Fall 2023

Investigation of Discrepancies in South Carolina Traffic Collision Forms

Jackson Wegmet

INVESTIGATION OF DISCREPANCIES IN SOUTH CAROLINA TRAFFIC COLLISION FORMS

by

Jackson Wegmet

Bachelor of Science in Engineering
University of South Carolina, 2022

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in

Civil Engineering

College of Engineering and Computing

University of South Carolina

2023

Accepted by:

Nathan Huynh, Director of Thesis

Yuche Chen, Reader

Chowdhury Siddiqui, Reader

Ann Vail, Dean of the Graduate School

ABSTRACT

The aim of this thesis is to improve the accuracy of information recorded in the South Carolina traffic collision forms. To accomplish this, it examines 200 forms containing information about fatal crashes in work zones between 2014 and 2020 to determine how many discrepancies exist between the written narrative and other fields. In addition to obtaining these statistics, this thesis seeks to identify factors that influence discrepancies. To test the hypothesis that crash complexity and weather influence the investigating officer's level of processing (a theory developed by Craik and Lockhart in 1972), and consequentially his/her ability to complete the traffic collision form accurately, a structural equation model (SEM) is developed. The SEM is used to explain the relationships between measured variables and latent variables and the relationships between latent variables (crash characteristics, weather conditions, and level of processing). SEM results show that increases in collision speed, number of units, number of events, and temperature resulted in an increase in the number of words and characters written in the narrative, whereas increases in precipitation and humidity resulted in a decrease in the number of words and characters written in the narrative. Notably, the number of discrepancies was not statistically significant, suggesting crash and weather-related factors do not affect an officer's reporting accuracy. A multiple linear regression model is also developed to identify factors that influence a form field's frequency of discrepancies. The form field's level of difficulty and its number of inputs are found to be statistically significant.

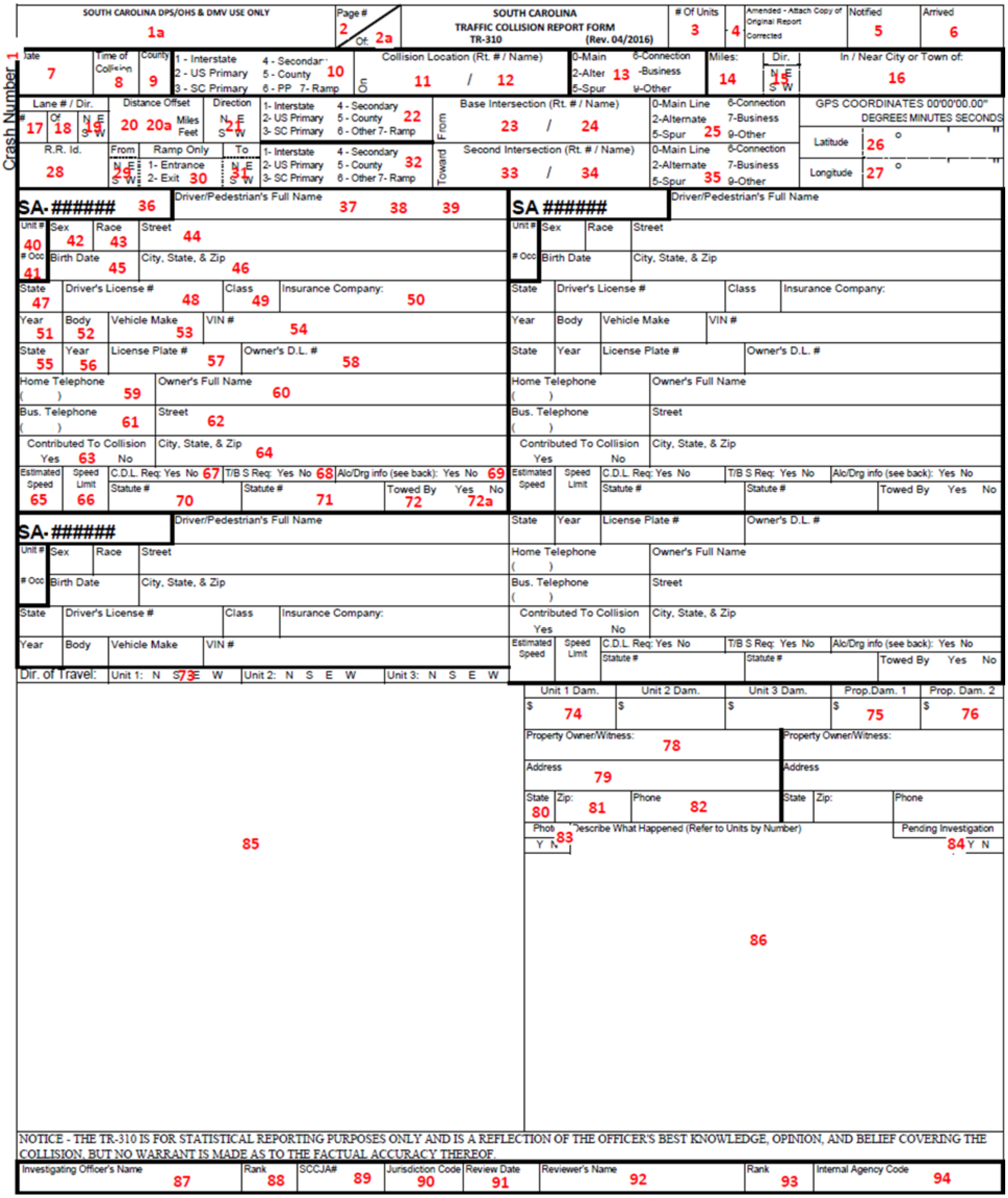
TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The study of misclassification in police crash reports is well-documented (1-6). Misclassification is commonly defined as any instance of incorrect reporting, including an officer misunderstanding the report format, misunderstanding the crash itself, and making errors during the data entry process (1-3). This thesis addresses one type of misclassification: discrepancies. The term “discrepancies” as used in this thesis means that what is written in the narrative by the investigating officer is inconsistent with the other fields recorded in the same traffic collision form. Figure 1.1 shows the front side of the South Carolina traffic collision form, which includes the narrative in Field 86, and Figure 1.2 shows the back side of the collision form, which contains the coded fields the narrative was compared against. An example of a discrepancy is shown in Figure 1.3 and Figure 1.4. The narrative describes Unit 2 as moving and Unit 3 as stopped in traffic, but the relevant form field has this information backward. From an applied perspective, misclassified data could lead to incorrect conclusions, and from a theoretical perspective, it could lead to severe bias in coefficient estimates and error rates from parametric and non-parametric models (4, 7-9).



Crash Number 1

SOUTH CAROLINA DPS/OHS & DMV USE ONLY 1a | Page # 2 Of: 2a | SOUTH CAROLINA TRAFFIC COLLISION REPORT FORM TR-310 (Rev. 04/2016) | # Of Units 3 | Amended - Attach Copy of Original Report / Corrected 4 | Notified 5 | Arrived 6

Date 7 | Time of Collision 8 | County 9 | 1 - Interstate 2 - US Primary 3 - SC Primary 4 - Secondary 5 - County 6 - PP 7 - Ramp 10 | On | Collision Location (Rt. # / Name) 11 / 12 | 0-Main 2-Alter 5-Spur 6-Connection -Business 9-Other 13 | Miles: 14 | Dir. N E S W 15 | In / Near City or Town of: 16

Lane # / Dir. # 17 Of 18 N E S W 19 | Distance Offset 20 20a Miles Feet | Direction N E S W 21 | 1- Interstate 2- US Primary 3- SC Primary 4 - Secondary 5 - County 6 - Other 7- Ramp 22 | From | Base Intersection (Rt. # / Name) 23 / 24 | 0-Main Line 2-Alternate 5-Spur 6-Connection 7-Business 9-Other 25 | GPS COORDINATES 00°00'00.00" DEGREES MINUTES SECONDS | Latitude 26 ° ' "

R.R. Id. 28 | From N E S W 29 | Ramp Only 1- Entrance 2- Exit 30 | To N E S W 31 | 1- Interstate 2- US Primary 3- SC Primary 4 - Secondary 5 - County 6 - Other 7- Ramp 32 | Toward | Second Intersection (Rt. # / Name) 33 / 34 | 0-Main Line 2-Alternate 5-Spur 6-Connection 7-Business 9-Other 35 | Longitude 27 ° ' "

SA-###### 36 | Driver/Pedestrian's Full Name 37 38 39 | SA ###### | Driver/Pedestrian's Full Name

Unit # 40 | Sex 42 | Race 43 | Street 44 | Unit # | Sex | Race | Street

Occ 41 | Birth Date 45 | City, State, & Zip 46 | # Occ | Birth Date | City, State, & Zip

State 47 | Driver's License # 48 | Class 49 | Insurance Company: 50 | State | Driver's License # | Class | Insurance Company:

Year 51 | Body 52 | Vehicle Make 53 | VIN # 54 | Year | Body | Vehicle Make | VIN #

State 55 | Year 56 | License Plate # 57 | Owner's D.L. # 58 | State | Year | License Plate # | Owner's D.L. #

Home Telephone ( ) 59 | Owner's Full Name 60 | Home Telephone ( ) | Owner's Full Name

Bus. Telephone ( ) 61 | Street 62 | Bus. Telephone ( ) | Street

Contributed To Collision Yes 63 No | City, State, & Zip 64 | Contributed To Collision Yes No | City, State, & Zip

Estimated Speed 65 | Speed Limit 66 | C.D.L. Req: Yes No 67 | T/B S Req: Yes No 68 | Alc/Drg info (see back): Yes No 69 | Statute # 70 | Statute # 71 | Towed By 72 | Yes No 72a | Estimated Speed | Speed Limit | C.D.L. Req: Yes No | T/B S Req: Yes No | Alc/Drg info (see back): Yes No | Statute # | Statute # | Towed By | Yes No

SA-###### | Driver/Pedestrian's Full Name | State | Year | License Plate # | Owner's D.L. #

Unit # | Sex | Race | Street | Home Telephone ( ) | Owner's Full Name

Occ | Birth Date | City, State, & Zip | Bus. Telephone ( ) | Street

State | Driver's License # | Class | Insurance Company: | Contributed To Collision Yes No | City, State, & Zip

Year | Body | Vehicle Make | VIN # | Estimated Speed | Speed Limit | C.D.L. Req: Yes No | T/B S Req: Yes No | Alc/Drg info (see back): Yes No | Statute # | Statute # | Towed By | Yes No

Dir. of Travel: | Unit 1: N S 73 E W | Unit 2: N S E W | Unit 3: N S E W

85

Unit 1 Dam. $ 74 | Unit 2 Dam. $ | Unit 3 Dam. $ | Prop.Dam. 1 $ 75 | Prop. Dam. 2 $ 76

Property Owner/Witness: 78 | Property Owner/Witness:

Address 79 | Address

State 80 | Zip: 81 | Phone 82 | State | Zip: | Phone

Photo Y N 83 | Describe What Happened (Refer to Units by Number) | Pending Investigation 84 Y N

86

NOTICE - THE TR-310 IS FOR STATISTICAL REPORTING PURPOSES ONLY AND IS A REFLECTION OF THE OFFICER'S BEST KNOWLEDGE, OPINION, AND BELIEF COVERING THE COLLISION, BUT NO WARRANT IS MADE AS TO THE FACTUAL ACCURACY THEREOF.

Investigating Officer's Name 87 | Rank 88 | SCCJA# 89 | Jurisdiction Code 90 | Review Date 91 | Reviewer's Name 92 | Rank 93 | Internal Agency Code 94

Figure 1.1 South Carolina Traffic Collision form TR-310, Front Side

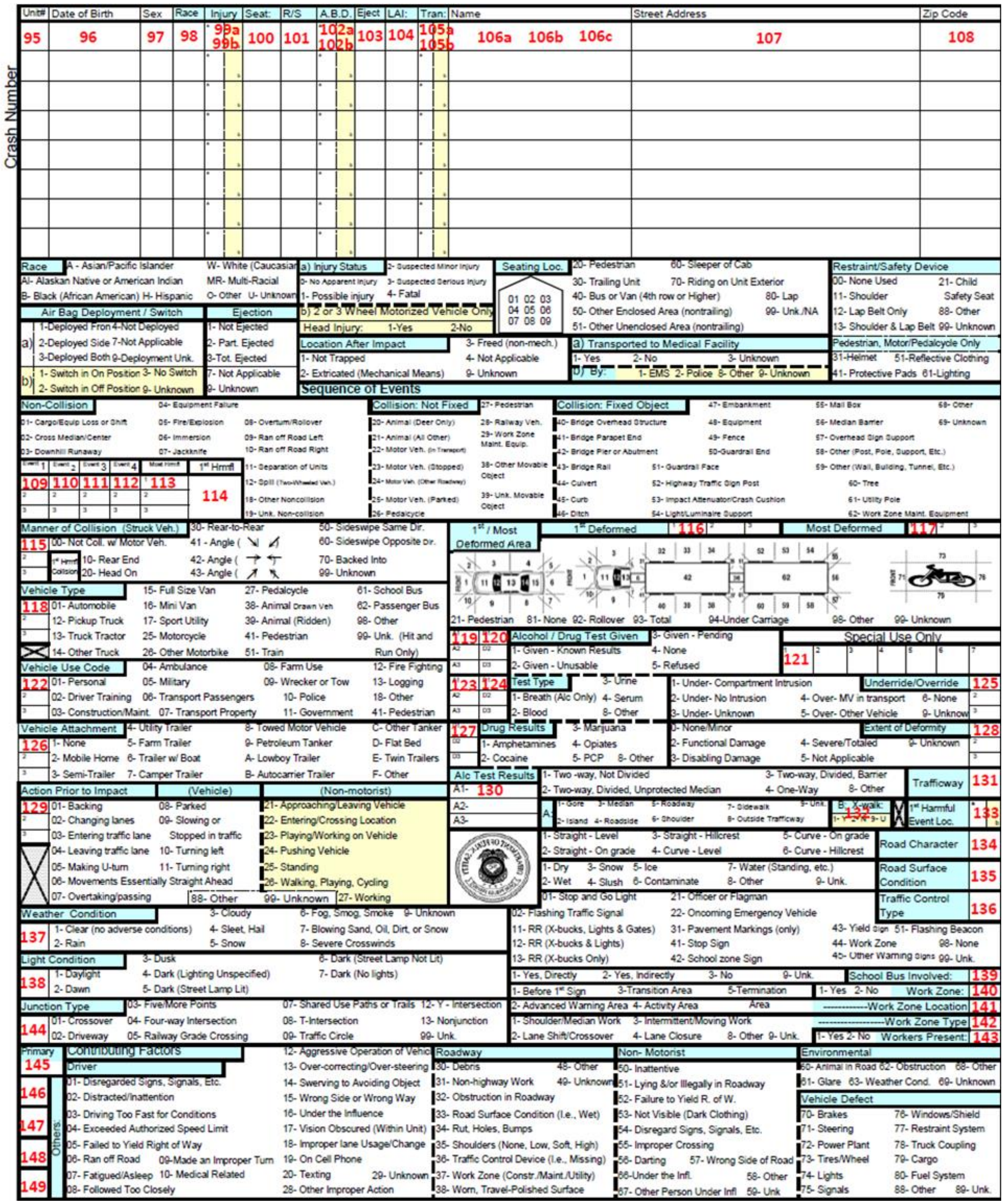


Figure 1.2 South Carolina Traffic Collision form TR-310, Back Side

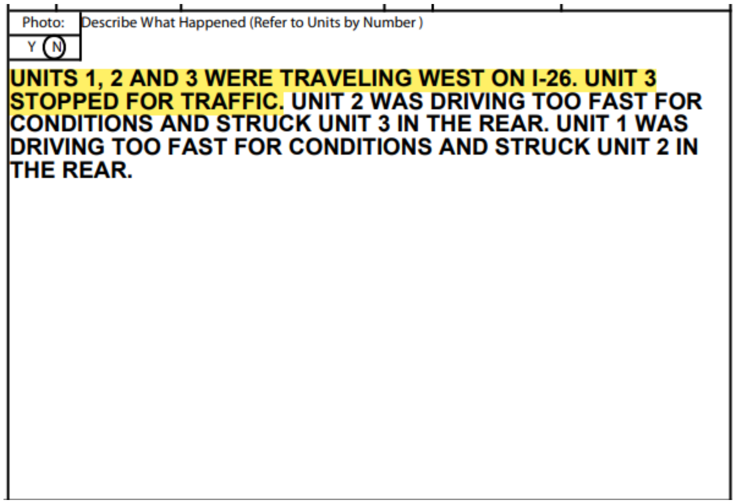
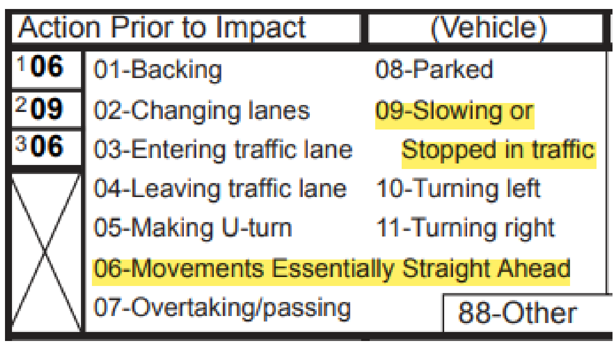


Photo: Describe What Happened (Refer to Units by Number)

Y N

UNITS 1, 2 AND 3 WERE TRAVELING WEST ON I-26. UNIT 3 STOPPED FOR TRAFFIC. UNIT 2 WAS DRIVING TOO FAST FOR CONDITIONS AND STRUCK UNIT 3 IN THE REAR. UNIT 1 WAS DRIVING TOO FAST FOR CONDITIONS AND STRUCK UNIT 2 IN THE REAR.

Figure 1.3 Discrepancy Example – Written Narrative (Field 86)



Action Prior to Impact		(Vehicle)
1 06	01-Backing	08-Parked
2 09	02-Changing lanes	09-Slowing or
3 06	03-Entering traffic lane	Stopped in traffic
	04-Leaving traffic lane	10-Turning left
	05-Making U-turn	11-Turning right
	06-Movements Essentially Straight Ahead	
	07-Overtaking/passing	88-Other

Figure 1.4 Discrepancy Example – Action Prior to Impact (Field 129)

Existing literature on the topic of misclassification primarily focuses on comparing information from traffic collision forms to external sources to identify misclassification. Some authors compared police crash reports to crash data sets collected by other agencies to assess their accuracy and comprehensiveness (2, 10, 11). Others compared traffic collision forms against unique sources, such as medical data and independent assessors, to determine their validity (1, 3). In cases where external sources were not used, the researchers relied on suggestive rather than explicit evidence to determine if misclassification occurred (12-14). To date, no studies have directly compared the narrative to other fields in traffic collision forms to understand the nature of discrepancies and their potential sources.

What distinguishes this thesis from previous misclassification studies is the determination of discrepancies within individual traffic collision forms by comparing the narrative text (Field 86) to information recorded in the form fields (see Table 3.2). In this thesis, the text in the narrative field is considered to have higher fidelity and is treated as the ground truth. Discrepancies between the narrative and form fields suggest that there are internal and external factors that affected the officer's cognitive ability to recall information and record it in a consistent manner. To this end, this thesis seeks to determine the level of discrepancies in South Carolina traffic collision forms and to identify factors that may have contributed to the discrepancies. The authors postulate that weather conditions and crash characteristics affect the process of recording crash information for the investigating officer. For example, the greater the number of vehicles involved in a crash, the more complex the situation, thereby requiring a higher level of processing by the officer to accurately fill out the form. The levels of processing theory states that the way information is encoded affects how well it is remembered. The deeper the level of processing, the easier the information is to recall (15). The psychology-based approach to understanding discrepancies in traffic collision forms is unique to this paper.

This thesis's objective is to improve the South Carolina Highway Patrol's (SCHP) accuracy of crash data reporting by identifying inaccuracies and their contributing factors. By identifying contributing factors, this thesis provides guidance on which areas of the reporting process the SCHP can focus on to improve its accuracy. To this end, it examines 200 randomly selected traffic collision forms out of 300 which involved a fatal crash in a work zone in South Carolina between 2014 and 2020. Error rates are determined for various fields on the form. Additionally, both structural equation modeling (SEM) and multiple linear regression (MLR) are used to identify factors that may have contributed to the discrepancies. Specifically, SEM is used to investigate the relationships between latent variables and level of processing, and MLR is used to investigate factors that affect the frequency of discrepancies in form fields.

CHAPTER 2: LITERATURE REVIEW

To my knowledge, this thesis is the first to consider the levels of processing concept in the context of crash data analysis. However, the levels of processing theory has been applied in other fields. A brief review is provided below. Regarding methods, SEM and MLR are used in this thesis. A review of related SEM studies is provided. As for MLR, its use in the area of crash data analysis is extensive. Previous work includes the prediction of injury severity score (16, 17), crash frequency (18, 19), and accident mortality rates (20, 21). Readers are referred to the work of Jiang et al. (22) and Edries and Alomari (23) for additional information regarding the application of MLR in crash data analysis.

Craik and Lockhart (1972) developed the levels of processing theory, in which information is understood through either shallow processing or deep processing, with deep processing leading to better memory retention (15). In linguistics, levels of processing theory is applied to language acquisition and recognition in native and non-native English speakers (24-27). In neuroscience, brain activity is measured for its reaction to different levels of processing tasks (28-30). In advertising, levels of processing theory is applied to memory retention from advertisement campaigns (31-33). In this thesis, levels of processing theory is applied to understand the complexity of a crash and how it affects discrepancies.

SEM has been applied in several traffic safety studies. Lee et al. (2018) used SEM to investigate how road, traffic and human, and rain and water-depth factors affected levels of accident severity of crashes in Seoul, South Korea (34). Boonyoo et al. (2021) applied

SEM to investigate how driver, road, environmental, and rear-end crash-specific factors affected the severity of rear-end collisions in Thailand (35). Kashani et al. (2021) used SEM to understand the influence of pedestrian, vehicle, environment, and road factors on measures of accident size for pedestrian-related crashes in Iran (36). Wang and Qin (2014) developed three SEM models to evaluate the factors affecting single-vehicle crash severity (37). Dong et al. (2022) used SEM to investigate how COVID-19 affected driver aggressiveness and inattentiveness, which in turn affected crash severity (38). The terms "accident size" and "crash severity" used in these studies refer to an overall measure of damage and injury caused by the crash which include injury severity. Yao and Wu (2012) used survey responses from E-bike riders in China to create an SEM model linking riders' safety perception and risk perception to aberrant riding behavior (39).

This thesis is the first to apply SEM to understand how the exogenous latent variables, crash characteristics and weather conditions, affect the endogenous latent variable, level of processing. The aim is to understand the relationship between the exogenous and endogenous variables and determine whether the level of processing has an impact on the observed discrepancies.

CHAPTER 3: DATA DESCRIPTION

Traffic collision forms (TR-310 forms) of fatal crashes occurring within work zones from 2014 to 2020 were provided by the South Carolina Department of Transportation (SCDOT) in PDF format as shown in Figures 1.1 and 1.2. Fields containing personal information were removed from the reports by the SCDOT. The information in the collision forms has been digitized by the SCDOT, and the digitized data were provided in a spreadsheet format. From the provided 300 traffic collision forms, 200 were randomly selected for review of discrepancies between the written narrative and the form fields on the traffic collision form. When information in a form field does not match the written narrative, the entire traffic collision form is classified as having a discrepancy. A summary of the frequency of discrepancies at the form level is shown in Table 3.1. It can be seen that 63.5%, 31%, and 5.5% of the forms contained 0, 1, and 2 discrepancies, respectively.

Table 3.1 Number of Forms with Discrepancies between Form Fields and Narrative

Number of Discrepancies	Traffic Collision Form Count
0	127
1	62
2	11

In addition to classifying discrepancies at the form level, the discrepancies were also counted at the field level. When multiple items in a field contain incorrect information, they were treated as a single discrepancy. For example, Fields 109 to 112 in Figure 1b capture the sequence of events following the action prior to impact. If the officer left out an event described in the written narrative, a correction would affect the entire sequence of fields. If only a single event was omitted, it is counted as one discrepancy. The discrepancies by form field are shown in Table 3.2. 17 distinct fields were investigated based on what information was included in the narrative. Given the reporting officers' conciseness in their descriptions, some narratives may not have contained information that could be compared to some of the 17 fields. As such, the selected fields represent the most common information available in the narrative, but not all fields could be compared to the narrative in every case. The fields with the most discrepancies were the sequence of events, action prior to impact, manner of collision, and contributing factors; their discrepancy rates are 31.0%, 21.4%, and 13.1%, respectively. Many of the fields had 0, 1, or 2 discrepancies.

Table 3.2 Number of Discrepancies by Form Fields

Discrepancy Type	Form Field Number(s)	Error Count
Sequence of Events	109-112	26
Most Harmful Event	113	1
First Harmful Event	114	1
Manner of Collision	115	11
Deformed Areas	116-117	7
Vehicle Type	118	0
Vehicle Attachments	126	1
Extent of Deformity	128	2
Action Prior to Impact	129	18
Trafficway Type	131	0
First Harmful Event Location	133	1
Road Character	134	0
Traffic Control Type	136	1
Work Zone Type	142	0
Worker Presence	143	0
Junction Type	144	2
Contributing Factors	145-149	11

The data set used for SEM considered each traffic collision form as an observation. Fields hypothesized to affect crash complexity include the number of units involved, the number of events describing the collision, collision speed, the number of alcohol or drug tests administered, and license class of the at-fault driver. The level of processing is operationalized by the number of discrepancies, the number of words in the narrative, and the number of characters in the narrative. This information was extracted from the traffic collision forms and the digitized data set. Additionally, weather station data for each crash was acquired from Local Climatological Data on a website managed by the National Oceanic and Atmospheric Administration (NOAA). A spreadsheet containing each station’s observations with date and time was obtained through the NOAA’s Geoportal. The weather station closest to the crash location was selected for each crash, and weather readings for the observation time closest to the police arrival time were used. The complete list of variables and their data types used for SEM analysis are shown in Table 3.3. It should be noted that because the SCDOT dataset was limited to only fatal work zone crashes, crash severity and work zone presence could not be used as variables, although they may indeed affect reporting accuracy.

Table 3.3 Variables Used for SEM Analysis

Data Source	Variable Name	Variable Type
Form TR-310	Number of Discrepancies	Discrete
	Number of Characters in Narrative	Discrete
	Number of Words in Narrative	Discrete
	Number of Units (Vehicles or Pedestrians) Involved in Crash	Discrete
	Number of Events (for all Units) in Crash	Discrete
	Collision Speed (mph)	Continuous
	Number of Alcohol/Drug Test Administered	Discrete
	License Class	Nominal
Weather Station Data from LCD	Dry Bulb Temperature (F)	Continuous
	Precipitation (in)	Continuous
	Relative Humidity (%)	Continuous
	Wind Speed (mph)	Continuous

The data set used for MLR considered each form field discrepancy to be an observation. With the help of experts from SCDOT, each observation was assigned a level of difficulty, with 0 denoting a relatively simple field, requiring only visual comprehension, and 1 a more complex field, requiring deeper comprehension. For instance, form fields 116-117 (Deformed Areas) were assigned a 0 due to their visual nature, whereas form fields 109-112 (Sequence of Events) were assigned a 1 due to the complexity of sequentially ordering the crash-related events. Each observation was also assigned a count of inputs and options. The input count was defined as the number of individual boxes within the field the officer could fill out. The option count was defined as the number of possible options the officer could select from. For example, in Figure 3.1, form field 126 (Vehicle Attachment) has an input count of 3 (for each of the boxes on the left) and an option count of 15 (for each of the options the officer can select from). Because the narrative only includes information regarding the crash and not personal driver information, only 17 form fields can be compared to the narrative. The data set used to estimate the MLR model is shown in Table 3.4.

Vehicle Attachment		4- Utility Trailer	8- Towed Motor Vehicle	C- Other Tanker
126	1- None	5- Farm Trailer	9- Petroleum Tanker	D- Flat Bed
2	2- Mobile Home	6- Trailer w/ Boat	A- Lowboy Trailer	E- Twin Trailers
3	3- Semi-Trailer	7- Camper Trailer	B- Autocarrier Trailer	F- Other

Figure 3.1 Form Field 126 (Vehicle Attachment)

Table 3.4 Reports by Type of Error

Form Location	Level of Difficulty	Error Count	Input Count	Option Count
109-112	1	26	12	51
113	1	1	3	12
114	1	1	1	12
115	1	11	3	11
116-117	0	7	6	61
118	0	0	3	18
126	0	1	3	15
128	0	2	3	6
129	1	18	3	20
131	0	0	1	5
133	1	1	2	11
134	0	0	1	6
136	0	1	1	16

CHAPTER 4: METHODOLOGY

4.1 Structural Equation Modeling (SEM)

SEM allows the relationship between different latent variables to be modeled. In this thesis, latent variables represent the different factors that could affect an officer's comprehension of the crash. These are weather conditions, crash characteristics, and level of processing. Latent variables are inherently unmeasurable and must be measured using observed variables. In this thesis, the observed variables are those shown in Table 3. These variables are not uniform in value. For example, the variable "Character Count" has values ranging from 56 to 761, while "Precipitation" has values ranging from 0 to 0.06 inches. Before proceeding with the SEM analysis, the variables' values were homogenized to the Likert scale with values ranging between 1 to 5, where 1 denotes the worst condition and 5 denotes the best condition.

First, hypothesized relationships between the observed variables shown in Table 3 and the latent variables were developed. The weather conditions factor is operationalized by wind speed, temperature, humidity, and precipitation. The crash characteristics factor is operationalized by the number of units, number of events, collision speed, license class, and the number of alcohol and/or drug tests administered. The level of processing factor is operationalized by the number of words in the narrative, the number of characters in the narrative, and the number of discrepancies in the form. Once the latent factors and their associated observed variables were defined, confirmatory factor analysis (CFA) was performed to test whether the data fit the hypothesized relationships. Once results were obtained from CFA, the SEM could be developed.

SEM consists of a structural model (the paths between latent variables) and measurement models (the relationship between each latent variable and its respective observed variables). Latent variables are called endogenous when they are dependent on another latent variable and exogenous when they are independent of other latent variables. For this thesis, the endogenous latent variable is level of processing, whereas the weather conditions and crash characteristics are exogenous. These factors were confirmed using Exploratory Factor Analysis (EFA) with Promax rotation. Each latent variable has a measurement model composed of the factor and its indicators. The exogenous variable measurement models can be expressed by the following equation.

$$x = \Lambda_x \xi + \delta$$

where x is a $(q \times 1)$ column vector of observed exogenous variables. δ is a $(q \times 1)$ column vector of measurement error terms for the observed variables in x. ξ is an $(n \times 1)$ column vector of latent exogenous variables. Λ_x is a $(q \times n)$ matrix of structural coefficients corresponding to the effects of the latent exogenous variables on their observed variables. The endogenous variable measurement model can be expressed by the following equation.

$$y = \Lambda_y \eta + \varepsilon$$

where y is a $(p \times 1)$ column vector of observed endogenous variables. ε is a $(p \times 1)$ column vector of measurement error terms for the observed variables in y. η is an $(m \times 1)$ column vector of the latent endogenous variable. Λ_y is a $(p \times m)$ matrix of structural coefficients corresponding to the effects of the latent endogenous variable on its observed variables.

The structural model consists of the exogenous variables weather conditions and crash characteristics, and the endogenous variable level of processing. Intuitively, this model resembles the levels of processing theory. Crash factors will affect crash complexity, and weather factors will likely have an impact on the officers' decision on how long to spend at the crash site. Both of these factors affect the level of processing the officer undergoes when filling out the traffic collision form. The structural model can be expressed by the following equation.

$$\eta = \beta\eta + \Gamma\xi + \zeta$$

where β is an $(m \times m)$ matrix of coefficients for the effects between latent endogenous variables. Since this thesis uses only one latent endogenous variable, the $\beta\eta$ term is zero. Γ is an $(m \times n)$ matrix of coefficients for the effects of latent exogenous variables on the latent endogenous variables. ζ is an $(m \times 1)$ column vector of error terms.

Three measures of model fit were used to assess the model: root mean squared error of approximation (RMSEA), Tucker-Lewis Index (TLI), and comparative fit index (CFI). The RMSEA measures goodness of fit based on the Chi-Square (χ^2) statistic and degrees of freedom (35, 37, 38). RMSEA is computed using the following equation.

$$RMSEA = \sqrt{\frac{\chi_M^2 - df_M}{df_M(N-1)}}$$

where χ_M^2 is the chi-squared test statistic for the model, df_M is the is the degrees of freedom, and N is the sample size. There are differing opinions on the maximum acceptable RMSEA value, but even the more stringent cutoffs agree a value less than 0.05 indicates good model fit (35, 37, 40-44). TLI and CFI are relative fit indices that compare to a baseline model to assess fit, but they differ in how they are affected by model complexity (45, 46). The equation for TLI is shown below.

$$TLI = \frac{\chi_B^2/df_B - \chi_M^2/df_M}{\chi_B^2/df_B - 1}$$

The equation for CFI is shown below.

$$CFI = 1 - \frac{max\ (\chi_M^2 - df_M, 0)}{max\ (\chi_B^2 - df_B, 0)}$$

where χ_B^2 and df_B are the χ^2 and degrees of freedom for the baseline model, respectively. Both CFI and TLI fall between 0 and 1, and values greater than 0.90 indicate the model to have good relative fit (47-49).

4.2 Multiple Linear Regression (MLR)

For MLR, the following assumptions are made: the residuals are normally distributed, there is a linear relationship between the dependent and independent variables, the variance of errors is consistent across independent variables (homoskedasticity), and the independent variables are independent (50-52). The data set used for the MLR model was assessed and found to satisfy the assumption criteria. A MLR model was created to assess the effect of level of processing, number of inputs, and number of options on the number of discrepancies by field type. The MLR model can be expressed as follows.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$$

where y is the expected value for the dependent variable (discrepancies), and x_i is the list of independent variables (level of difficulty, number of inputs, and number of options). β_0 is the value of y when the independent variables are all zero, and β_1 through

β_i are the regression coefficients for the independent variables x_i. ε is the error between the predicted and observed value for the dependent variable, or residual.

To assess goodness of fit, R-squared and adjusted R-squared were used. These values indicate the amount of variance explained by the model and range from 0 to 1, with a value of 1 indicating all variance can be explained by the model. Adjusted R-squared compensates for the addition of variables into a model (53-55).

CHAPTER 5: RESULTS

5.1 Structural Equation Modeling (SEM)

First, CFA was conducted to assess the fit of the proposed model. The results indicated good model fit, so the SEM model was developed. Both CFA and SEM analysis were performed using SPSS Amos. Figure 5.1 shows the SEM model results for the 200 traffic collision forms with coefficients standardized. The fit indices indicate that the SEM model is statistically significant, meaning its null hypothesis (crash characteristics and weather conditions affect level of processing) cannot be rejected: χ^2/df = 1.119 (<3), CFI = 0.986 (>0.9), TLI = 0.981 (>0.9), and RMSEA = 0.024 (<0.05). Overall, 76% of the variance in level of processing is explained by crash characteristics and weather conditions.

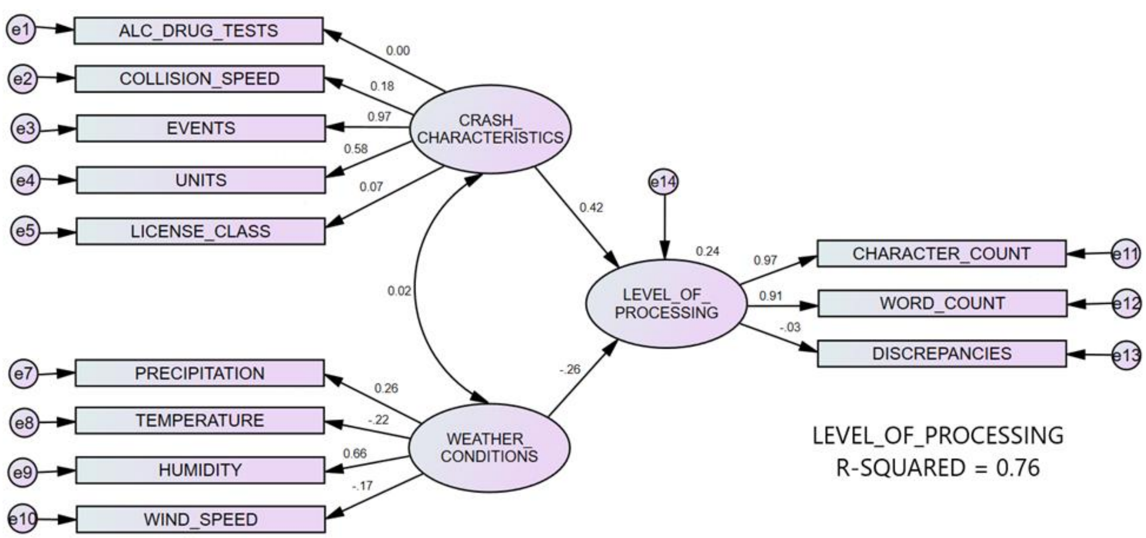


Figure 5.1 SEM Results

Due to the relatively small sample size, the 90% confidence level was used. At this threshold, several variables are significant. The structural model indicates the expected relationships between the latent variables. The coefficient estimate for the latent crash characteristics (0.42) indicates that it has a strong positive effect on the level of processing, whereas the coefficient estimate for the latent weather conditions (-0.26) indicates that it has a negative impact on the level of processing, meaning as the measure of poor weather conditions increases, the level of processing decreases. Since both of these variables are statistically significant, it can be concluded that crash characteristics and weather conditions positively and negatively affect the level of processing, respectively, with crash characteristics having a more significant role.

The measurement models indicate which observed variables are significant to the model. Out of the statistically significant variables affecting crash characteristics, the number of events, the number of units, and collision speed all have a positive effect on crash characteristics (0.97, 0.58, and 0.18, respectively). The number of events has the strongest effect. A higher value for any of these variables will result in an increase in the level of processing. Multiplying the coefficient estimate for any of these variables by the coefficient estimate for crash characteristics will give the effect of the variable on level of processing. Humidity and precipitation have positive effects on weather conditions (0.26 and 0.66, respectively), and thus will lower level of processing with an increase in value due to the negative relationship between weather conditions and level of processing. Temperature has the opposite effect because it has a negative relationship with weather conditions (-0.22), which in turn has a negative relationship with level of processing; an increase in temperature will increase the level of processing. Multiplying their coefficients shows a positive impact of temperature on level of processing. The number of words and characters in the narrative both have positive relationships with the latent variable level of processing (0.91 and 0.97, respectively), although the number of characters has a slightly stronger impact. As the level of processing increases, both the number of words and number of characters in the narrative will increase. To find the direct impact of any variable on the number of characters or words, simply multiply the coefficients forming the path between the variables. For example, the effect of precipitation on number of words would be the product of the coefficients 0.26, -0.26, and 0.91.

The variables not found to be statistically significant were number of alcohol and/or drug tests administered, license class, wind speed, and, most notably, the number of discrepancies (p = 0.87, 0.35, 0.23, and 0.78, respectively). For these variables, the model failed to reject the null hypothesis that each was not related to their respective latent variables. As such, it can be concluded that no variables in the model affect the occurrence of discrepancies. This result suggests that poorer weather conditions and crashes with a higher measure of complexity result in a longer written narrative (and vice versa), but these factors do not contribute to form discrepancies. Form discrepancies may be explained through the results of the MLR model examined below.

5.2 Multiple Linear Regression (MLR)

The MLR model estimation results are shown in Table 5.1. The model's R-squared and adjusted R-squared are 0.752 and 0.695, respectively, indicating very good model fit. At the 90% confidence level, the level of difficulty ($p = 0.054$) and input count ($p = 0.007$) are statistically significant. Their positive coefficients indicate that as the level of difficulty and/or input count increases, so will the number of discrepancies. That is, when the level of difficulty is complex instead of simple, the number of discrepancies can be expected to increase by 4.678. When the input count is increased by 1, the number of discrepancies can be expected to increase by 1.928. These findings correspond to intuition. That is, a field that is more difficult or requires more information to be entered is more likely to have discrepancies.

Table 5.1 MLR Model Estimation Results

Variable	β	Std. Error	*t*-value	*p*-value
(Intercept)	-2.670	1.592	-1.677	0.117
Level of Processing	4.678	2.213	2.114	0.054
Input Count	1.928	0.603	3.194	0.007
Option Count	-0.005	0.084	-0.064	0.950

CHAPTER 6: CONCLUSION

This thesis analyzed 200 traffic collision forms from fatal work zone crashes occurring in South Carolina between 2014 to 2020 to determine the number of discrepancies and potential contributing factors. A SEM model was developed to test whether factors related to weather conditions and crash complexity affected an officer's level of processing (indicated by the narrative length and presence of discrepancies) when filling out the traffic collision form. The SEM results showed that an increase in the number of units, number of events, collision speed, and temperature resulted in an increased number of characters and words written in the narrative, and that the narrative length is shortened with an increase in precipitation and humidity. Notably, the number of discrepancies occurring within the form was not found to be statistically significant. In addition, an MLR model was developed to test whether factors related to the structure of form fields affect the frequency of discrepancies. This thesis's findings indicated that the frequency of discrepancies in a form field will increase with additional inputs or if it has a higher level of difficulty. Additional study may be required to draw definitive conclusions regarding discrepancies in traffic collision forms.

Based on this thesis's findings, it can be concluded that officers in South Carolina are doing their job well in filling out the traffic collision forms. That is, they do not let the circumstances surrounding the crash, such as its complexity and weather conditions, affect their ability to process information and record it. Should the traffic collision form need to be modified in the future, the new fields should be kept as simple as possible with minimum input boxes. This thesis has several limitations that need to be considered when interpreting its findings. First, the provided traffic collision forms are limited to fatal crashes occurring within work zones. Analyzing traffic collision forms of other injury severity levels may yield different results. Along this line, crashes occurring within work zones are a relatively small subset of all traffic crashes. Future work that analyzes traffic collision forms not occurring within work zones may yield different results. Second, the narrative text does not allow for all fields to be validated. Thus, the number of discrepancies is likely to be more than what was identified in this thesis. Third, police officers used an app to fill out the traffic collision form rather than a pen and paper. As such, discrepancies could be due to errors in inputting the information rather than the inability to accurately recall the crash information. Fourth, because personal information was removed from the forms by SCDOT, this thesis did not investigate how demographic factors (i.e., age, gender, or race of involved drivers) affect the officer's level of processing. Fifth, in some cases, officers may not include enough information in their narratives to compare to all 17 form fields. Subsequently, some inaccuracies may have been unidentified because the officer omitted information that could result in a discrepancy. Lastly, because officer training varies across states, the findings in this thesis cannot be generalized to the entire nation. Future work could compare discrepancy rates across different states to assess officer training quality.

REFERENCES

1. Fotios S, Robbins C. Incorrect categorisation of ambient light level at the time of a road traffic collision. Lighting Research & Technology 2022: 14771535211069028.

2. Shinar D, Treat JR, McDonald ST. The validity of police reported accident data. Accident Analysis & Prevention 1983; 15: 175-191.

3. Amoros E, Martin J-L, Chiron M, Laumon B. Road crash casualties: characteristics of police injury severity misclassification. Journal of Trauma and Acute Care Surgery 2007; 62: 482-490.

4. Hausman JA, Abrevaya J, Scott-Morton FM. Misclassification of the dependent variable in a discrete-response setting. Journal of econometrics 1998; 87: 239-269.

5. Ahmed A, Sadullah AFM, Yahya AS. Errors in accident data, its types, causes and methods of rectification-analysis of the literature. Accident Analysis & Prevention 2019; 130: 3-21.

6. Ullman GL, Scriba TA. Revisiting the Influence of Crash Report Forms on Work Zone Crash Data. Transportation Research Record 2004; 1897: 180-182.

7. Bi Y, Jeske DR. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. Journal of Multivariate Analysis 2010; 101: 1622-1637.

8. Meyer BD, Mittag N. Misclassification in binary choice models. Journal of Econometrics 2017; 200: 295-311.

9. Höfler M. The effect of misclassification on the estimation of association: a review. International journal of methods in psychiatric research 2005; 14: 92-101.

10. Farmer CM. Reliability of police-reported information for determining crash and injury severity. 2003.

11. Blackman R, Debnath AK, Haworth N. Understanding vehicle crashes in work zones: Analysis of workplace health and safety data as an alternative to police-reported crash data in Queensland, Australia. Traffic injury prevention 2020; 21: 222-227.

12. Filtness AJ, Armstrong KA, Watson A, Smith SS. Sleep-related crash characteristics: Implications for applying a fatigue definition to crash reports. Accident Analysis & Prevention 2017; 99: 440-444.

13. Yahaya M, Fan W, Fu C, Li X, Su Y, Jiang X. A machine-learning method for improving crash injury severity analysis: A case study of work zone crashes in Cairo, Egypt. International journal of injury control and safety promotion 2020; 27: 266-275.

14. Sayed MA, Qin X, Kate RJ, Anisuzzaman D, Yu Z. Identification and analysis of misclassified work-zone crashes using text mining techniques. Accident Analysis & Prevention 2021; 159: 106211.

15. Craik FI, Lockhart RS. Levels of processing: A framework for memory research. Journal of verbal learning and verbal behavior 1972; 11: 671-684.

16. Majdzadeh R, Eshraghian MR, Khalagi K, Motevalian A, Naraghi K. Crash-related factors associated with the severity of road traffic injuries in Iran. International journal of injury control and safety promotion 2011; 18: 175-180.

17. Ryb GE, Dischinger PC. Injury severity and outcome of overweight and obese patients after vehicular trauma: a crash injury research and engineering network (CIREN) study. Journal of Trauma and Acute Care Surgery 2008; 64: 406-411.

18. Chliaoutakis JE, Demakakos P, Tzamalouka G, Bakou V, Koumaki M, Darviri C. Aggressive behavior while driving as predictor of self-reported car crashes. Journal of safety Research 2002; 33: 431-443.

19. Sivak M, Schoettle B, Reed MP, Flannagan MJ. Body-pillar vision obstructions and lane-change crashes. Journal of safety research 2007; 38: 557-561.

20. O'Neill B, Kyrychenko SY. Use and Misuse of Motor-Vehicle Crash Death Rates in Assessing Highway-Safety Performance. Traffic Injury Prevention 2006; 7: 307-318.

21. La Torre G, Van Beeck E, Quaranta G, Mannocci A, Ricciardi W. Determinants of within-country variation in traffic accident mortality in Italy: a geographical analysis. International Journal of Health Geographics 2007; 6: 49.

22. Jiang Z-H, Yang X-G, Sun T, Wang T, Yang Z. Investigating the relationship between traffic violations and crashes at signalized intersections: an empirical study in China. Journal of advanced transportation 2021; 2021: 1-8.

23. Edries B, Alomari AH. Forecasting the Fatality Rate of Traffic Accidents in Jordan: Applications of Time-Series, Curve Estimation, and Multiple Linear Regression Models. Journal of Engineering Science & Technology Review 2022; 15.

24. Ayçiçegi-Dinn A, Caldwell-Harris CL. Emotion-memory effects in bilingual speakers: A levels-of-processing approach. Bilingualism: Language and Cognition 2009; 12: 291-303.

25. Hui B, Godfroid A. Testing the role of processing speed and automaticity in second language listening. Applied Psycholinguistics 2021; 42: 1089-1115.

26. Bird S. Expert knowledge, distinctiveness, and levels of processing in language learning. Applied Psycholinguistics 2012; 33: 665-689.

27. Morris CD, Bransford JD, Franks JJ. Levels of processing versus transfer appropriate processing. Journal of verbal learning and verbal behavior 1977; 16: 519-533.

28. Kapur S, Craik FI, Tulving E, Wilson AA, Houle S, Brown GM. Neuroanatomical correlates of encoding in episodic memory: levels of processing effect. Proceedings of the National Academy of Sciences 1994; 91: 2008-2011.

29. Van Veen V, Cohen JD, Botvinick MM, Stenger VA, Carter CS. Anterior cingulate cortex, conflict monitoring, and levels of processing. Neuroimage 2001; 14: 1302-1308.

30. Rose NS, Craik FI, Buchsbaum BR. Levels of processing in working memory: Differential involvement of frontotemporal networks. Journal of cognitive neuroscience 2015; 27: 522-532.

31. Nordhielm CL. The influence of level of processing on advertising repetition effects. Journal of consumer research 2002; 29: 371-382.

32. Wilson RT, Lohmeier JH, Lustick DS, Chen RF. Using transit advertising to improve public engagement with social issues. International Journal of Advertising 2021; 40: 783-809.

33. Saegert J. A demonstration of levels-of-processing theory in memory for advertisements. ACR North American Advances 1979.

34. Lee J, Chae J, Yoon T, Yang H. Traffic accident severity analysis with rain-related factors using structural equation modeling–A case study of Seoul City. Accident Analysis & Prevention 2018; 112: 1-10.

35. Boonyoo T, Champahom T, Ratanavaraha V. ANALYSIS OF FACTORS AFFECTING REAR-END CRASH SEVERITY USING STRUCTURAL EQUATION MODELING. Suranaree Journal of Science & Technology 2022; 29.

36. Kashani AT, Jafari M, Bondarabadi MA. A new approach in analyzing the accident severity of pedestrian crashes using structural equation modeling. Journal of injury and violence research 2021; 13: 23.

37. Wang K, Qin X. Use of structural equation modeling to measure severity of single-vehicle crashes. Transportation Research Record 2014; 2432: 17-25.

38. Dong X, Xie K, Yang H. How did COVID-19 impact driving behaviors and crash Severity? A multigroup structural equation modeling. Accident Analysis & Prevention 2022; 172: 106687.

39. Yao L, Wu C. Traffic safety for electric bike riders in China: attitudes, risk perception, and aberrant riding behaviors. Transportation research record 2012; 2314: 49-56.

40. Shi J, Bai Y, Tao L, Atchley P. A model of Beijing drivers' scrambling behaviors. Accident Analysis & Prevention 2011; 43: 1540-1546.

41. Champahom T, Jomnonkwao S, Karoonsoontawong A, Hantanong N, Beeharry R, Ratanavaraha V. Modeling user perception of bus service quality: case study of Mauritius. Songklanakarin J Sci Technol 2020; 42: 660-670.

42. Steiger JH. Understanding the limitations of global fit assessment in structural equation modeling. Personality and Individual Differences 2007; 42: 893-898.

43. Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit. Sociological Methods & Research 1992; 21: 230-258.

44. Hair J, Black W, Babin B, Anderson R, Tatham R. Multivariate Data Analysis 6th Edition Pearson Education Inc. USA: 898pp 2006.

45. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 1980; 88: 588-606.

46. Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin 1990; 107: 238-246.

47. Yu C-Y. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes: University of California, Los Angeles; 2002.

48. Hamdar SH, Schorr J. Interrupted versus uninterrupted flow: a safety propensity index for driver behavior. Accident analysis & prevention 2013; 55: 22-33.

49. Hassan HM, Abdel-Aty MA. Exploring the safety implications of young drivers' behavior, attitudes and perceptions. Accident Analysis & Prevention 2013; 50: 361-370.

50. Uyanık GK, Güler N. A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences 2013; 106: 234-240.

51. Williams MN, Grajales CAG, Kurkiewicz D. Assumptions of multiple regression: Correcting two misconceptions. Practical Assessment, Research, and Evaluation 2013; 18: 11.

52. Osborne JW, Waters E. Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation 2019; 8: 2.

53. Eberly LE. Multiple linear regression. Topics in Biostatistics 2007: 165-187.

54. Jobson J. Multiple linear regression. Applied multivariate data analysis: Regression and experimental design 1991: 219-398.

55. Fávero LP, Belfiore P, de Freitas Souza R. Chapter 14 - Simple and multiple regression models. In: Fávero LP, Belfiore P, de Freitas Souza R, editors. Data Science, Analytics and Machine Learning with R: Academic Press; 2023: p. 237-258.