink?

1 ☐ Beer 2 ☐ Wine 3 ☐ Liquor 4 ☐ None

20. Think about the alcohol you drank during the past 28 days (4 weeks). A drink is defined as one 12-ounce can of beer, one mixed drink, or one glass of wine. How many of the 28 days did you have…

	Number of days
No drinks	_____ 1
1 drink	_____ 2
2 drinks	_____ 3
3 drinks	_____ 4
4 drinks	_____ 5
5 drinks	_____ 6
6 drinks	_____ 7
7 or more drinks	_____ 8

21. In the last 28 days (4 weeks), how many times did you drive a motor vehicle **within two hours** after drinking an alcoholic beverage?

1 ☐ 0 2 ☐ 1 3 ☐ 2 4 ☐ 3–5

5 ☐ 6–10 6 ☐ 11–15 7 ☐ 16–20 8 ☐ Over 20

22. In the last 12 months, how often did you drive after using marijuana or other illegal drugs?

1 ☐ Never 2 ☐ Rarely 3 ☐ Sometimes

4 ☐ Often 5 ☐ Almost always

THANK YOU FOR YOUR ASSISTANCE!

You may make additional comments on a separate piece of paper.

Please return the completed survey in the pre-paid envelope provided.

Appendix B

Wave 1 Survey Contact Letter

DEPARTMENT OF MOTOR VEHICLES

RESEARCH AND DEVELOPMENT BRANCH
P.O. BOX 932382 MS: F-126
SACRAMENTO, CA 94232-3820

2010 California Driver Survey

July 5, 2010

Doug Driver
P.O. Box 9999
Anytown, Ca. 99999

Dear Doug,

We are asking a small group of California drivers about their driving habits. You have been selected at random to participate in this survey.

The information we collect will help us improve our service and develop more effective driver safety programs. Your answers will be confidential and have no effect on your driving privileges now or in the future.

Please take a few minutes to complete this survey. When you are done, mail it back to us in the enclosed envelope. No postage is needed if you use the envelope provided.

Thank you for completing this survey and helping us find better ways to improve the safety of driving in California and better serve the public.

Sincerely,

DAVID DeYOUNG, Chief
Research and Development Branch

Enclosure

#####

Appendix C

Wave 2 Survey Contact Letter

DEPARTMENT OF MOTOR VEHICLES

RESEARCH AND DEVELOPMENT BRANCH
P.O. BOX 932382 MS: F-126
SACRAMENTO, CA 94232-3820

2010 California Driver Survey

August 10, 2010

Doug Driver
P.O. Box 9999
Anytown, Ca. 99999

Dear Doug,

A few weeks ago, we sent a questionnaire to you and a small number of other drivers selected at random. You are representing many drivers who are similar to yourself in this survey, so your experiences and opinions are very important to us.

Since we have not heard from you, we are sending you another questionnaire in case you did not receive or lost the first one. If you have already returned the questionnaire, please disregard this letter.

If you have not already completed and returned the questionnaire, we hope you will take a few minutes to do so now. Your answers will be completely confidential and will have no impact on your driving privilege.

When you have completed the survey, please place it in the enclosed envelope and drop it in the mail. No postage is needed if you use the envelope provided.

Thank you for completing this survey and helping us find better ways to improve the safety of California drivers and improve our service to the public.

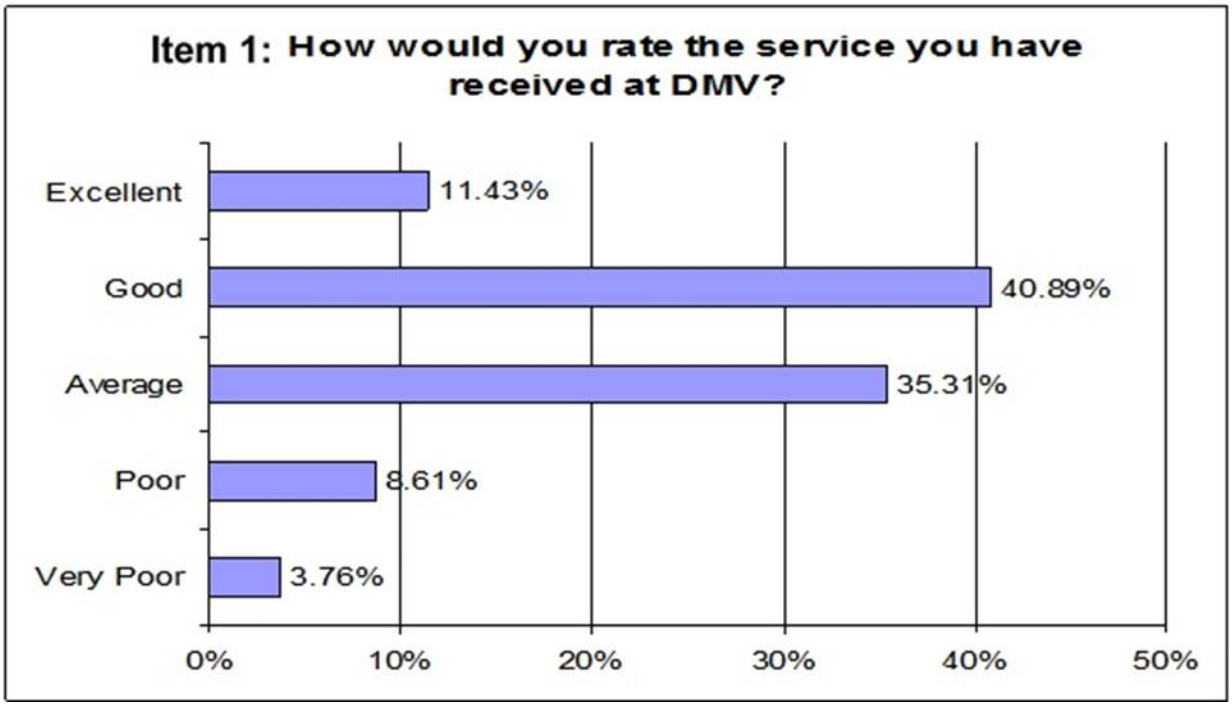
Sincerely,

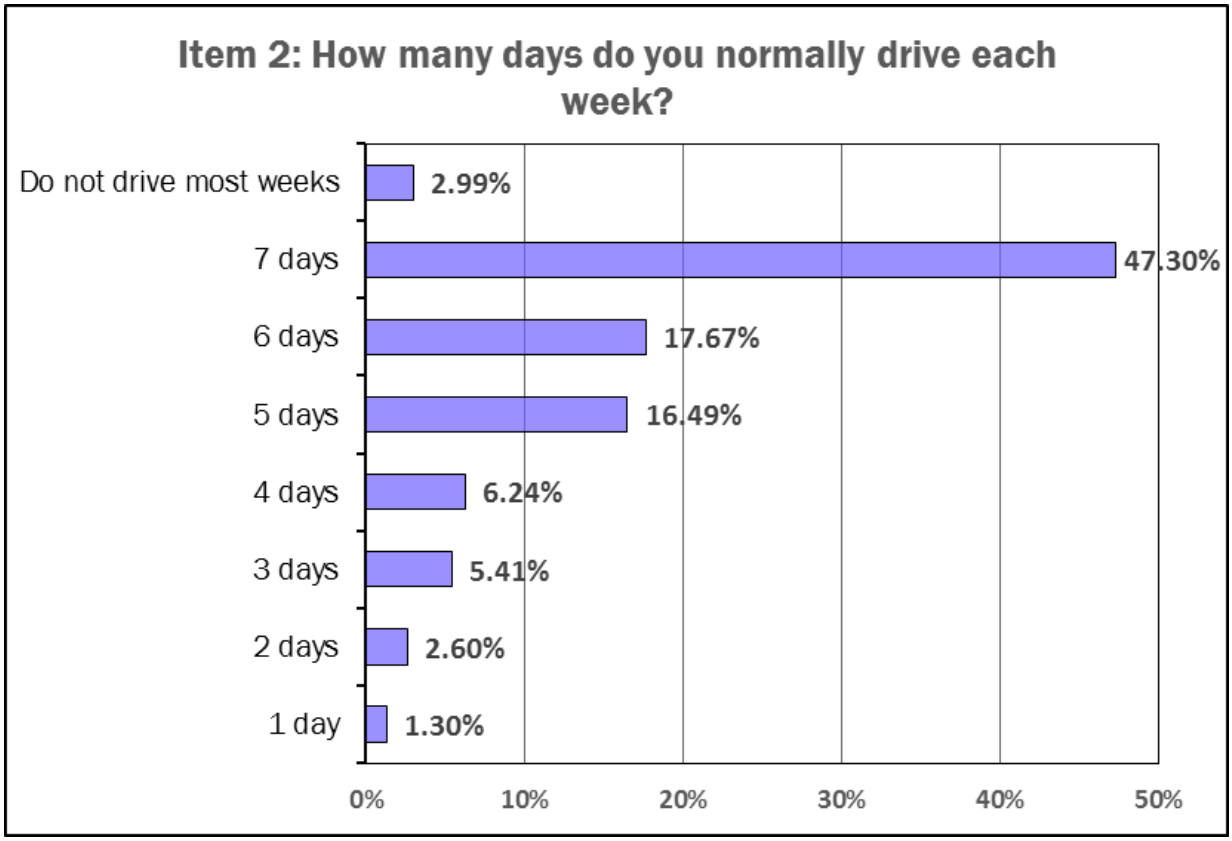
DAVID DeYOUNG, Chief
Research and Development Branch
Enclosure

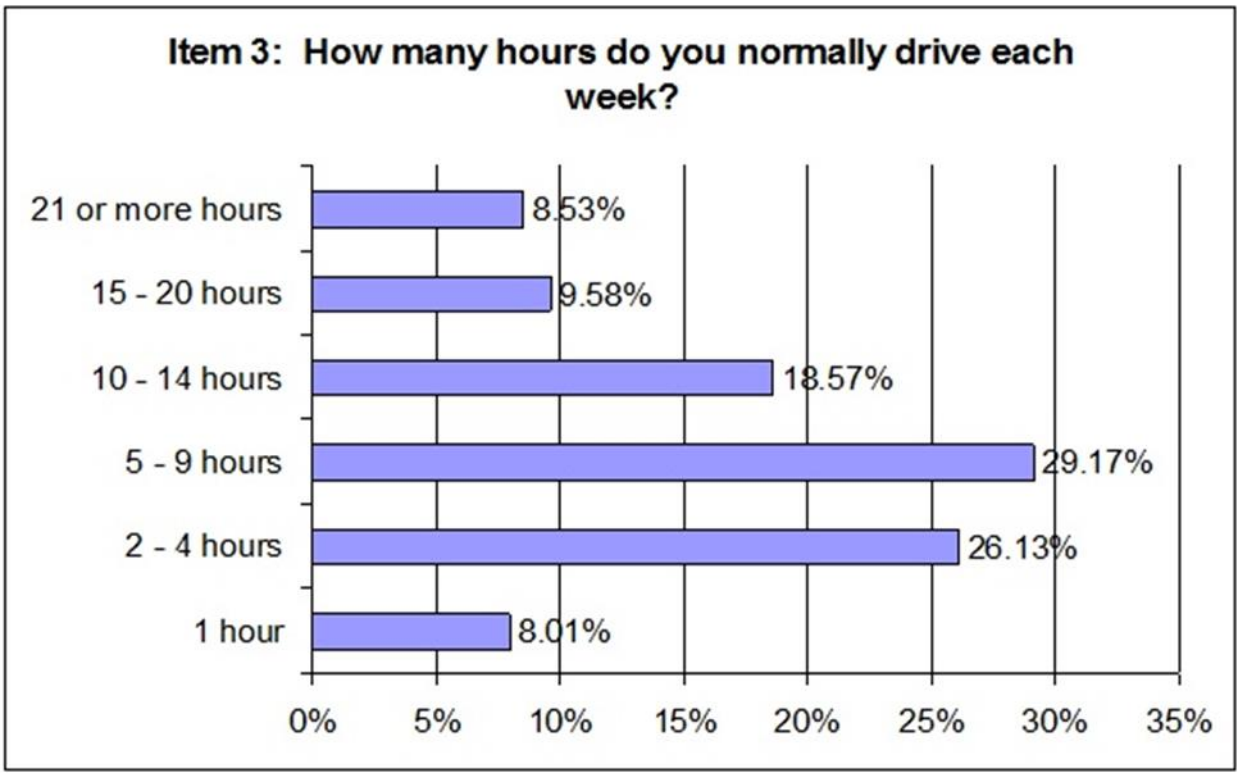
#####

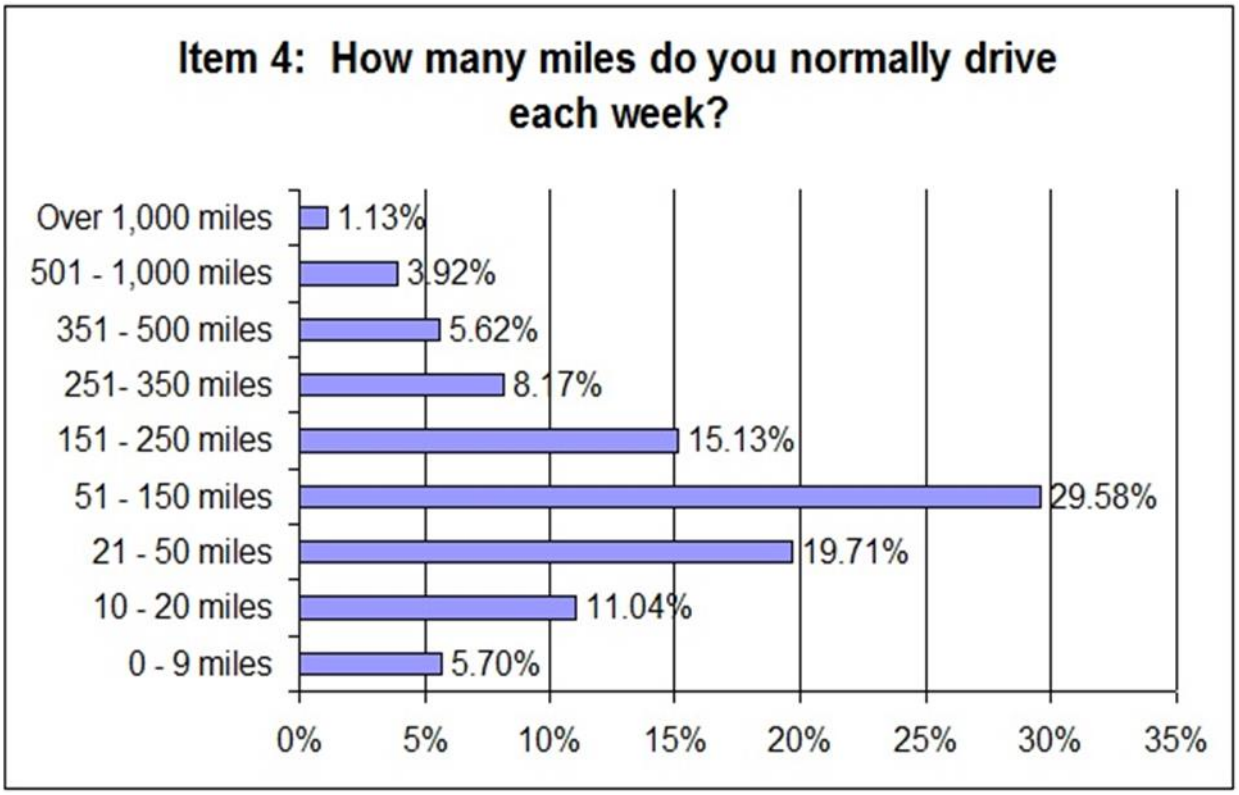
Appendix D

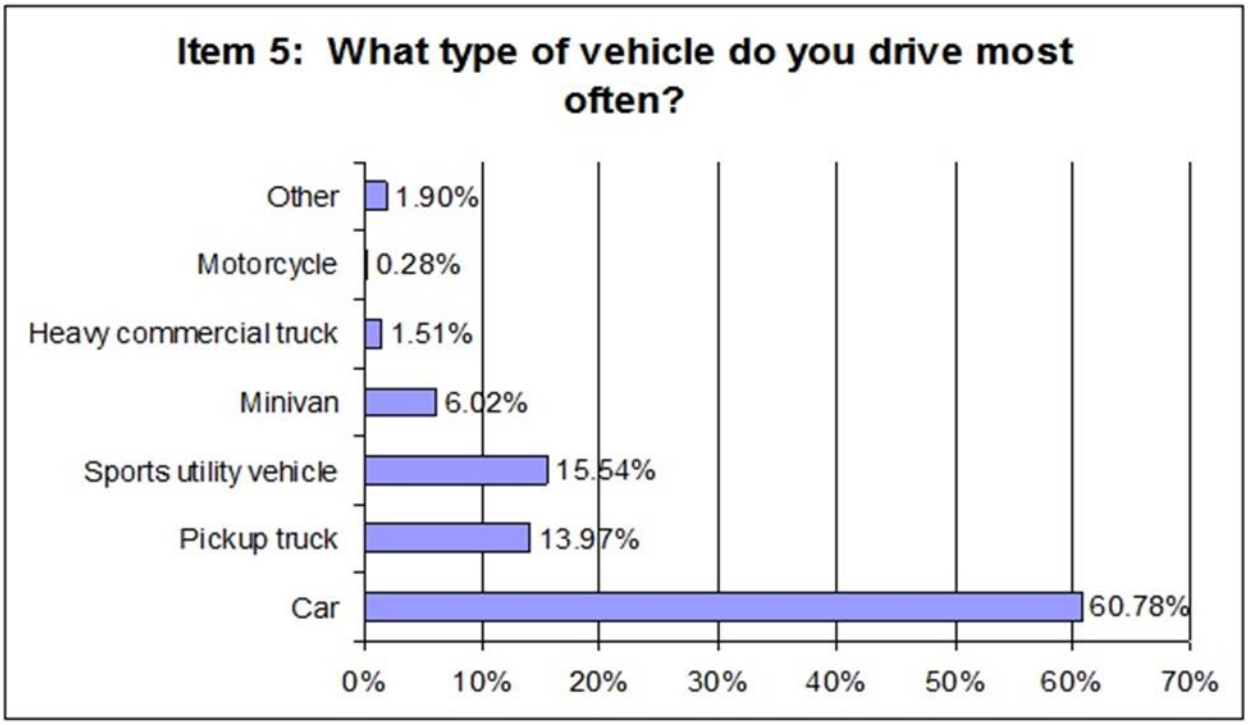
Item Response Distribution from California Driver Survey

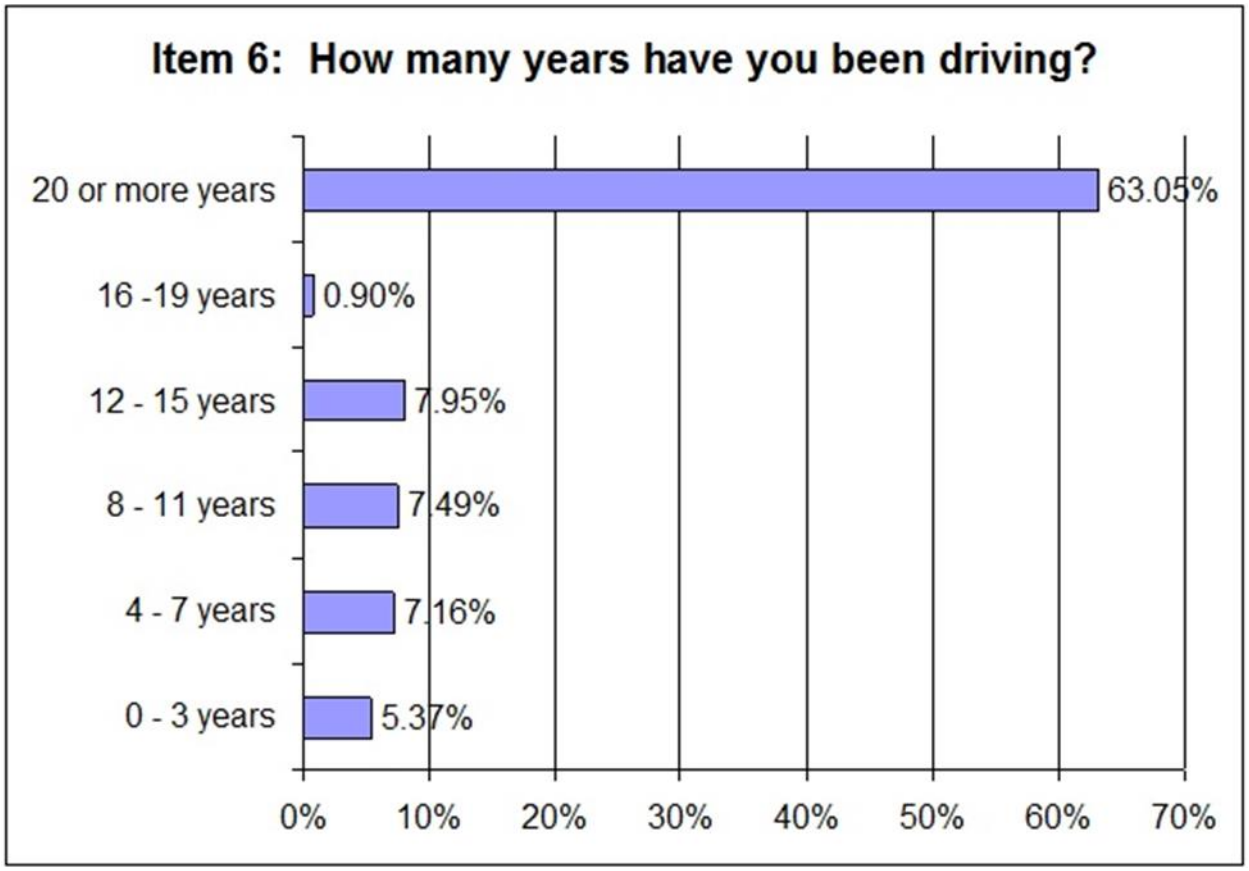

Item 1: How would you rate the service you have received at DMV?
Excellent
11.43%
Good
40.89%
Average
35.31%
Poor
8.61%
Very Poor
3.76%
0%
10%
20%
30%
40%
50%

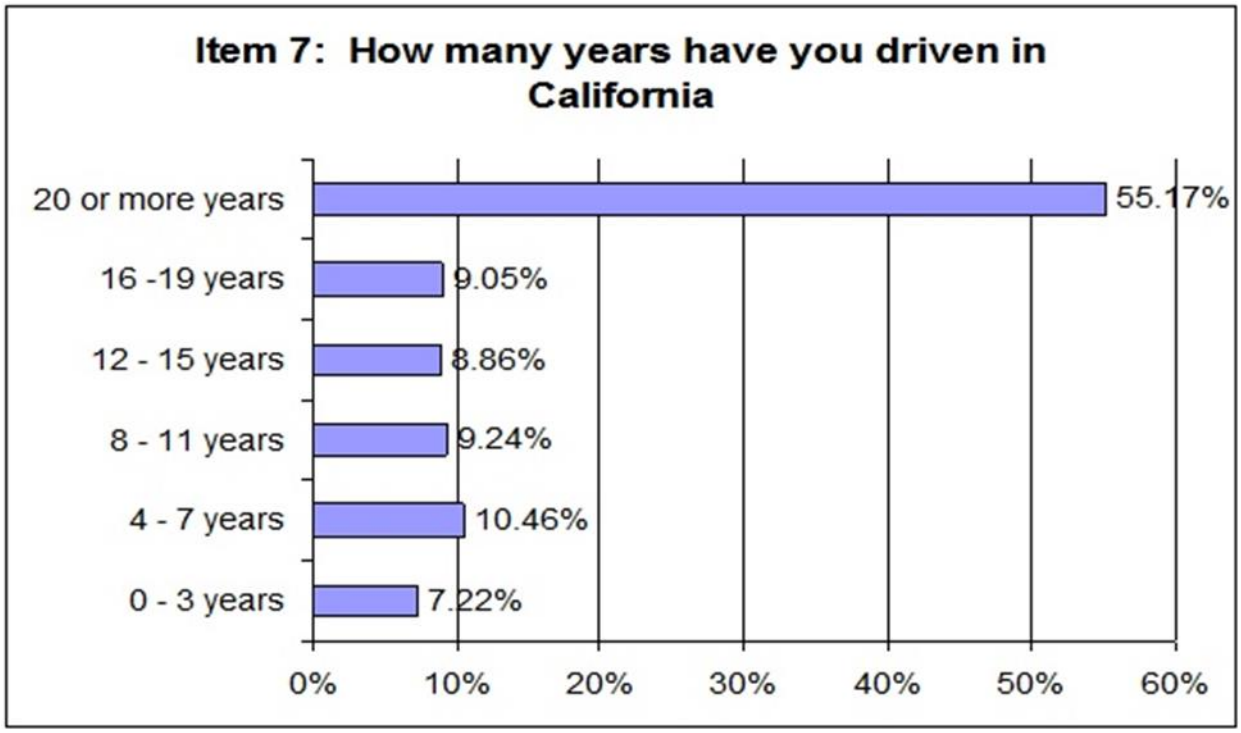

Item 2: How many days do you normally drive each week?
Do not drive most weeks
2.99%
7 days
47.30%
6 days
17.67%
5 days
16.49%
4 days
6.24%
3 days
5.41%
2 days
2.60%
1 day
1.30%
0%
10%
20%
30%
40%
50%

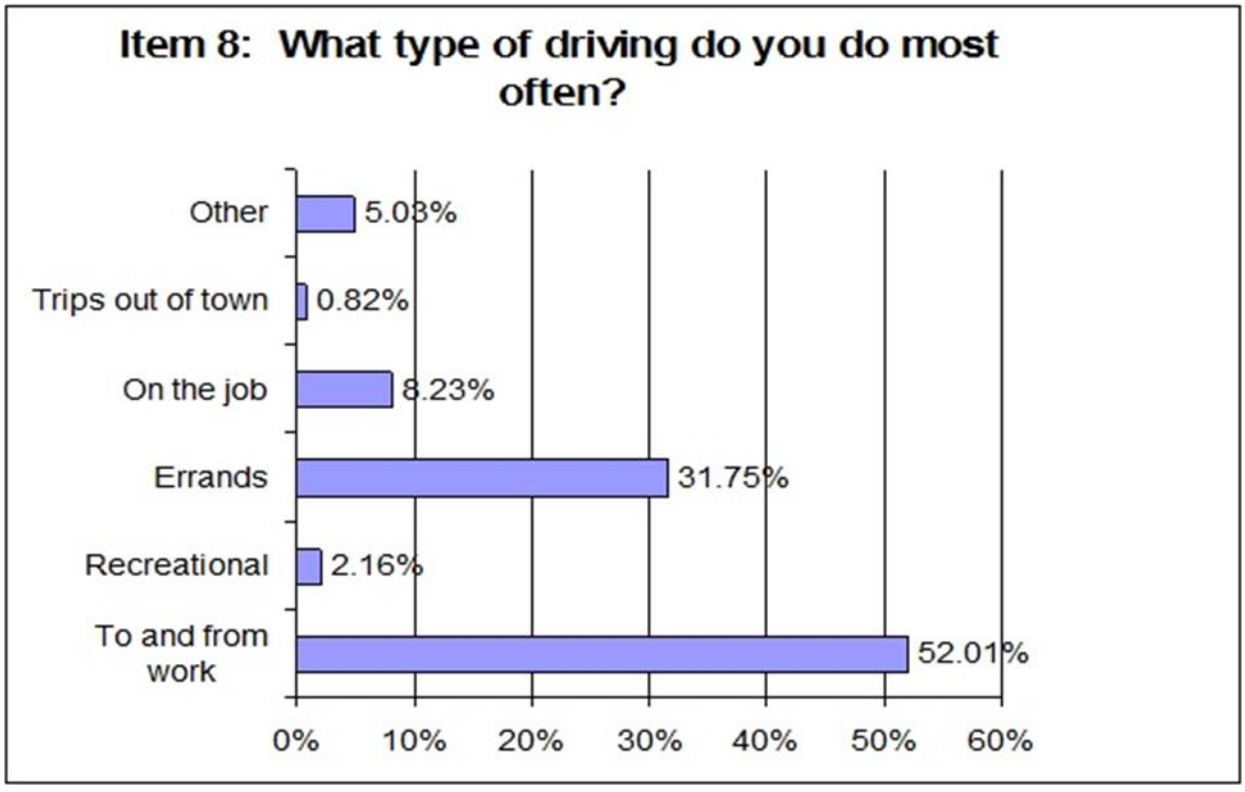

Item 3: How many hours do you normally drive each week?
21 or more hours
8.53%
15 - 20 hours
9.58%
10 - 14 hours
18.57%
5 - 9 hours
29.17%
2 - 4 hours
26.13%
1 hour
8.01%
0%
5%
10%
15%
20%
25%
30%
35%

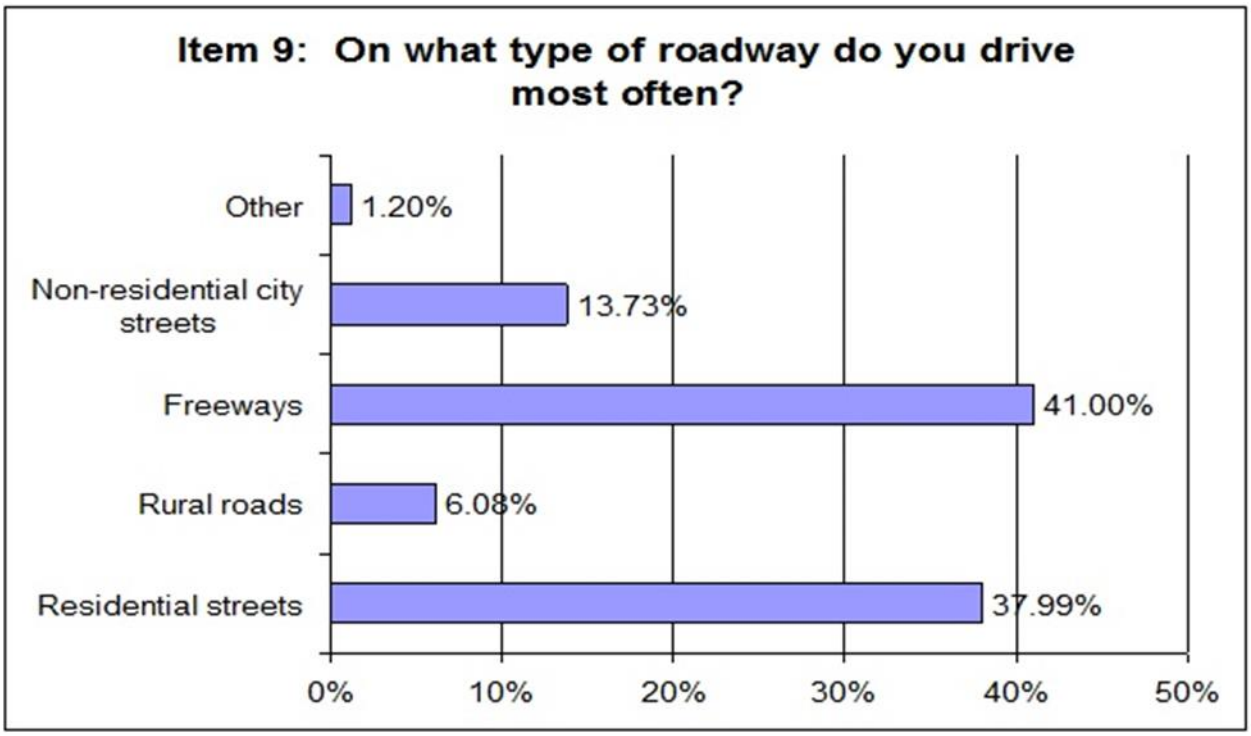

Item 4: How many miles do you normally drive each week?
Over 1,000 miles
1.13%
501 - 1,000 miles
3.92%
351 - 500 miles
5.62%
251- 350 miles
8.17%
151 - 250 miles
15.13%
51 - 150 miles
29.58%
21 - 50 miles
19.71%
10 - 20 miles
11.04%
0 - 9 miles
5.70%
0%
5%
10%
15%
20%
25%
30%
35%

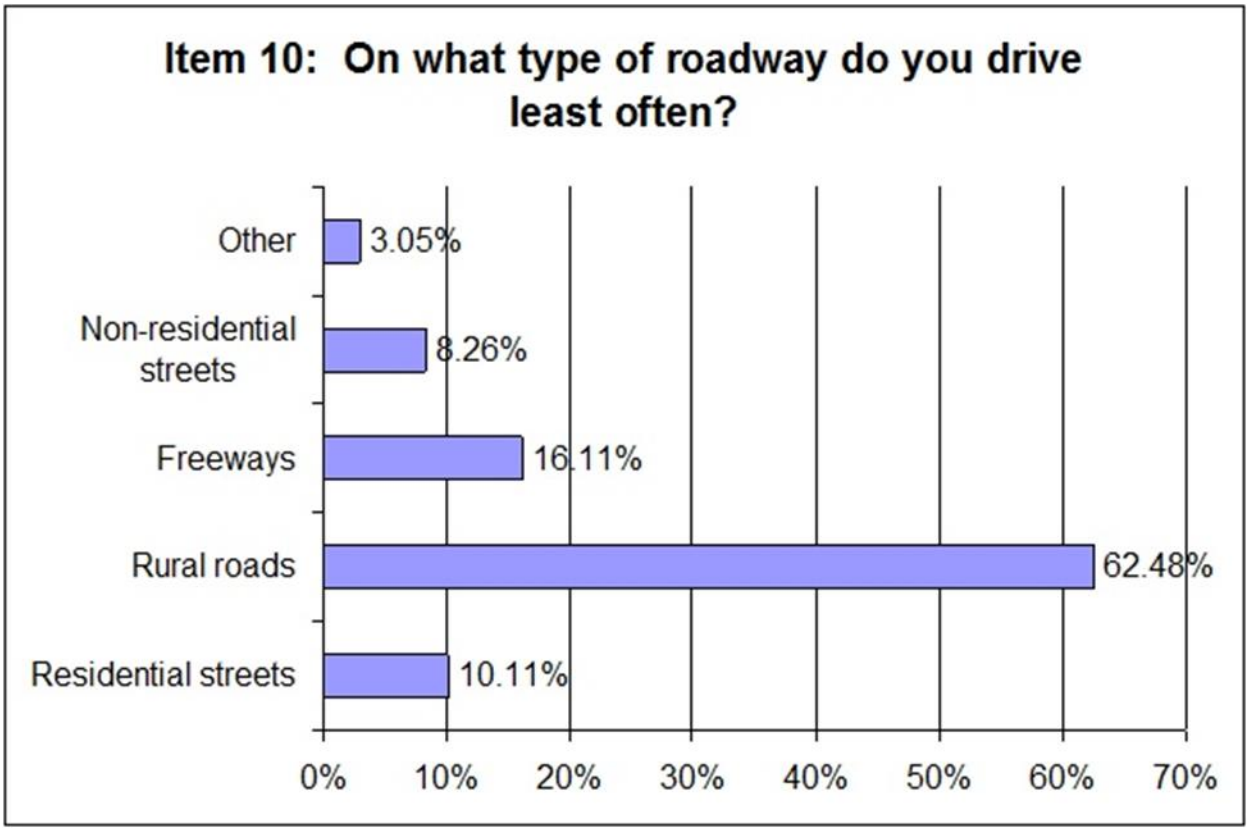

Item 5: What type of vehicle do you drive most often?
Other
1.90%
Motorcycle
0.28%
Heavy commercial truck
1.51%
Minivan
6.02%
Sports utility vehicle
15.54%
Pickup truck
13.97%
Car
60.78%
0%
10%
20%
30%
40%
50%
60%
70%

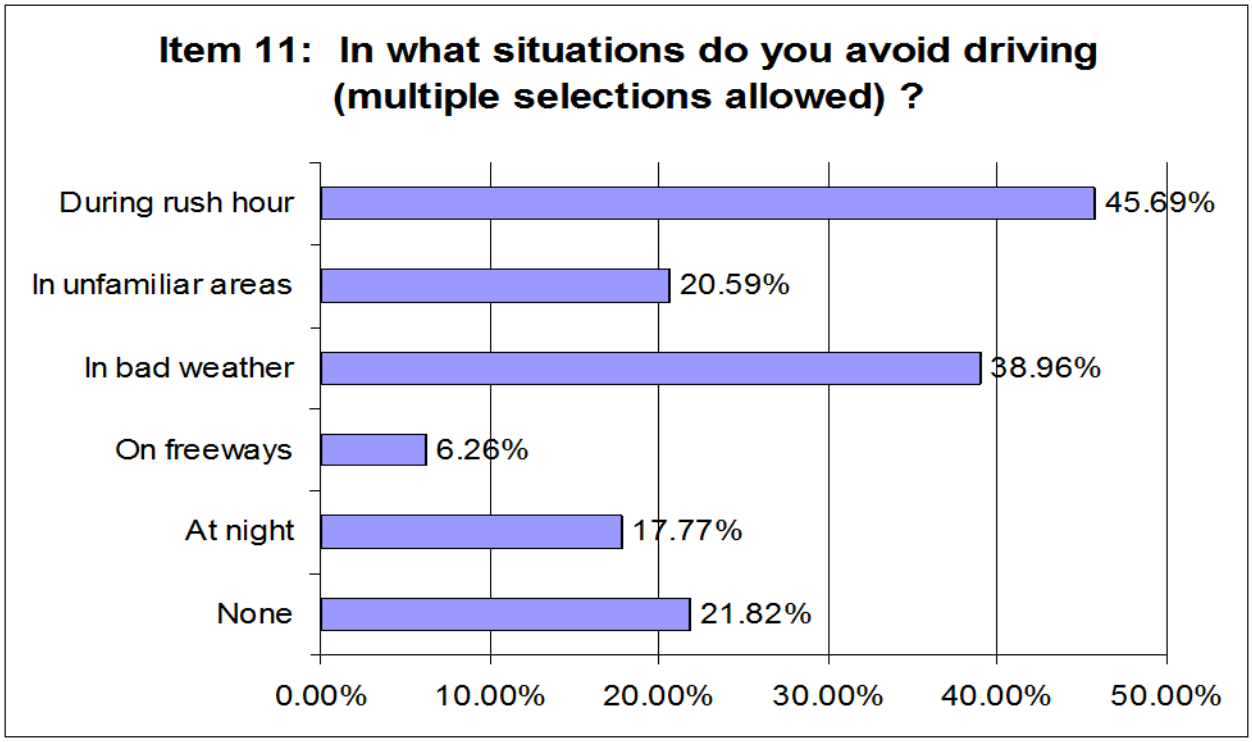

Item 6: How many years have you been driving?
20 or more years
63.05%
16 -19 years
0.90%
12 - 15 years
7.95%
8 - 11 years
7.49%
4 - 7 years
7.16%
0 - 3 years
5.37%
0%
10%
20%
30%
40%
50%
60%
70%

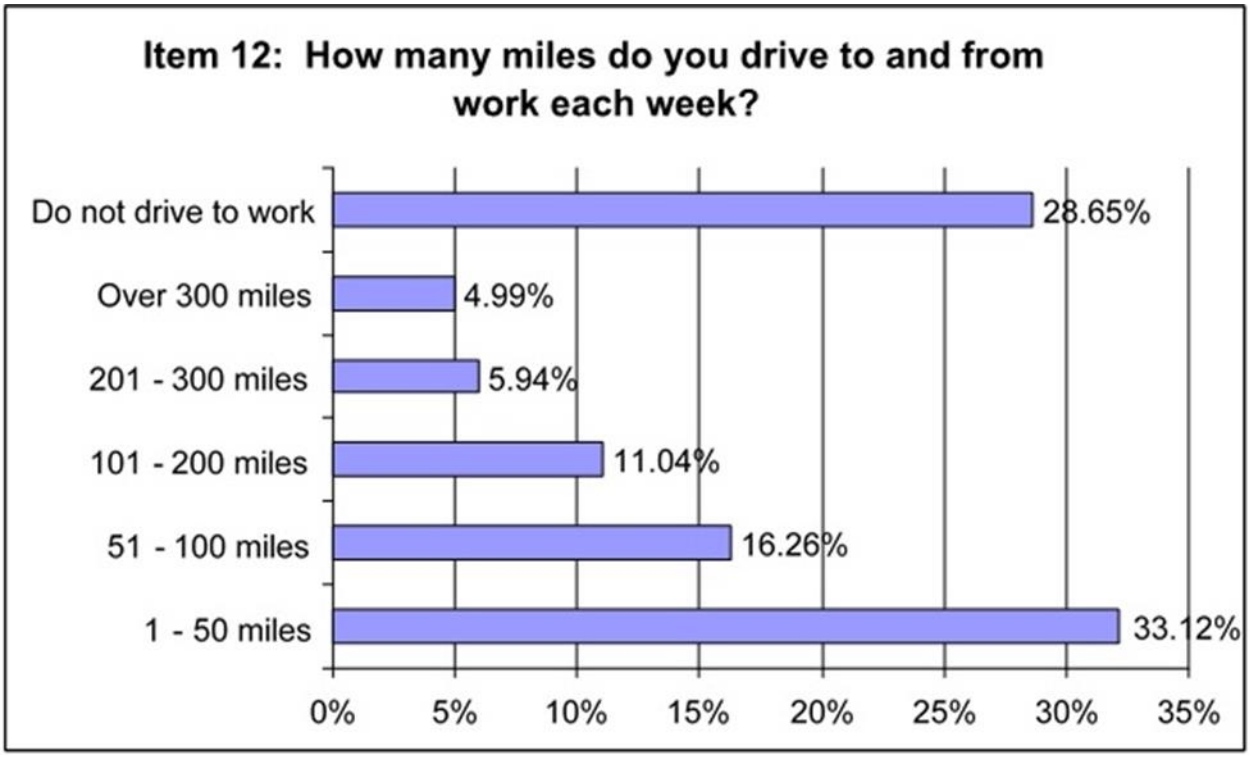

Item 7: How many years have you driven in California
20 or more years
16 -19 years
12 - 15 years
8 - 11 years
4 - 7 years
0 - 3 years
55.17%
9.05%
8.86%
9.24%
10.46%
7.22%
0%
10%
20%
30%
40%
50%
60%

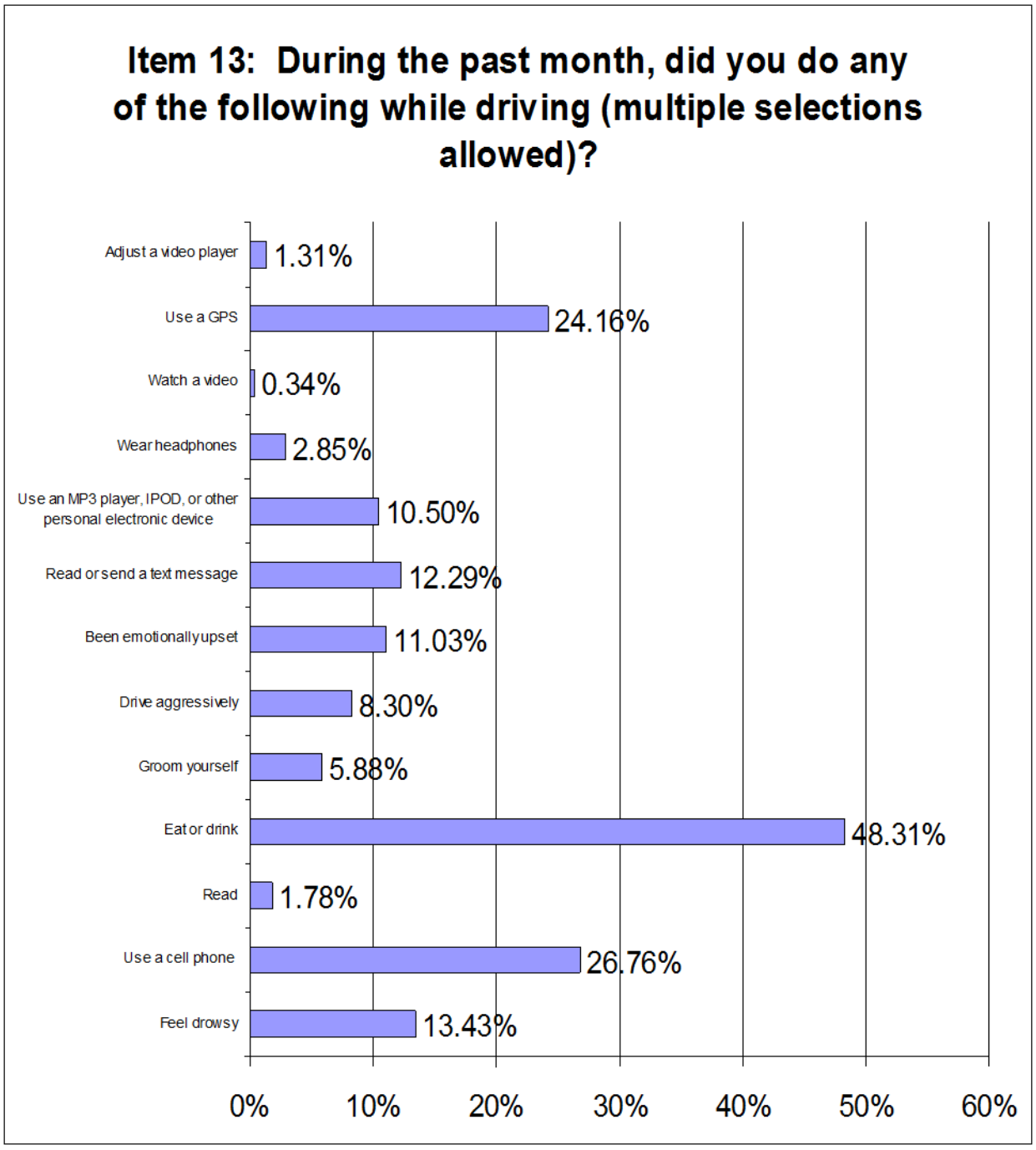

Item 8: What type of driving do you do most often?
Other
Trips out of town
On the job
Errands
Recreational
To and from work
5.08%
0.82%
8.23%
31.75%
2.16%
52.01%
0%
10%
20%
30%
40%
50%
60%

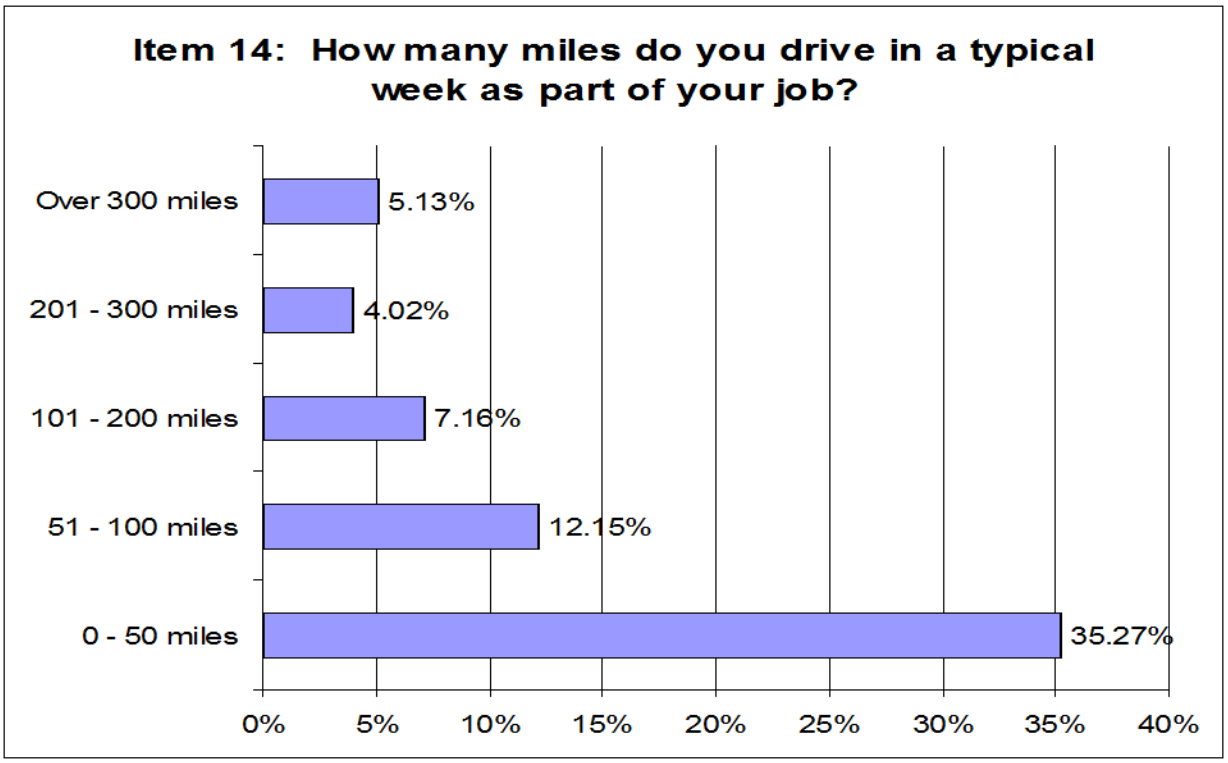

Item 9: On what type of roadway do you drive most often?
Other
1.20%
Non-residential city streets
13.73%
Freeways
41.00%
Rural roads
6.08%
Residential streets
37.99%
0%
10%
20%
30%
40%
50%

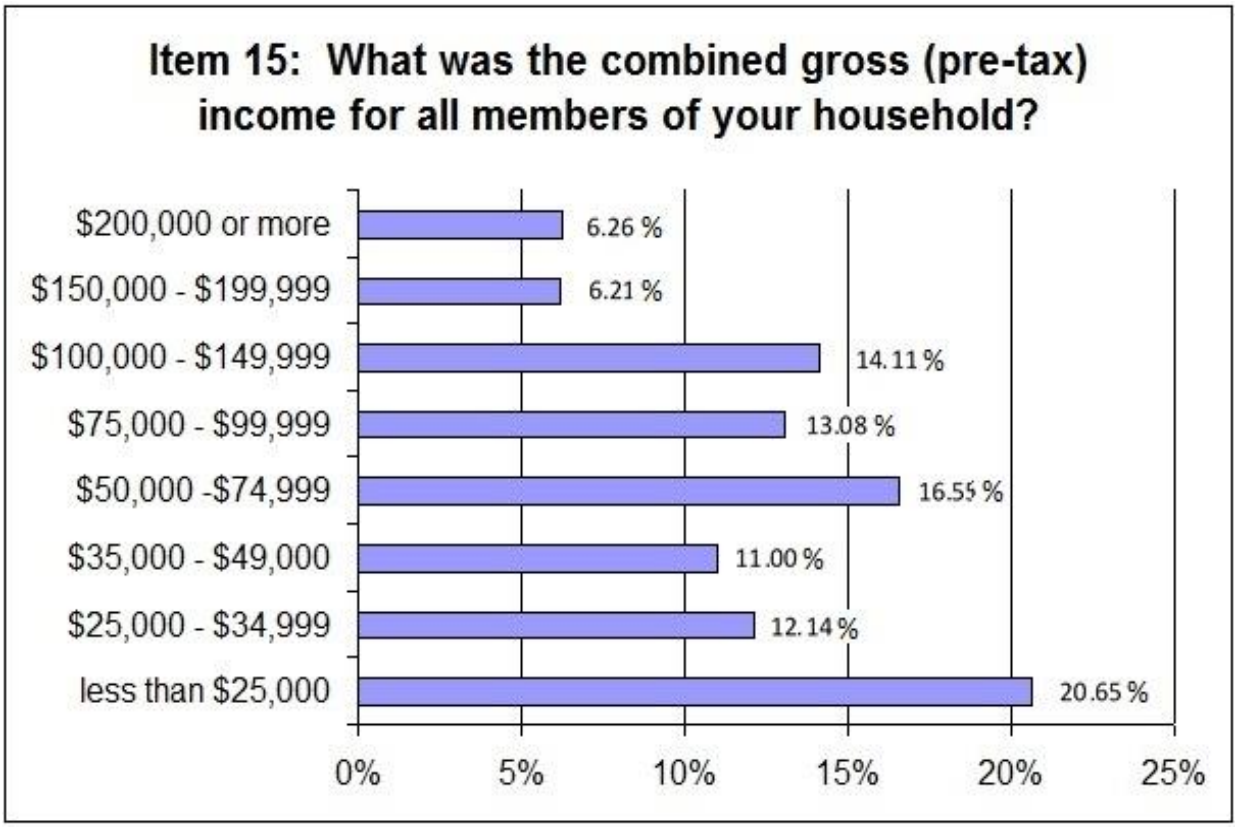

Item 10: On what type of roadway do you drive least often?
Other
3.05%
Non-residential streets
8.26%
Freeways
16.11%
Rural roads
62.48%
Residential streets
10.11%
0%
10%
20%
30%
40%
50%
60%
70%

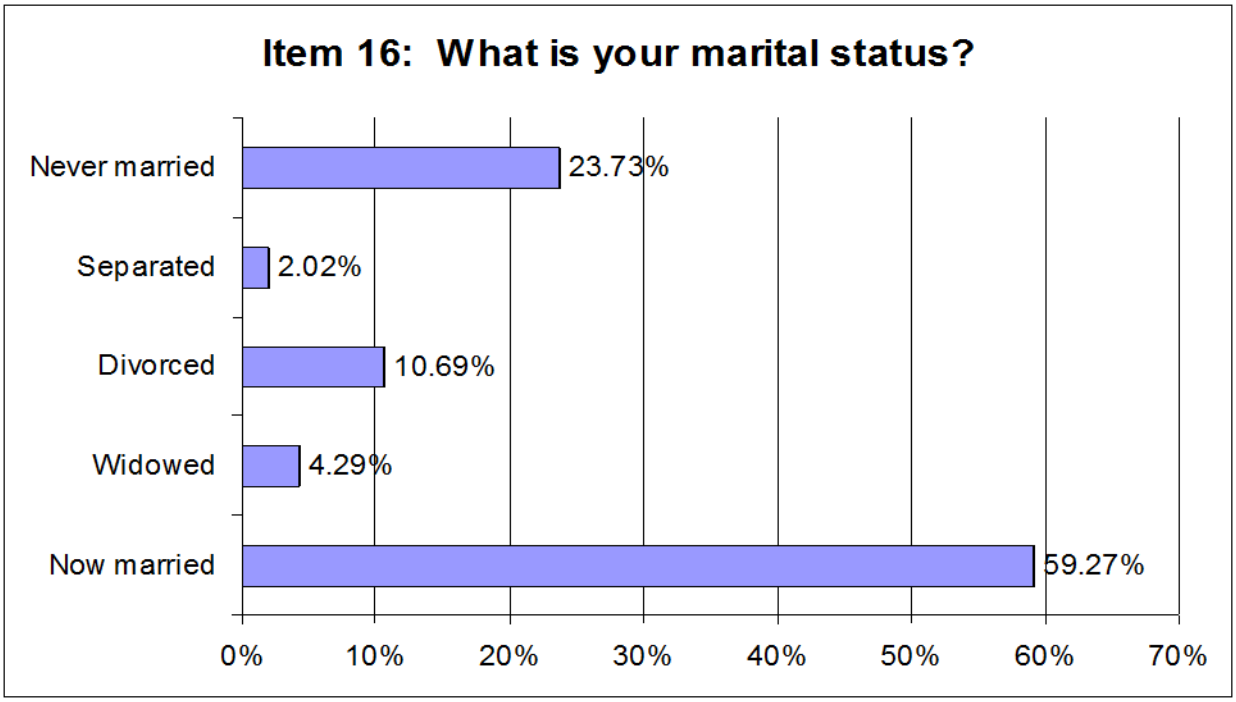

Item 11: In what situations do you avoid driving (multiple selections allowed) ?
During rush hour
45.69%
In unfamiliar areas
20.59%
In bad weather
38.96%
On freeways
6.26%
At night
17.77%
None
21.82%
0.00%
10.00%
20.00%
30.00%
40.00%
50.00%

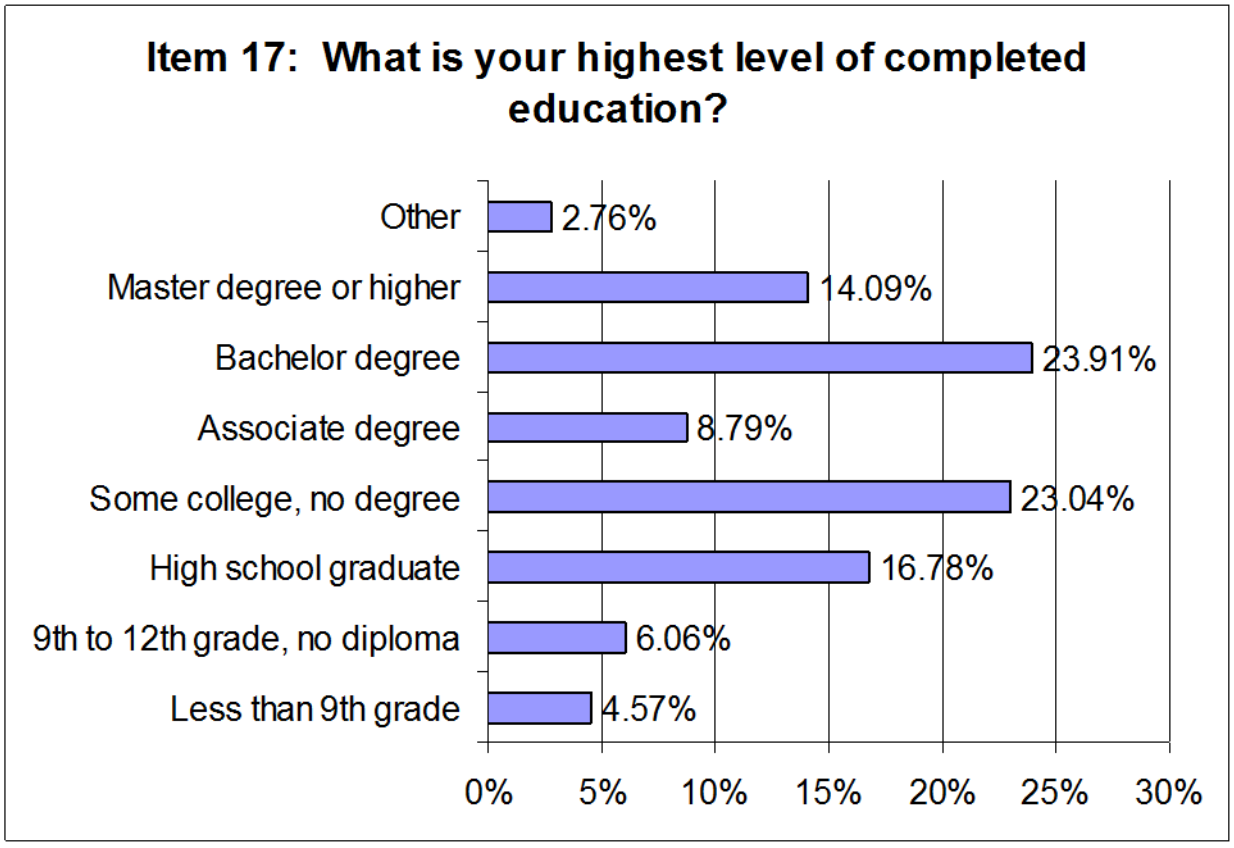

Item 12: How many miles do you drive to and from work each week?
Do not drive to work
28.65%
Over 300 miles
4.99%
201 - 300 miles
5.94%
101 - 200 miles
11.04%
51 - 100 miles
16.26%
1 - 50 miles
33.12%
0%
5%
10%
15%
20%
25%
30%
35%

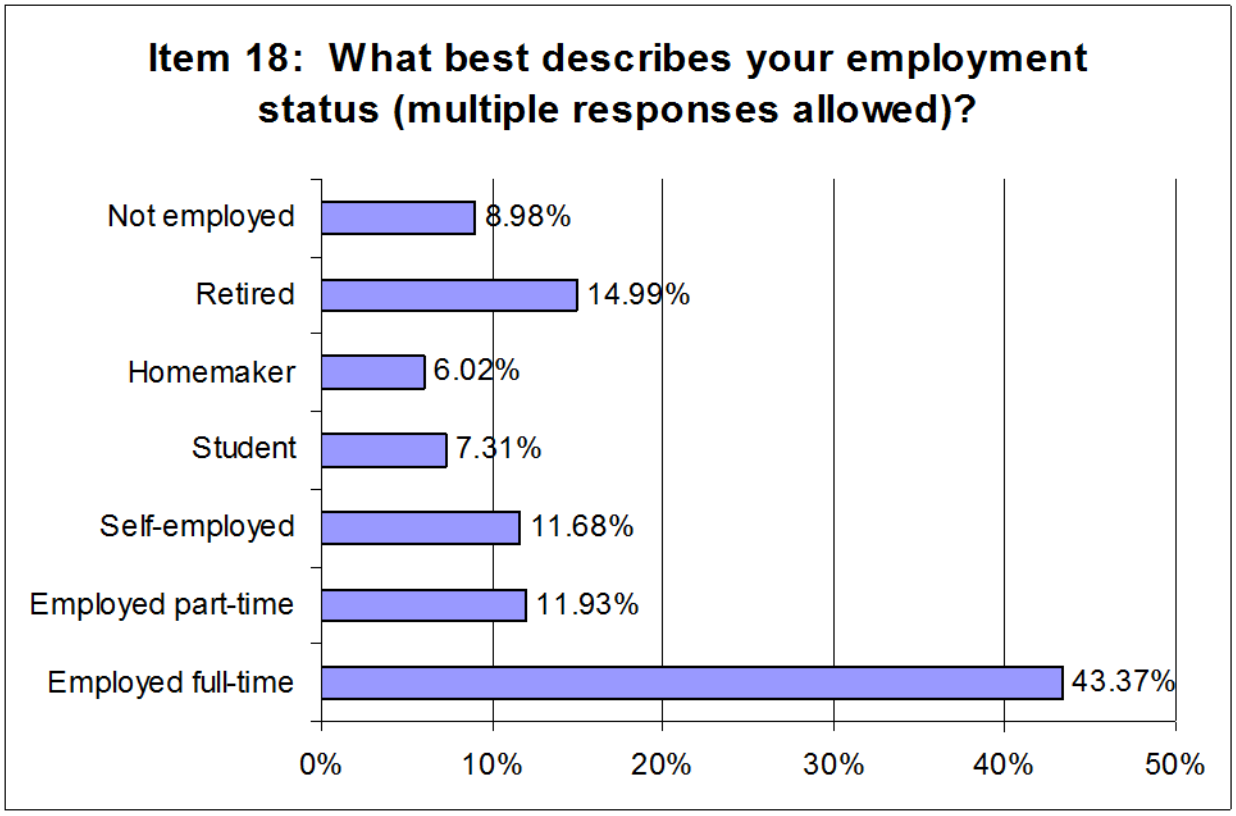

Item 13: During the past month, did you do any of the following while driving (multiple selections allowed)?
Adjust a video player
1.31%
Use a GPS
24.16%
Watch a video
0.34%
Wear headphones
2.85%
Use an MP3 player, IPOD, or other personal electronic device
10.50%
Read or send a text message
12.29%
Been emotionally upset
11.03%
Drive aggressively
8.30%
Groom yourself
5.88%
Eat or drink
48.31%
Read
1.78%
Use a cell phone
26.76%
Feel drowsy
13.43%
0%
10%
20%
30%
40%
50%
60%

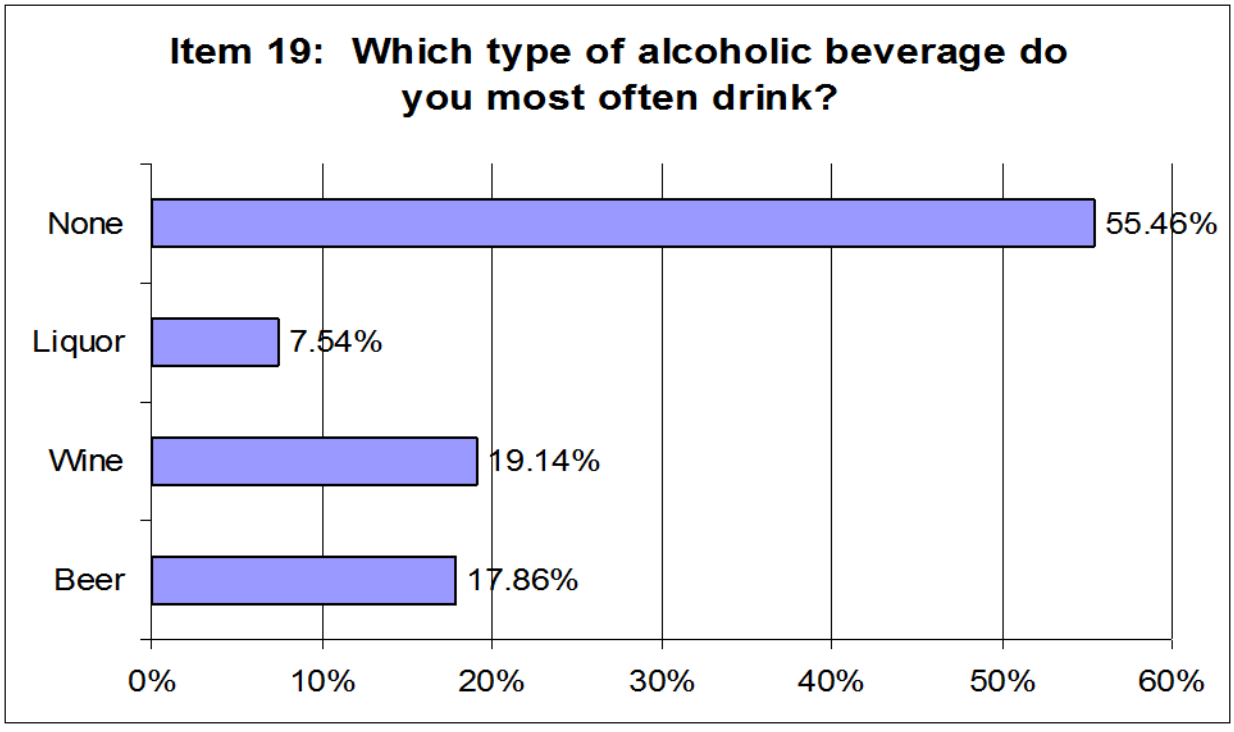

Item 14: How many miles do you drive in a typical week as part of your job?
Over 300 miles
5.13%
201 - 300 miles
4.02%
101 - 200 miles
7.16%
51 - 100 miles
12.15%
0 - 50 miles
35.27%
0%
5%
10%
15%
20%
25%
30%
35%
40%

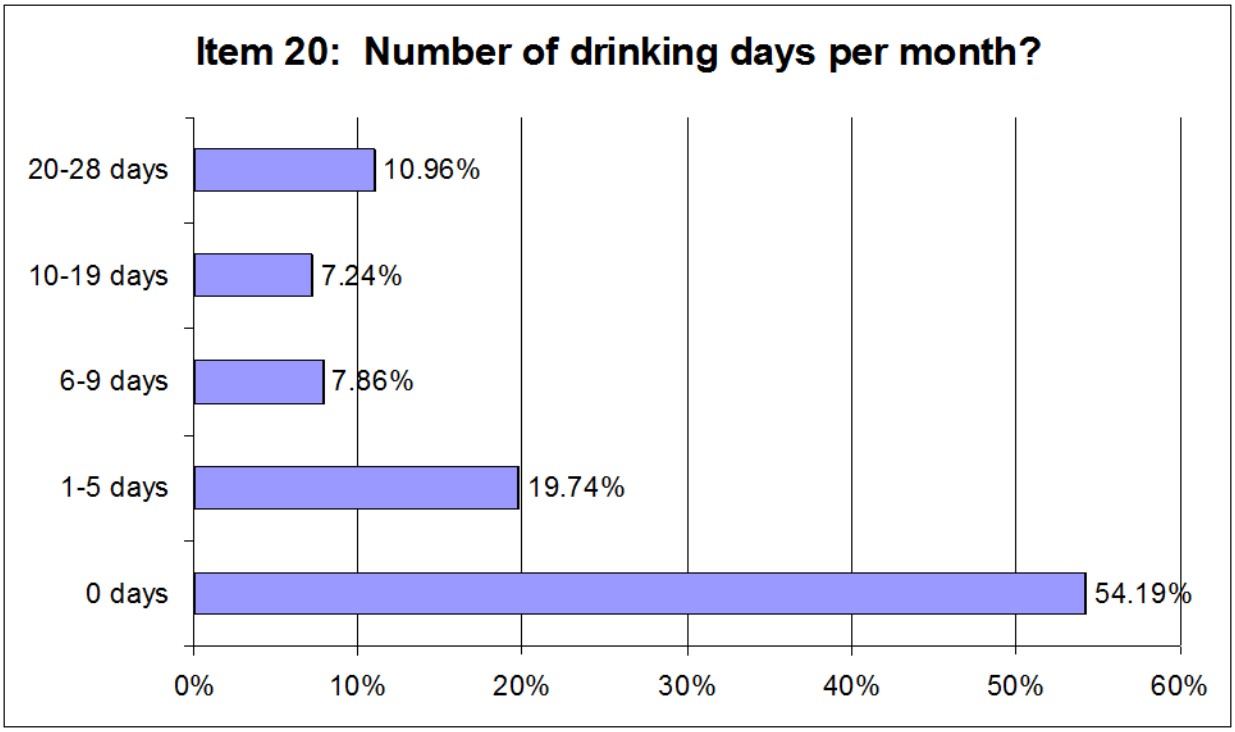

Item 15: What was the combined gross (pre-tax) income for all members of your household?
$200,000 or more
6.26 %
$150,000 - $199,999
6.21 %
$100,000 - $149,999
14.11 %
$75,000 - $99,999
13.08 %
$50,000 -$74,999
16.55 %
$35,000 - $49,000
11.00 %
$25,000 - $34,999
12.14 %
less than $25,000
20.65 %
0%
5%
10%
15%
20%
25%

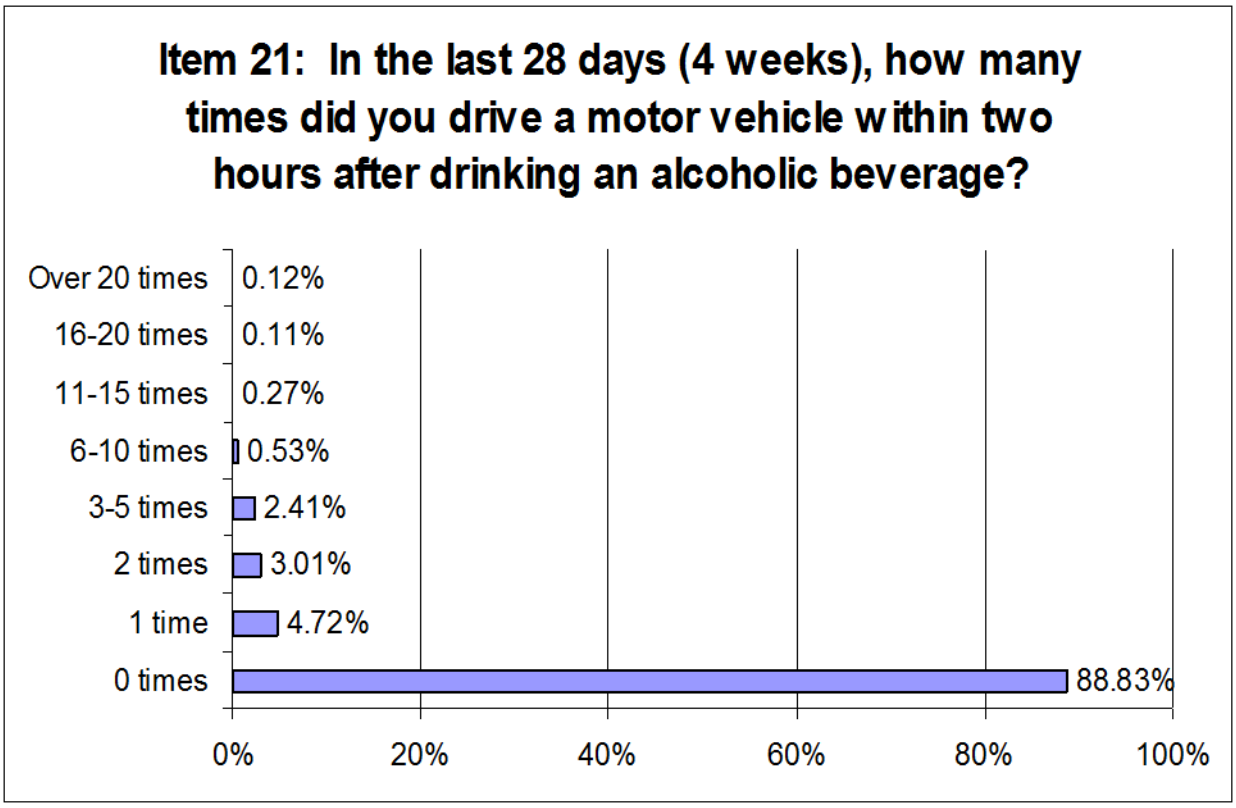

Item 16: What is your marital status?
Never married
23.73%
Separated
2.02%
Divorced
10.69%
Widowed
4.29%
Now married
59.27%
0%
10%
20%
30%
40%
50%
60%
70%

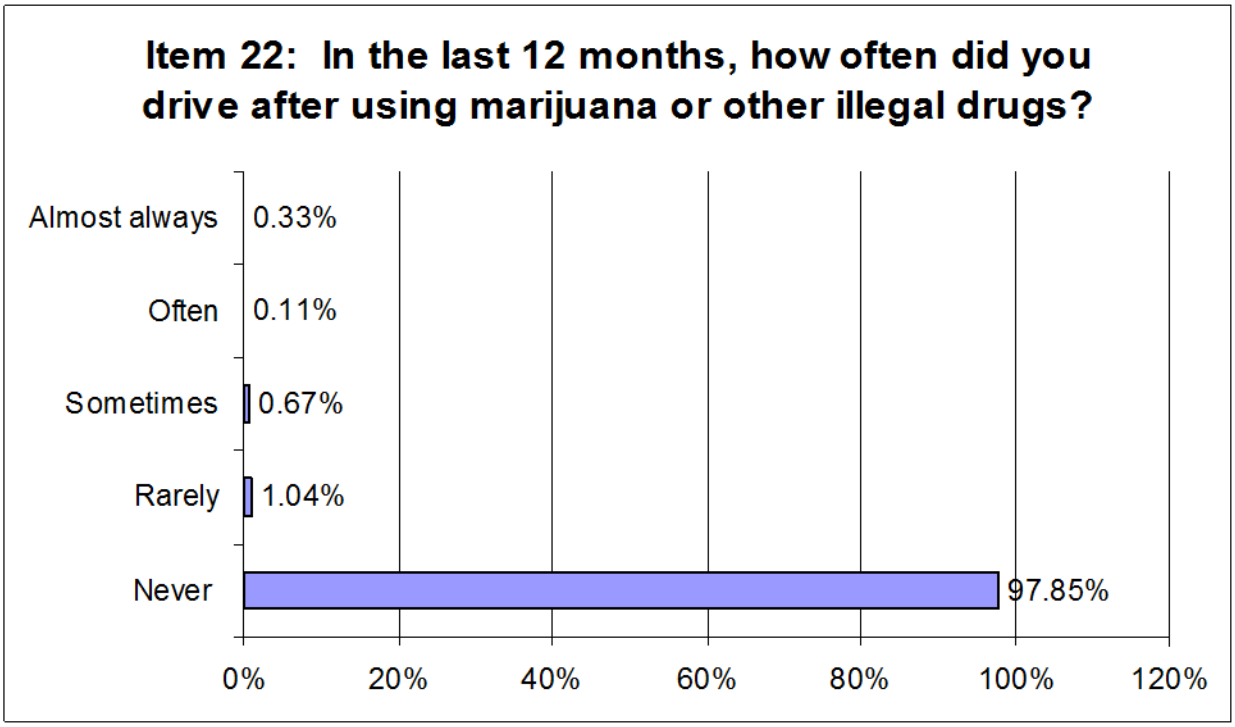

Item 17: What is your highest level of completed education?
Other
2.76%
Master degree or higher
14.09%
Bachelor degree
23.91%
Associate degree
8.79%
Some college, no degree
23.04%
High school graduate
16.78%
9th to 12th grade, no diploma
6.06%
Less than 9th grade
4.57%
0%
5%
10%
15%
20%
25%
30%


Item 18: What best describes your employment status (multiple responses allowed)?
Not employed
8.98%
Retired
14.99%
Homemaker
6.02%
Student
7.31%
Self-employed
11.68%
Employed part-time
11.93%
Employed full-time
43.37%
0%
10%
20%
30%
40%
50%


Item 19: Which type of alcoholic beverage do you most often drink?
None
55.46%
Liquor
7.54%
Wine
19.14%
Beer
17.86%
0%
10%
20%
30%
40%
50%
60%


Item 20: Number of drinking days per month?
20-28 days
10.96%
10-19 days
7.24%
6-9 days
7.86%
1-5 days
19.74%
0 days
54.19%
0%
10%
20%
30%
40%
50%
60%


Item 21: In the last 28 days (4 weeks), how many times did you drive a motor vehicle within two hours after drinking an alcoholic beverage?
Over 20 times
0.12%
16-20 times
0.11%
11-15 times
0.27%
6-10 times
0.53%
3-5 times
2.41%
2 times
3.01%
1 time
4.72%
0 times
88.83%
0%
20%
40%
60%
80%
100%


Item 22: In the last 12 months, how often did you drive after using marijuana or other illegal drugs?
Almost always
0.33%
Often
0.11%
Sometimes
0.67%
Rarely
1.04%
Never
97.85%
0%
20%
40%
60%
80%
100%
120%

Appendix E

Data Dictionary for the Multiple Poisson Regression Models Constructed Using the Survey Sample

Data Dictionary for the Multiple Poisson Regression Models Constructed Using the Survey Sample

Criterion: Number of total crashes from 01/01/09 through 05/31/10
Predictors:
Gender (ref: Male)
Female = 1 if female, 0 otherwise
Age = the age of the driver in years as of 12/31/08
License class (ref: Non commercial)
Commercial license = 1 if driver holds a commercial license, 0 otherwise
Prior 3-year total citations = Number of total citations on record 01/01/06 through 12/31/08
Prior 3-year total crashes = Number of total crashes on record 01/01/06 through 12/31/08
Territorial composite index = zip code total composite index 01/01/06 through 12/31/08
Territorial total crash index = zip code total crash index 01/01/06 through 12/31/08
Exposure: Amount of driving
Number of days driven each week
(ref: Do not drive in most weeks)
1 day = 1 if selected, 0 otherwise
2 days = 1 if selected, 0 otherwise
3 days = 1 if selected, 0 otherwise
4 days = 1 if selected, 0 otherwise
5 days = 1 if selected, 0 otherwise
6 days = 1 if selected, 0 otherwise
7 days = 1 if selected, 0 otherwise
Number of hours driven each week
(ref: 1 hour)
2-4 hours = 1 if selected, 0 otherwise
5-9 hours = 1 if selected, 0 otherwise
10-14 hours = 1 if selected, 0 otherwise
15-20 hours = 1 if selected, 0 otherwise
21 or more hours = 1 if selected, 0 otherwise
Number of miles driven each week
(ref: 0-9 miles)
10-20 miles = 1 if selected, 0 otherwise
21-50 miles = 1 if selected, 0 otherwise
51-150 miles = 1 if selected, 0 otherwise
251-350 miles = 1 if selected, 0 otherwise
351-500 miles = 1 if selected, 0 otherwise
501-1,000 miles = 1 if selected, 0 otherwise
over 1,000 miles = 1 if selected, 0 otherwise
Total years driving (ref: 20 or more years)
0-3 years = 1 if selected, 0 otherwise
4-7 years = 1 if selected, 0 otherwise
8-11 years = 1 if selected, 0 otherwise
12-15 years = 1 if selected, 0 otherwise
16-19 years = 1 if selected, 0 otherwise
Total miles driven to and from work each week
(ref: Do not drive to work)
251-350 miles = 1 if selected, 0 otherwise
351-500 miles = 1 if selected, 0 otherwise

Predictors:

501-1,000 miles = 1 if selected, 0 otherwise
over 1,000 miles = 1 if selected, 0 otherwise
Total years driving (ref: 20 or more years)
0-3 years = 1 if selected, 0 otherwise
4-7 years = 1 if selected, 0 otherwise
8-11 years = 1 if selected, 0 otherwise
12-15 years = 1 if selected, 0 otherwise
16-19 years = 1 if selected, 0 otherwise
Total miles driven to and from work each week (ref: Do not drive to work)
1-50 miles = 1 if selected, 0 otherwise
51-100 miles = 1 if selected, 0 otherwise
101-200 miles = 1 if selected, 0 otherwise
201-300 miles = 1 if selected, 0 otherwise
over 300 miles = 1 if selected, 0 otherwise
Total miles driven in a typical week as part of job (ref: Do not drive on the job)
1-50 miles = 1 if selected, 0 otherwise
51-100 miles = 1 if selected, 0 otherwise
101-200 miles = 1 if selected, 0 otherwise
201-300 miles = 1 if selected, 0 otherwise
over 300 miles = 1 if selected, 0 otherwise

Exposure: Type of driving

Type of driving done most often (ref: To and from work)
recreational = 1 if selected, 0 otherwise
errands = 1 if selected, 0 otherwise
on the job = 1 if selected, 0 otherwise
trips out of town = 1 if selected, 0 otherwise
other = 1 if selected, 0 otherwise
Type of roadway driven most often (ref: Freeways)
residential streets = 1 if selected, 0 otherwise
rural streets = 1 if selected, 0 otherwise
non-residential city streets = 1 if selected, 0 otherwise
other streets = 1 if selected, 0 otherwise
Type of roadway driven least often (ref: Freeways)
residential streets = 1 if selected, 0 otherwise
rural streets = 1 if selected, 0 otherwise
non-residential city streets = 1 if selected, 0 otherwise
other streets = 1 if selected, 0 otherwise
Type of situations avoided
none = 1 if selected, 0 otherwise
at night = 1 if selected, 0 otherwise
on freeways = 1 if selected, 0 otherwise
in bad weather = 1 if selected, 0 otherwise
in unfamiliar areas = 1 if selected, 0 otherwise
during rush hour = 1 if selected, 0 otherwise

Predictors:

Vehicle

Type of vehicle driven most often

(ref: Car)

pickup truck = 1 if selected, 0 otherwise
sports utility vehicle = 1 if selected, 0 otherwise
minivan = 1 if selected, 0 otherwise
heavy commercial vehicle = 1 if selected, 0 otherwise
motorcycle = 1 if selected, 0 otherwise
other = 1 if selected, 0 otherwise

Distracted, drowsy, and aggressive driving

During the past month, did the following while driving

feel drowsy = 1 if selected, 0 otherwise
use a cell phone = 1 if selected, 0 otherwise
read = 1 if selected, 0 otherwise
eat or drink = 1 if selected, 0 otherwise
groom yourself = 1 if selected, 0 otherwise
drive aggressively = 1 if selected, 0 otherwise
been emotionally upset = 1 if selected, 0 otherwise
read or send a text message = 1 if selected, 0 otherwise
use an MP3 player, IPOD, or other personal electronic device = 1 if selected, 0 otherwise
wear headphones = 1 if selected, 0 otherwise
watch a video = 1 if selected, 0 otherwise
use a GPS = 1 if selected, 0 otherwise
adjust a video player = 1 if selected, 0 otherwise

Socio economic

Annual combined household gross income

(ref: More than $200,000)

less than $25,000 = 1 if selected, 0 otherwise
$25,000 - $34,999 = 1 if selected, 0 otherwise
$35,000 - $49,999 = 1 if selected, 0 otherwise
$50,000 - $74,999 = 1 if selected, 0 otherwise
$75,000 - $99,999 = 1 if selected, 0 otherwise
$100,00 - $149,999 = 1 if selected, 0 otherwise
$150,000 - $199,999 = 1 if selected, 0 otherwise

Marital status (ref: Now married)

widowed = 1 if selected, 0 otherwise
divorced = 1 if selected, 0 otherwise
separated = 1 if selected, 0 otherwise
never married = 1 if selected, 0 otherwise

Highest education level obtained

(ref: Master degree or higher)

less than 9th grade = 1 if selected, 0 otherwise
9th to 12th grade, no diploma = 1 if selected, 0 otherwise
high school graduate = 1 if selected, 0 otherwise
some college, no degree = 1 if selected, 0 otherwise
associate degree = 1 if selected, 0 otherwise
bachelor degree = 1 if selected, 0 otherwise
other = 1 if selected, 0 otherwise

Predictors:

Employment status
Employed full-time = 1 if selected, 0 otherwise
Employed part-time = 1 if selected, 0 otherwise
Self-employed = 1 if selected, 0 otherwise
Student = 1 if selected, 0 otherwise
Homemaker = 1 if selected, 0 otherwise
Retired = 1 if selected, 0 otherwise
Not employed = 1 if selected, 0 otherwise

Alcohol usage

Alcohol beverage drank most often
(ref: None)
beer = 1 if selected, 0 otherwise
wine = 1 if selected, 0 otherwise
liquor = 1 if selected, 0 otherwise

Number of drinking days per month
(ref: 0 drinking days)
1 - 5 drinking days = 1 if selected, 0 otherwise
6 - 9 drinking days = 1 if selected, 0 otherwise
10 - 19 drinking days = 1 if selected, 0 otherwise
20 - 28 drinking days = 1 if selected, 0 otherwise

Number of times in last month drove within two hours of drinking (ref: 0 times)
1 time = 1 if selected, 0 otherwise
2 times = 1 if selected, 0 otherwise
3 - 5 times = 1 if selected, 0 otherwise
6 - 10 times = 1 if selected, 0 otherwise
over 10 times = 1 if selected, 0 otherwise

Drug usage

In last 12 months, how often drove after using marijuana or other illegal drugs (ref: Never)
rarely = 1 if selected, 0 otherwise
sometimes = 1 if selected, 0 otherwise
often or almost always = 1 if selected, 0 otherwise