Final Report

Reducing Crash Risk at Work Zones in South Carolina

Sponsoring Agencies:

South Carolina Department of Transportation

Federal Highway Administration

Principal Investigator:

Nathan Huynh

Department of Civil and Engineering

University of South Carolina

June 2024

Technical Report Documentation Page

1. Report No. FHWA-SC-24-04	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Reducing Crash Risk at Work Zones in South Carolina	5. Report Date June 2024
	6. Performing Organization Code
7. Author(s) Nathan Huynh, Mahyar Madarshahian, and Jackson Wegmet	8. Performing Organization Report No. University of South Carolina
9. Performing Organization Name and Address University of South Carolina Department of Civil and Environmental Engineering 300 Main Street Columbia, SC 29208	10. Work Unit No. (TRAIS) 11. Contract or Grant No. SPR No. 760
12. Sponsoring Agency Name and Address South Carolina Department of Transportation PO Box 191 Columbia, SC 29202-0191	13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This project investigated several aspects related to work zone safety. First, it examined the accuracy of the information recorded in the South Carolina (SC) traffic collision report forms (TR-310). A total of 200 fatal crashes in work zones between 2014 and 2020 were examined to determine how many discrepancies exist between the written narrative and other fields in the traffic collision forms. Of the 200 forms, 63.5%, 31%, and 5.5% contained 0, 1, and 2 discrepancies, respectively. Second, it determined factors that contributed to injury in work zone-related crashes in SC on roads with 60 mph or higher speed limits and those with speed limits less than 60 mph using crash data from 2014 to 2020 in SC. Results from mixed binary (injury or no injury) logit models indicate there are common factors: dark lighting conditions, female (at-fault) drivers, and excessive speed. Factors that contributed to injury on roads with speed limits less than 60 mph are work zone on an SC or US primary road, work zone activity area, at-fault drivers under 35, sideswipe collisions, and worker presence. Contributing factors on roads with 60 mph or higher speed limits are number of vehicles involved, rear-end collisions, proximity to first work zone signs, and weekdays. Additionally, a mixed binary (injury or no injury) logit model with heterogeneity in mean and variance was used to determine factors that contributed to injury severity of work zone rear-end crashes with collision speeds over 35 mph. Factors that increased injury severity include multi-vehicle involvement, airbag deployment, dark conditions, and truck involvement. Conversely, factors that reduced injury severity are late-night and dawn/dusk conditions, advanced warning areas, work zone activity areas, lane shifts/crossover work zones, and young and middle-aged at-fault drivers. Third, it assessed the effectiveness of the presence of law enforcement at work zones. Using a split-plot design with blocking, eight models were explored. These models use two types of response variables, average speed in the entire work zone and average speed in the transition area. For each type of response, two variations were assessed. The first is to subtract the temporary posted speed limit from the average speed, and the other variation is to include traffic volume as a covariate. The results indicated that traffic speeds at work zones were lower when law enforcement was present. Fourth and lastly, it developed a predictive work zone risk assessment tool to enable SCDOT engineers to determine crash risk and benefit-cost of implementing countermeasures. This tool was implemented in Excel using VBA and estimation results from a zero-inflated negative binomial model to predict the expected number of crashes given a work zone's length, duration, and AADT.

17. Key Words Work zone injury contributing factors, police crash report form discrepancies, law enforcement effectiveness, countermeasures benefit-cost	18. Distribution Statement No restrictions.

19. Security Classif. (of this report) Unclassified.	20. Security Classif. (of this page) Unclassified.	21. No. of Pages 74	22. Price

Disclaimer

The contents of this report reflect the views of the author who is responsible for the facts and the accuracy of the data presented. The contents do not necessarily reflect the official views of the South Carolina Department of Transportation or Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

The State of South Carolina and the United States Government do not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the object of this report.

Acknowledgments

The authors greatly appreciate the guidance and assistance from the following Project Steering and Implementation Committee Members:

- Dr. Chowdhury Siddiqui (Project Chair)
- Joey Lucas
- Emily Thomas
- Andrew Stokes
- Jeremy Yuhas
- James Remsey
- Carolyn Fisher
- Merrill Zwanka

Executive Summary

Stroup et al. (2018) sought to review work zone-related crash reports to verify the reported information, Fotios and Robbins (2024) identify factors that contribute to work zone-related crashes in South Carolina, Shinar et al. 1983 identify countermeasures based on said factors, Amoros et al. (2007) understand the impact of the presence of law enforcement at work zones, and Hausman et al. (1998) develop a predictive work zone risk assessment tool to proactively assess the risk at the beginning and during the lifespan of a project. Different statistical models were developed to achieve each objective.

A total of 200 forms containing information about fatal crashes in work zones between 2014 and 2020 were analyzed to determine how many discrepancies exist between the written narrative and other fields. 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) was developed. SEM results show that increases in collision speed, number of units, number of events, and temperature increased the number of words and characters written in the narrative, whereas increases in precipitation, humidity, and poor weather conditions 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 was 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 were found to be statistically significant.

Utilizing crash data spanning from 2014 to 2020, two models were developed using mixed logit models to find contributing factors affecting crash injury (versus no injury): one tailored for non-interstate roads with speed limits below 60 miles per hour (mph), and another tailored for interstates with speed limits of 60 mph or higher. The findings indicated the necessity for separate models based on speed. Common factors contributing to injury across both models encompass dark lighting conditions, female (at-fault) drivers, and driving too fast for conditions. Furthermore, factors impacting injury on non-interstate roadways include SC or US primary roadways, work zone activity area, at-fault drivers under 35, sideswipe collisions, presence of workers, and collisions with fixed objects. Conversely, factors affecting injury on interstates include the number of vehicles involved, rear-end collisions, proximity to the first work zone sign, and crashes occurring on weekdays.

To determine factors influencing injury (versus no injury) in a work zone, rear-end crashes with collision speeds over 35 mph, a mixed binary logit model with heterogeneity in both mean and

variance was developed. Significant factors contributing to injury included multi-vehicle involvement, airbag deployment, dark conditions, and crashes involving trucks. Conversely, late-night and dawn/dusk conditions, along with variables such as advanced warning areas, activity zones, lane shifts/crossovers, and the presence of young and middle-aged at-fault drivers were associated with no injury.

A split plot design with blocking was used to investigate the effectiveness of law enforcement on speed reduction in South Carolina work zones. The analysis used speed as the response variable, seasons as the main plots, the presence of law enforcement as subplots, and traffic volume as a covariate. Using data from 2019, eight alternative models were explored to determine whether the speed for the entire work zone should be considered or just the speed at the locations where troopers were stationed. Additionally, the models examined whether the average speed of traffic or the speed exceeding the temporary posted speed limit (excess speed) should be used. These combinations were considered with and without traffic volume as a covariate. All eight models showed a reduction in traffic speed when law enforcement was present. The model with the best fit is the one that considers excess speed for the entire work zone without having traffic volume as a covariate. However, the ANCOVA analysis found the covariate to be efficient. Thus, it is recommended that traffic volume be included in future analyses. Seasonal analysis indicated that throughout the entire work zone, there is no difference in average traffic speed and excess speed between seasons. However, for the transition area, the average traffic speeds and excess speed were lower in the winter compared to fall and summer. In the absence of troopers, there is no variation in speeds between seasons.

The work zone assessment tool was developed in Excel using Visual Basic for Applications (VBA) to enable SCDOT engineers to determine crash risk and the benefit/cost of implementing countermeasures at a work zone. Countermeasures and their associated Crash Modification Factors (CMFs) developed specifically for work zones by researchers from the University of Missouri were adopted. To determine the benefit-to-cost ratio, specifically, the estimated crash cost savings divided by the cost of implementing the measures, a crash prediction model was developed to estimate the expected number of crashes in work zones based on their length, duration, and Annual Average Daily Traffic (AADT). This model used work zone data manually extracted from ProjectWise for four types of work zones: widening, rehabilitation, reconstruction, and preservation. Work zone length and duration were computed from project descriptions, while AADT was determined by averaging traffic counts from count stations within work zone boundaries. Due to the high number of work zones experiencing zero crashes, a zero-inflated Negative Binomial model was developed instead of the traditional Negative Binomial model. The models indicated that work zone length, duration, and log of AADT were significant in predicting crash counts across different work zone types.

Table of Contents

List of Figures

List of Tables

1. Introduction

The 2020–2024 South Carolina Strategic Highway Safety Plan (SHSP) identified 12 emphasis areas based on a detailed analysis of statewide crash data. Among these are work zones due to highway workers being vulnerable users. Work zones alter the normal traffic flow requiring motorists to change their speeds, process information from roadside signs, make merging maneuvers, and travel next to cones or barricades. These activities can lead to vehicular crashes and injury to motorists and workers in the work zone. Figure 1 shows the trend in the total number of crashes and the number of fatalities in South Carolina work zones from 2014 to 2020.

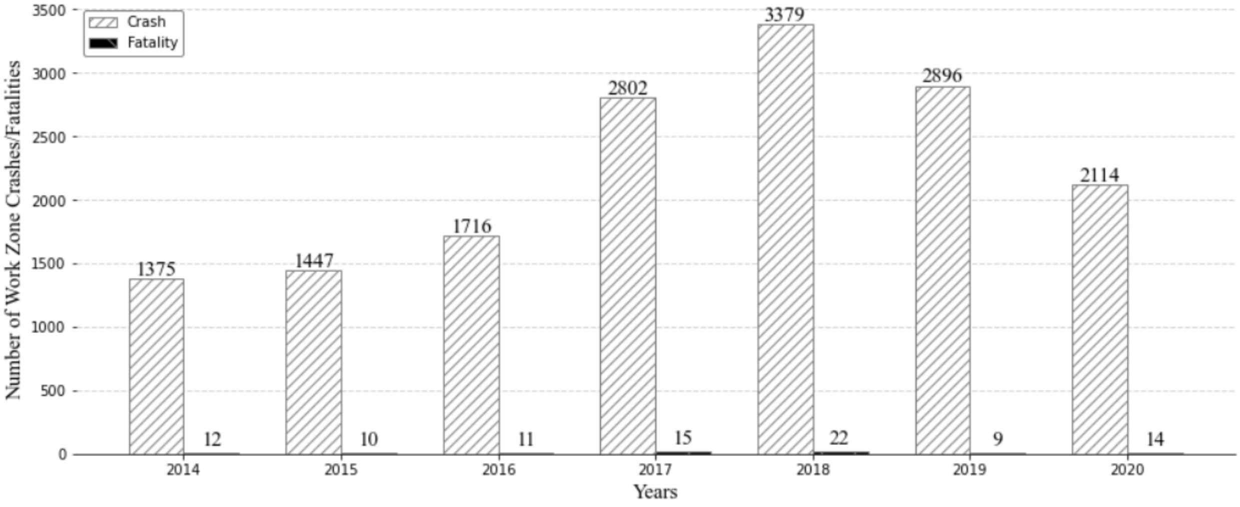


Figure 1. Total number of work zone crashes and fatalities from 2014 to 2020 in South Carolina.

The South Carolina Department of Transportation (SCDOT) has made several concerted efforts to improve work zone safety. On June 2, 2006, the agency entered into an agreement with the South Carolina Department of Public Safety where Highway Patrol Troopers would devote their time to selective, concentrated, and strict enforcement of the state's traffic laws at work zones. In 2016, the SCDOT created the Procedures and Guidelines for Work Zone Traffic Control Design document aimed at reducing work zone collisions. Despite these efforts and the introduction of the Workers' Safety Act (House Bill 4033) in 2017 where penalties include fines, jail time, and points assessed against an offender's driving record, the number of work zone-related crashes has remained high as shown in Figure 1, which suggest that more can be done to improve work zone safety. This fact is recognized by the 2020–2024 SHSP and it outlined several strategies for the SCDOT and South Carolina Department of Public Safety (SCDPS) to implement to get closer to its "target zero" goal. These strategies include improving data collection for work zone-related collisions, improving driver compliance with work zone traffic controls, and increasing public knowledge and awareness of work zones.

This project sought to contribute to the mission and commitment of the SCDOT and SCDPS to increase work zone safety through the following five research objectives:

1. Review work zone-related crash reports to verify the reported information.
2. Identify factors that contribute to work zone-related crashes in SC.
3. Identify countermeasures based on factors identified in objective 2.
4. Understand the impact of the presence of law enforcement at work zones.
5. Develop a predictive work zone risk assessment tool to proactively assess the risk at the beginning/during the lifespan of a project.

The aim of research objective 1 is to enhance the collection of crash data at work zones so that the SCDOT can identify high-risk work zone locations and activity areas, and to improve the accuracy of the South Carolina Traffic Collision Fact Books. While it is known from the 2020–2024 SHSP that "the most frequently reported contributing factors in work zone-related fatal and serious injury collisions are driving too fast for conditions and failure to yield right of way," research objective 2 sought to provide the SCDOT with a more comprehensive understanding of the contributing factors by examining multiple data sources (i.e., crash, unit, and occupant). Research objectives 3 and 4 aim to identify strategies that have a high likelihood of being successful in South Carolina. Lastly, research objective 5 aims to provide the SCDOT with a practical, useful, and impactful work zone risk assessment tool.

The next chapter (Chapter 2) presents a literature review of related work and results from a survey of state DOTs on data collection for work zone crashes. Chapter 3 describes the procedures used to synthesize the data for analysis and the methods used to model the acquired data. Chapter 4 presents the findings from the statistical models. Chapter 5 includes a discussion of the model findings and explains the project deliverables. Lastly, Chapter 6 presents this study's conclusions, recommendations, and implementation plan.

2. Literature Review

Many studies have investigated factors that contribute to work zone crashes. The majority can be categorized as focusing on either injury severity (most often defined as the most severely injured person involved in the crash) or crash frequency (the rate of occurrence). Some studies considered both aspects, while others considered neither injury severity nor crash frequency. The authors from prior studies employed a variety of parametric and non-parametric methods to understand injury severity and crash frequency of work-zone-related crashes. Consequently, the work zone literature was grouped as follows: injury severity, crash frequency, combined severity and frequency, and additional contributing factors.

2.1. Injury Severity

The following includes studies that investigated contributing factors to injury severity in work zone-related crashes. The sub-sections (2.1.1, 2.1.2, 2.1.3, and 2.1.4) group papers focused on work zone injury severity as it relates to truck involvement, lighting conditions, work zone type, and area within the work zone respectively.

Li and Bai (2008) developed a crash severity index (CSI) to evaluate the risk of an accident being fatal within a work zone. Data were obtained from the Kansas Department of Transportation (KDOT) database and included both fatal crashes from 1998 to 2004 and injury crashes from 2003 to 2004. The procedure to develop work zone CSI models began with identifying contributing factors to work zone crashes, and then the models themselves were developed using logistic regression. Last, prediction accuracy was tested using the most recent crash data. Four models were developed: two models were first created depending on the identified contributing factors, resulting in driver-independent and driver-dependent models. Simplified models for each of the two types were created by dropping non-statistically significant variables from each of the models. Model validation found that crash severity prediction was generally accurate for injury work zone crashes, but less so for fatal crashes.

Weng and Meng (2011) developed a tree-based logistic regression model for work zone crashes to assess the vehicle occupant's casualty risk. Work zone crash injury data were collected from a database maintained by the University of Michigan Transportation Institute. The decision tree was built using sampled data. Based on the tree structure, the sample data was split into separate groups. A logistic regression model was then built for each of the groups. It was found that interacting variables were airbag, occupant identity, and gender. The tree-based logistic regression model was found to give more accurate predictions for injury events when compared to a pure logistic regression model.

Liu et al. (2016) investigated how pre-crash behavior affects injury severity in work zone crashes. Data were acquired from the Virginia 2013 statewide crash database, which itself was derived from Virginia police crash reports. A hierarchical modeling methodology was applied to focus on pre-crash driver actions, which are nested within driver-vehicle characteristics. The study found that improper actions, such as following too closely or speeding, had a high correlation with injuries. Additionally, not using seat belts and driving under the influence of alcohol or drugs were associated with severe injuries.

Zhang et al. (2018) developed a hybrid approach of combining factor analysis with ordered probit model to determine the most significant factors affecting work zone crash severity in Egypt. Data were pulled from a database maintained by Egypt's General Authority for Roads, Bridges and Land Transport (GARBLT) from 2010 to 2015. The factor analysis first determined both main and common factors in determining crash severity. With these results, the ordered probit model was calibrated using three levels of crash severity (no injury, injury, and fatal injury) to determine the most influential factors. Four factors, the most influential being weather conditions, were found to significantly affect work zone crash severity.

Ghasemzadeh and Ahmed (2019) utilized a probit-classification tree to identify factors contributing to injury severity of work zone crashes in adverse weather conditions. Crash data were extracted from the Strategic Highway Research Program 2 (SHRP2) Roadway Information Dataset (RID) in Washington State from 2006 to 2013. This technique combined the conventional parametric probit regression model with a nonparametric classification tree model to compensate for the disadvantages of both individual models. Relevant factors usually in the conventional probit model were included, such as vehicle type and age, lighting, and weather conditions. It was found that the presence of a traffic control device and lighting conditions were significant interacting variables, and the authors recommended installing countermeasures to compensate for weather conditions.

Sze and Song (2019) examined which risk factors contributed to work zone crashes involving fatalities or severe injuries in New Zealand. Data were extracted from New Zealand's Crash Analysis System, where work zone crashes were selected from November 25, 2008, to November 25, 2013, for locations where the speed limit was temporarily reduced. Notably, data did not include information on the type of road, road environment, or characteristics of the work zone itself. A multinomial logistic regression model using a 20% level of significance was applied to determine contributing factors, and injuries were grouped into fatal/serious, minor, and non-injury. Day of week, time of day, and involvement of motorcycles/bicycles/pedestrians were found to affect the likelihood of fatal/serious and minor injury.

Zhang and Hassan (2019a) investigated the contributing factors leading to work zone rear-end crashes in Egypt. Data on crashes were acquired from Egypt's Ministry of Transport for 12 long-term (longer than one year) work zones in Egypt from 2010 to 2017. Six categories were acquired from the data: driver information, vehicle information, crash time, road characteristics, work zone information, and environmental conditions. A random parameter ordered probit model, which allows for unobserved heterogeneity, was utilized, and injury severity was categorized as no injury, injury, and fatal. Findings include that unexpected maneuvers and young male drivers traveling at nighttime on weekends both increased the chances of fatality, and injury severity is higher during asphalt surface construction than milling surface construction. The authors recommended driver training programs and intelligent transportation systems (ITS) technologies as countermeasures to reduce the number of rear-end crashes.

Zhang and Hassan (2019b) aimed to determine the difference in injury severity and contributing factors in work zone crashes during daytime and nighttime. The data used were from ten long-term (longer than one year) work zones in Egypt from 2010 to 2016. Separate mixed multinomial logit models were used for day and night, with each separating injuries into three categories: property damage only, injury, and fatality. Likelihood ratio tests statistically justified the usage of separate models for daytime and nighttime. It was found that significant factors had substantial differences between day and night models. Even in cases where variables were significant in both, they displayed different directions or magnitudes of effect across models.

Mokhtarimousavi et al. (2019) compared the performance of two different approaches, one parametric and one non-parametric, in predicting the injury severity of work zone crashes. Crash records from 2013 to 2017 in Miami-Dade County were obtained from the Florida Signal Four analytics tool. The parametric approach utilized a mixed logit modeling framework to predict crashes. The non-parametric approach utilized support vector machine (SVM) modeling and applied three unique optimization algorithms to determine which most improved results. The base SVM model and all three applied algorithms outperformed the mixed logit model in correctly predicting observed crashes. The SVM model including the harmony search algorithm was found to perform best, with an accuracy of 83.5% compared to the mixed logit model's 67.2%.

Yu et al. (2020) analyzed the injury severity of rear-end work zone crashes and their contributing factors. Work zone crash data in North Carolina from 2010 to 2013 were acquired from the Federal Highway Administration Highway Safety Information System (HSIS) and then split into two-year periods (2010-11 and 2012-13). Likelihood ratio tests confirmed that the two time periods should indeed be modeled separately, thus implying temporal instability. The random parameters logit approach with heterogeneity in mean and variance was selected with three injury severity levels (injury, possible injury, and property damage only). Akaike's and Bayesian information criteria (AIC and BIC) demonstrate that this method outperforms the random parameters logit approach. Contributing factors were found to vary with time, although involvement of alcohol or drugs and full access control have similar outcomes over both periods.

Islam et al. (2020) investigated the severity of work-zone related crashes over a six-year period, from January 1, 2012, to December 31, 2017. Data were retrieved from the Florida Crash Analysis Reporting (CAR) data system and combined with a vehicle dataset, resulting in a comprehensive dataset of single-vehicle work zone crashes. To account for possible unobserved heterogeneity, a random parameters logit model was used. Likelihood-ratio tests rejected the null hypothesis that parameters are equal in all years, so differences in injury severity data by year were deemed statistically significant. Only two variables were found to be statistically significant over all years. The authors noted that variations in work zones, which by nature are temporary, are a source of the temporal instability that has been observed.

Hosseini et al. (2021) developed a Multiple Correspondence Analysis approach to finding significant contributing factors influencing crash severity in New Jersey. Work zone crash data from 2016 to 2018 in New Jersey were utilized for the study. A total of 20 independent variables, categorized into crash, road, temporal, driver, and environmental characteristics were selected, with injury severity as the dependent variable. The results illustrated that the most significant factors were lighting conditions, time, vehicles involved in crashes, and crash type. Subsequently, the authors suggested installing lighting equipment, providing speed limits and enforcement, implementing Variable Speed Limit technology, increasing fines for offending drivers, and providing education as effective countermeasures to reduce the rate and severity of crashes.

Mokhtarimousavi et al. (2021) investigated factors affecting crash severity and its relation to time of day. Crash data from the S4 crash database were obtained for Florida from 2015 to 2017. Separate binary mixed logit models were utilized to determine contributing factors for daytime and nighttime conditions. Furthermore, SVM models trained by the Cuckoo Search (CS) algorithm were used to explore nonlinear relationships among crash severity levels. In both daytime and nighttime models, driver alcohol involvement, rainy weather, wet surfaces, multiple vehicle occupants, and distraction were found to be the most significant contributors to injury severity in work zone crashes. Additionally, the CS-SVM models were found to more accurately predict crashes in comparison to the SVM models, which themselves outperformed the logit models.

Ashqar et al. (2021) investigated the impact of different risk factors on work zone crash severity. Data were pulled from work zone crashes along highway I-94 in Michigan for the 2016 calendar year. Frequency analyses, logistic regression statistics, and a machine learning Random Forest (RF) algorithm were all used to identify and model risk factors. Driver, crash, road, and environmental specifications were considered as independent variables, and crash severity was used as the dependent variable. Based on the results, the authors suggested potential countermeasures to reduce work zone crashes, including traffic calming before the work zone, improving illumination during the work zone, as well as education and awareness measures for high-risk driver groups. For small sample sizes, the RF algorithm was proposed as a more effective approach to crash data analysis when compared to logistic regression.

Islam (2022) proposed a model to identify contributing factors to injury severity in work zone motorcyclist crashes. Data contained Florida work zone crashes involving motorcycles from 2012 to 2016 and were obtained from the CAR system. The resulting dataset contained a variety of potential factors, including motorcycle type, speed, helmet usage, roadway characteristics, crash characteristics, and both spatial and temporal characteristics. The random parameter multinomial logit model with heterogeneity in mean and variance was utilized for both single and multi-vehicle motorcycle crashes. It was found that license endorsement and partial ejection were the only variables statistically significant in both single- and multi-vehicle crashes. Based on the results, the authors suggested lighting at night, shoulder widening, increasing signage on surface conditions, improving helmet usage, and increasing motorcycle endorsement (education-based) as potential countermeasures to decrease the number and severity of crashes of work zone crashes involving motorcycles.

2.1.1. Truck Involvement

Khattak and Targa (2004) investigated how work zone characteristics affect total harm and the most seriously injured occupant in crashes with a distinction between truck-involved and non-truck-involved collisions. Data were obtained from HSIS and combined with police crash reports from the State of North Carolina in 2000. An ordered probit model was used with consideration of the ordinal and categorical nature of injury severity. A new total harm variable was created by assigning an economic value to injury severity and summing all injuries. The ordinary least-square log-transformed model was used for total harm. It was found that truck-involved, multivehicle crashes were most harmful and injurious under a variety of conditions.

Osman et al. (2016) analyzed which causal factors contributed to injury severity of large truck crashes in work zones. Data were collected from 2003 to 2012 in Minnesota from HSIS. A variety of unordered and ordered modeling methods were utilized and compared, with the generalized ordered response logit (GORL) model outperforming the rest based on the Bayesian Information Criterion (BIC) test statistic. The most significant variables increasing the risk of severe outcomes in work zone crashes involving large trucks were found to be daytime crashes, no control of access, higher speed limit, and rural principal arterial road classification. The authors suggested that lowering speed limits and using warning signs to inform motorists or large trucks of work zones can lower occurrences of crashes.

2.1.2. Lighting Condition

Dias and Dissanayake (2016) identified which factors contributed to higher injury severity in work-zone crashes and compared nighttime and daytime crashes in work zones. Data were obtained from KDOT for all work zone crashes within the state from 2010 to 2013. Crash severity was considered as the dependent variable with five categories: fatal, incapacitating injury, non-incapacitating injury, possible injury, and not injured. Ordered probit models, one for daytime and one for nighttime, were produced. Some factors increased or decreased crash severity consistently over both day and night, but other variables, such as work zone area of crash occurrence and driver age, had differing effects at day and night.

Wei et al. (2017) analyzed injury severity in work zone crashes under different lighting conditions. Data from 2003 to 2015 were pulled from the Enhanced Tennessee Roadway Information Management System. Although five unique light conditions were described in this data, they were grouped into three categories: daylight, dark-lighted, and dark-not-lighted. Using the Classification and Regression Trees (CART) algorithm, decision trees for each of the three light conditions were generated to determine contributing factors to work zone crashes and severity. The study found that traffic control devices had differing effects depending on lighting conditions, implying they should be designed differently according to light conditions. Additionally, an increase in the number of lanes may increase crash severity in daylight conditions but have the opposite effect in dark-not-lighted conditions.

Al-Bdairi (2020) investigated the significance of time of day in highway work zone crashes. The time of day is separated into four groups: (1) Morning from 6:00 to 11:00 a.m., (2) Midday from 12:00 to 5:00 p.m., (3) Night from 6:00 to 11:00 p.m., and (4) Late night from 12:00 to 5:00 a.m. using data obtained from the Washington State Department of Transportation (WSDOT). A mixed logit model was used to account for unobserved heterogeneity and predict injury severity. Likelihood ratio tests reject the null hypothesis that estimated parameters are the same across holistic and separated models, so the different periods must be separately modeled. Some factors, such as lack of airbag deployment and rear-end collision, affect injury severity regardless of time of day (albeit their impact varies with time of day). Other factors are common in multiple periods, such as female drivers decreased the probability of no injury during morning and night. Some factors are significant in only one period, such as sober drivers increasing the probability of no injury during the morning.

2.1.3. Work Zone Type

Weng and Meng (2011) analyzed casualty risk for drivers in work zone crashes for different work zone types. The Fatality Analysis Reporting System (FARS) was used to obtain data on work zones within the United States between 2001 and 2006. The binary logistic regression model was used to predict either injury or non-injury; fatalities were not considered separately due to their relatively small contribution to the total number of crashes and thus grouped within the injury category. Models were created for each of the three work zone types: construction, maintenance, and utility. It was found that construction zones have the largest casualty risk and that several factors influence risk in all three types. The authors noted that the relatively small sample sizes for maintenance and utility work zones may affect these results.

Osman et al. (2018) investigated which factors contributed to injury severity in passenger-car crashes with specific attention to work zone configuration. The dataset used was collected from HSIS for work zone crashes in Minnesota from 2003 to 2012. Work zones were categorized into five types: lane closure, lane shift, crossover, shoulder or median, and intermittent/mobile. A mixed generalized ordered response probit model was utilized, which allows additional flexibility over the standard ORP model. Some variables were found to be significant across all work zone types, while others were type-specific. The author recommended collecting work zone-specific data such as duration, lane widths, and speed limits to improve findings.

Yu et al. (2020) investigated the factors affecting work zone crashes involving trucks in rural and urban areas. Data on truck-involved work zone crashes were obtained from HSIS from 2005 to 2014 in North Carolina. To account for unobserved heterogeneity, the mixed logit and partial proportional odds models are utilized and compared. Three injury severity levels (fatal/incapacitating/non-incapacitating injury, possible injury, and property damage only) were considered for the dependent variable. Conclusions include that usage of restraint and involvement of alcohol are contributing factors regardless of area. Visibility improvement and speed reductions would be effective in rural areas, while appropriate placement and design of indicator signs would be effective in urban areas.

2.1.4. Work Zone Area

Osman et al. (2019) investigated the different risk factors affecting driver injury severity in work zone crashes within different work zone areas. Work zone crash data from 2002 to 2013 in Minnesota were acquired from HSIS. Work zones are broken into four distinct areas: advance warning area, transition area, activity area, and termination area. Injury severity is classified into three categories: severe injury, injury, and no injury. A mixed generalized ordered response probit model is adopted, which accounts for unobserved heterogeneity and ordering of injury severity. Airbag deployment, alcohol involvement, ejection, seatbelt use, and partial control of access are all found to contribute to severe outcomes, and many covariates had varying effects across different work zone areas.

Koilada et al. (2020) examined how the odds of crash occurrence and its contributing factors change with work zone area. Work zones were split into four area types per the Manual on Uniform Traffic Control Devices (MUTCD): advance warning, transition, activity, and termination. Five years (2010-2014) of crash data from North Carolina were acquired from the Highway Safety

Information Systems. Four models were developed: proportional odds for injury severity in the transition area, and partial proportional odds models for work zone area types, injury severity in the advance warning area, and injury severity in the transition area. The results indicated specific variables affected the odds of crashes depending on the work zone area. Another notable result is that the odds of a crash increased in transition and activity areas when flexible post barriers were used as a median.

2.2. Crash Frequency

The following includes studies that investigated contributing factors to crash frequency in work zones. The sub-sections (2.2.1 and 2.2.2) group papers focused on work zone crash frequency as it relates to rear-end crashes and traffic control devices respectively.

Daniel et al. (2000) studied fatal work-zone crashes and compared them to fatal crashes at non-work-zone locations. Data were obtained from FARS and from the Georgia Department of Transportation for work zones in Georgia from 1995 to 1997. Different factors, such as manner of collision, light conditions, truck involvement, and roadway classification were examined and summarized. Statistical tests for independence were performed for each of the above factors, and the null hypothesis of independence between work-zone and non-work-zone crashes was rejected. Several findings were identified: crashes more often occurred in construction work zones rather than maintenance work zones, vehicles were more likely to be involved in fatal work zone crashes than fatal non-work-zone crashes, and rear-end crashes were a high proportion of work zone crashes.

Garber and Zhao (2002) studied the characteristics of crashes within different work zone areas. Data were acquired using Virginia police crash reports from 1996 through 1999. Crashes in work zones were classified as occurring in one of five areas: advance warning, transition, buffer, activity, and termination. Proportionality tests were performed to determine whether crash frequency, severity level, crashes by severity, and collision type had statistically significant differences in each work zone area. It was found that the activity area was the most prevalent accident location, property damage only was the common severity, and rear-end crashes were the most predominant collision type.

Arditi et al. (2007) investigated fatal work zone crashes to find a potential safety difference between daytime and nighttime safety. Data were collected from FARS and filtered into crashes occurring at work zones in Illinois from 1996 to 2000. The data were sorted into daytime based on the “Day” classification, and all other classifications were considered nighttime, which included dawn and dusk and was irrespective of light conditions. Calibration factors for traffic volume, number of work zones, and hours of daylight vs nighttime were combined to compensate for the difference between day and night, thus making the dataset more directly comparable. Using the Kruskal-Wallis test, the null hypothesis of no differences between daytime and nighttime fatal crash occurrences was rejected. As such, statistical evidence pointed to nighttime construction being five times more hazardous than daytime in regard to fatal work zone crashes.

Yang et al. (2013) investigated the relationship between explanatory variables and work zone crash frequency. Data for work zone crashes in New Jersey were assembled from independent sources within the New Jersey Department of Transportation. The negative binomial model was extended to incorporate effects arising from errors in the measurement of work zone length. This error is due to variability in work zone length as the project progresses. A new model coined MENB, accounted for this to better fit the data. It was found that work zone length and traffic volume were positively correlated with work zone crash occurrence.

Weng et al. (2015) investigated drivers' merging behavior as well as rear-end crash risk in work zone merging areas. The period considered begins with the start of the merging maneuver and ends with the vehicle fully entering the adjacent through lane. For calibration and validation, a case study with merging trajectory data from a work zone in Singapore was used. A mixed probit-based merging behavior model was developed to determine the probability a merging vehicle completes the merging maneuver. Two surrogate safety measures were selected to compute the rear-end crash risk between the merging vehicle and its neighboring vehicles. It was found that rear-end crashes were more likely to occur when the merging vehicle moves very slowly or quickly, and the probability of completing the merging maneuver increases over the time elapsed.

Weng et al. (2016) investigated casualty patterns in work zone crashes using association rules . The association rule approach is a data mining technique that can interpret relationships between a large number of variables; these relationships can then be easily described. By changing the support and confidence, different lift values, determining association strength, were found. A case study was performed with data including environmental characteristics, control information, crash information, and occupant information. Data were acquired from the University of Michigan Transportation Institute. This case study concluded that crashes were more likely on roads with more than four lanes where the speed limit was more than 40 miles per hour.

Weng et al. (2018) developed a time-varying mixed logit model for modeling vehicle merging behavior within work zones . The period considered begins with the start of the merging maneuver and ends with the vehicle fully entering the adjacent through lane. For calibration and validation, a case study with merging trajectory data from a work zone in Singapore was used. This new method was found to have a higher prediction accuracy than models utilizing vehicle speeds and gap sizes. Several factors were found to affect merging behavior, and specific scenarios in which vehicles were more likely to successfully complete the merging maneuver were defined.

Hou and Chen (2020) presented an integrated framework for work zone safety under adverse driving conditions . The framework is broken into four parts. In the first, work zone traffic is simulated using a cellular automaton model considering a variety of factors. In the second and third, multiple-vehicle and single-vehicle crash simulations respectively are used to determine the probability of each. In the fourth, overall safety is assessed using both single-vehicle and multiple-vehicle crashes. A case study was used to investigate safety under different weather conditions. The results of their study indicated rain and snow conditions lower work zone capacity and increase the probability of crashes, and the most prevalent type of crash varies with weather conditions. Limitations of their study include a lack of validation with actual crash data and not considering the possibility of multi-vehicle crashes caused by single-vehicle crashes.

Gupta et al. (2021) investigated work zone crashes resulting in fatalities and involving trucks . Crash data was pulled from the S4 database developed by the Florida Department of Highway Safety and Motor Vehicles over seven years. Data resampling was accomplished using the SMOTE-NC algorithm and random over-sampling, and then significant variables were extracted from decision trees to create tuned RF models. For truck crashes, pedestrian involvement, lighting conditions, safety equipment, driver condition, driver age, and work zone location were all identified as primary contributors. Some environmental and roadway-specific conditions notably did not show significant contribution to the model. Fatality patterns for pedestrian crashes showed different factors contributed when compared to non-pedestrian fatal crashes.

Santos et al. (2021) identified risk factors affecting work zone crashes and compared binary logistic (logit) to probit regression modeling methods . Data were collected from police crash reports in mainland Portugal from 2013 to 2015. Due to limitations related to the filling of data into crash report forms, the authors decided that performing modeling by road environment would compensate for missing information. Analysis was performed to determine risk factors for crash type, primary contributing factor, and driver age group. Logit and probit models were found to produce very similar results. The authors suggested that their differences may be due to the small sample size.

2.2.1. Rear-End Crashes

Meng and Weng (2011) evaluated the rear-end crash risk at work zones and investigated driver merging behavior . Data were obtained from two work zones in Singapore using a video camera to record vehicle trajectory. The primary measure to determine crash probability was the deceleration rate to avoid the crash (DRAC). Using the stepwise regression method, four models were produced: one for each of the two work zones data were collected by considering macroscopic contributing factors, one which combined the two work zones, and one which additionally considered twelve microscopic variables. Findings included the lane closest to the work zone, the expressway work zone, and trucks having the highest crash risk out of their respective categories.

Weng et al. (2015) evaluated the effect of vehicle-following patterns on rear-end crash risk in work zones . Four front vehicle-following vehicle patterns were used: car-car, car-truck, truck-car, and truck-truck. Data were collected using video cameras at two expressway work zone sites in Singapore. DRAC was the primary measure to determine rear-end crash risk, and tests for statistical significance proved each of the four patterns should be separately modeled. The highest risk for rear-end crashes in work zones was found to be the car-truck pattern.

2.2.2. Traffic Control Devices

Li and Bai (2009) investigated the effectiveness of different temporary traffic control devices in reducing severe work zone crashes . Data represented work zone crashes resulting in fatality or injury in Kansas occurring in 2003 or 2004. The binary logistic regression technique was used for evaluating effectiveness, which was measured by severity reduction and odds of crash occurrence. This was compared against the following human errors: inattentive driving, disregarding traffic control, following too closely, and exceeding speed limit/driving too fast for conditions. The most effective devices were found to be the presence of a flagger or officer, followed by having flashers or center/edge lines. Stop signs/signals and no passing zones were not found to be effective.

Rista et al. (2017) examined the impact of various temporary traffic control measures on work zone safety . Data included lane closure reports, AADT estimates, as well as traffic crash information and was provided by the Michigan Department of Transportation and state police crash database. Safety performance functions including site-specific information were developed and reinforced by count data models to account for unobserved heterogeneity. Sites were primarily compared to their respective locations before the implementation of work zones. Results found that there was no difference in crash rates in shoulder closure work zones when compared to pre-work-zone conditions, while other work zone types (single and multilane closures, lane shifts) showed higher crash rates.

Department of Transportation & Infrastructure Studies, Morgan State University, Baltimore, MD, USA et al. (2018) studied the impact of mobile barriers on driver behavior on arterial roads A driving simulator was utilized to replicate a one-mile stretch of Hillen Road, located in Baltimore, Maryland. Test drivers' throttle/brake control (speeding) behavior and steering handling (lateral movement) behavior were recorded along with pre- and post-simulation surveys to collect demographics and preferences on barrier type, respectively. Three types of work zone barriers were investigated: cone pylons, concrete jersey barriers, and metal barriers. Results found that participants drove faster next to concrete barriers than cone pylons but tended to move away from concrete barriers. The authors suggested that concrete jersey barriers may be more effective in improving safety on arterial roads.

2.3. Combined Injury Severity and Crash Frequency

Dias (2015) analyzed work zone crash characteristics and identified factors associated with crash severity and frequency data were acquired for the entire state of Kansas from 2010 to 2013 using a variety of databases, including the Kansas CAR and KANPLAN, a GIS portal for KDOT. A common crash severity model was developed alongside individual models for crash severity for daytime, nighttime, single-vehicle, and multi-vehicle work zone crashes. All injury severity models utilized ordered probit. Crash frequency negative binomial models were also developed to find crash characteristics related to crash frequencies. The author recommended that appropriate countermeasures be implemented based on the contributing factors leading to increased crash severity. He noted that the unavailability of a full work zone database to find proper information was a major difficulty in the study.

Khattak et al. (2002) analyzed the effect of work zone duration on the frequency and severity of crashes in California. Their analysis used data obtained from HSIS and project-level information obtained from CALTRANS. Using calculated crash rates before the work zone and during the work zone, the authors developed five negative binomial models. Crash frequency, non-injury crashes pre-work zone, non-injury crashes during work zone, injury crashes pre-work zone, and injury crashes during work zone were used as dependent variables for each of the models. The factors considered include average daily traffic (ADT), work zone duration, work zone length, traffic exposure, and urban setting of the work zone. The first model predicted that the total crash rate would increase by 21.5% during the work zone period compared to the pre-work zone period and that the non-injury crash rate has a larger increase than the injury crash rate. The other four models indicated that increased work zone length, work zone duration, and traffic exposure raise the frequency of non-injury and injury crashes in the work zone.

2.4. Analysis of Factors that Contribute to Work Zone Crashes

This section summarizes studies that investigated contributing factors in work zone crashes but do not specifically focus on injury severity or crash frequency. The subsection (2.4.1) focuses on papers that focused specifically on the misclassification of data in work zone crashes.

Debnath et al. (2015) performed a qualitative study of worker perceptions of common work zone hazards and countermeasures . Participants for the study were recruited from private and government organizations involved in road construction, maintenance, and traffic control in Queensland, Australia, and averaged over nine years of roadwork-related experience. Responses were split into three groups based on exposure to traffic: fully exposed, semi-exposed (usually behind some barrier or protection), and non-exposed (those working primarily from officers with only occasional visits to sites). The most frequent hazard mentioned was drivers exceeding work zone speed limits. Other significant factors included driver inattention and adverse environmental conditions. Respondents additionally mentioned driver aggression as a hazard, which is rarely investigated in studies. Police reinforcement and education measures were the most common suggested means of improving safety in work zones.

Yang et al. (2015) reviewed work zone modeling and safety-related analysis. Generally consistent results confirmed that work zones increase crash rates, crashes are not uniformly distributed across work zones, and rear-end crashes are the most common type of crash. The majority of studies utilized negative binomial and logistic regression models based on police crash report data, but the authors argue that these models cannot accurately incorporate work zone-specific factors, such as their inherent short-term nature and common lack of sufficient crash data at any given work zone. The authors recommended more advanced statistical modeling methods once more comprehensive data can be collected about work zone crashes.

Theofilatos et al. (2017) attempted to summarize the effect of work zones on road safety and crash frequency from other studies. Studies related to work zone crashes focusing on either length and/or duration which applied fixed effects negative binomial models were selected for this study. Meta-analysis and meta-regression techniques were utilized to provide a general estimate of coefficients for work zone duration and work zone length. After correction, duration was found to have a positive non-significant effect on work zone crash frequency, while length had a positive significant effect. The author noted the rather small selection of studies matching the criteria and the fundamental heterogeneity across different studies as limitations when applying the results.

Mannering (2018) explored the temporal instability of highway accident data and discussed its possible implications. The author drew from several fields, including cognitive science, economics, neuroscience, and psychology to conclude that decision-making, which applies to all drivers, is temporally unstable. He suggested that temporal instability is potentially a significant portion of unobserved heterogeneity in models and that many of the existing accident prediction methods are unable to account for this factor. Some potential methods of compensating for said instability are presented, such as developing a function that predicts how variables will change over time. Although appropriately modeling temporal instability remains a significant challenge, the author recommended further investigation and increasing awareness of this factor in current safety assessment practices.

Thapa and Mishra (2021) investigated the influence of external variables on the performance of Work Zone Intrusion Alert Systems (WZIAS) . Three different types of WZIAS were used over 525 trials to determine factors resulting in work zone crashes. The subsequent data were analyzed using survival analysis, including the non-parametric Kaplan Meir estimator and the semi-parametric Cox proportional hazard model. Intrusion speed, sensor-to-worker spacing, and system accuracy were all found to significantly influence crash occurrence. In addition to suggesting standardized deployment strategies for different WZIAS types, the authors recommended reducing speed limits and standardizing the length of the buffer space as work zone crash countermeasures.

Azimi et al. (2021) created a guideline to assist decision-makers in analyzing whether and what type of ITS technologies should be used in work zone projects through a four-step system . Based on interviews with practicing engineers and contractors involved in work zone projects, a scoring system and flowchart were developed to assess the feasibility of ITSs given work zone conditions. ITS candidates for the work zone were identified based on the intent of the ITS device and project characteristics. The ITS to be used was selected based on its potential benefits and associated costs. Last, the ITSs were deployed and evaluated for performance. A Texas Department of Transportation (TxDOT) highway improvement project was used as a case study for recommending potential ITSs to deploy.

2.4.1. Misclassification of Data in Work Zone Crashes

Ullman and Scriba (2004) presented how differences in state crash report forms can influence work zone crash data in FARS. Researchers reviewed and categorized crash report forms for each of the states based on whether work zone fields were included explicitly, indirectly, or not at all. These forms were then compared to their respective 1992 counterparts to determine which had added fields for work zones. Based on three years of crash data for each state (1998-2000), it was found that there is a statistically significant, linear relationship between the percentage of fatalities recorded in work zones and the way work zone data is included on crash report forms. Based on this analysis, the authors suggested that existing data may underreport work zone fatalities by up to 10%.

Yahaya et al. (2020) studied the effects of mislabeling in crash datasets using machine learning algorithms to identify misclassified information . A work zone crash injury severity dataset from Cairo, Egypt from 2010 to 2015 was acquired from Egypt's GARBLT and the Police Reports Accident Database. A new M-IPF algorithm, based on the Iterative Partitioning Filter (IPF), was proposed, which included additional sampling techniques. The additions were intended to minimize the incorrect deletion of minority class samples. M-IPF filter was compared to other state-of-the-art filtering algorithms for effectiveness and efficiency and was found to have superior performance; albeit the authors noted that the method is too likely to eliminate samples that are not mislabeled.

Sayed et al. (2021) developed a classifier to find unidentified work zone crashes in crash reports through text mining . Wisconsin crash reports from January 2017 to January 2018 were used as training data, and January to October 2019 data were used for testing. The classifier utilized the noisy-OR method to determine unigram and bigram work-zone-indicative “keywords” in the crash report narrative. In the top 450 cases identified, 201 were identified as missed work zone crashes, proving the unigram + bigram noisy-OR method classifier was effective at classifying missing work zone crashes. The authors utilized ad-hoc analysis of misclassified work zone crashes to find the reason for work zone missing crashes. It was found that work zone crashes were most often missed during the daytime (specifically during the 4-5 p.m. period), during summer months, and on urban city streets.

3. Methodology

3.1. Data Acquisition

Crash data were provided by the SCDOT and came in three separate CSV files: crash, unit, and occupant. These tables were joined using the field "ANO," the primary key/unique identifier for each collision as illustrated in Figure 2.

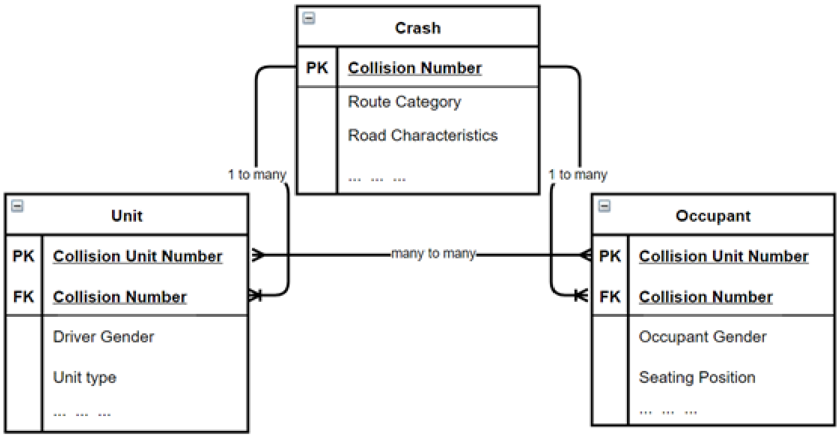


Figure 2. Merge procedure of SCDOT crash data tables using unique keys.

The available crash data contains information on time and day (such as day of the week and time of day), roadway and environmental conditions (such as functional classification, curve, and grade), crash attributes (such as number of vehicles involved, collision speed), work zones (such as configuration type), vehicles (such as airbag deployment), and drivers characteristics (such as age, gender). Note that these data pertain to the conditions observed by the reporting officer. For example, collision speed is not the actual or measured speed, but rather, an estimated speed. It is obtained based on answers provided by the drivers involved and based on the evidence at the crash scene (e.g., length of skid marks, deployment of airbags, extent of damages to vehicles).

Several data sets were developed for the different tasks associated with this project. The development processes for these sets are described below according to the tasks each set was used for. In this study, the injury severity used for a crash is the most severe one; there could be more than one injury in a crash.

3.1.1. Data for Identifying Contributing Factors

As previously noted, two separate analyses were conducted to identify the contributing factors, each utilizing a different dataset. The first analysis sought to identify contributing factors to injury for work zone-related crashes on roadways with 60 mph or higher speed limits (assumed to be predominantly interstates) and on roadways with less than 60 mph (assumed to be predominantly non-interstate). Mixed logit models were developed using South Carolina statewide work zone crash data from 2014 to 2020. This analysis focused on truck-involved crashes. The reason is that truck-involved crashes at work zones pose a greater risk for injuries and fatalities, and they are more serious in nature than crashes that occur in non-work zones Khattak and Targa (2004). Additionally, truck-involved crashes pose a greater economic impact as trucks carry high-value goods and require a longer incident clearance time. Despite initiatives implemented by SCDOT to improve work zone safety (e.g., National Work Zone Awareness Week), the number of truck-involved crashes at work zones increased from 189 in 2014 to 666 in 2019 (a 252.4% increase) with the peak occurring in 2018. The increasing trend from 2014 to 2019 is a concern, given that the number of work zones is expected to increase significantly due to the increase in funding for construction projects. It should be noted that due to the COVID-19 pandemic, the total number of crashes and truck-involved crashes at work zones in South Carolina decreased in 2020.

In total, there were 15,727 crashes: 93 were fatal crashes, 176 were serious injury crashes, 674 were minor injury crashes, 2,451 were possible injuries, and 12,333 were property damage only (PDO) crashes. To prepare the dataset for modeling, it was filtered to include only those crashes that involved at least one truck whether at fault or not. In the truck-involved work zone crash dataset, there were a total of 3064 crashes: 29 (0.95%) were fatal crashes, 36 (1.75%) were serious injury crashes, 121 (3.95%) were minor injury crashes, 381 (12.43%) were possible injury crashes, and 2,496 (81.46%) were PDO crashes. Due to the small number of observations per injury severity level, the five levels are combined into two, injury and PDO, where injury includes fatal, serious injury, minor injury, and possible injury. The final dataset contains 565 injury crashes and 2,488 PDO crashes. It should be noted that the final dataset has 11 fewer observations [3064 - (565 + 2488)] than the initial dataset because these crash records did not have posted speed limits. To evaluate the effect of the posted speed limit on the roadway where the work zone is located, the final dataset was first divided into three different speed limit categories: less than 40 mph, between 40 and 60 mph, and 60 mph or greater. The number of observations for the less than 40 mph category was only 312. For this reason, two speed limit categories were used, less than 60 mph and greater than or equal to 60 mph. The former category has 305 injury crashes and 1,443 PDO crashes, whereas the latter category has 260 injury crashes and 1,045 PDO crashes. The reason for choosing 60 mph as the demarcation speed is that interstates in South Carolina typically have a posted limit of 60 mph or higher and guidelines for setting up work zones on interstates are more stringent. Table 1 presents the injury severity level frequency and percentage distribution by speed limit categories.

Table 1. Injury severity level frequency and percentage distribution by posted speed limit levels.

Speed category	Total observation	Injury (%)	PDO (%)
Less than 60 mph	1,748	305 (17.45)	1,443 (82.55)
Greater than or equal to 60 mph	1,305	260 (19.92)	1,045 (80.08)

Descriptive statistics of explanatory variables used in this analysis are shown in Table 2. They include characteristics related to vehicles, crashes, roadways, work zones, day and time, environment, and drivers.

Table 2. Descriptive statistics of variables by two posted speed limit levels.

Variables	Speed < 60 mph	Speed ≥ 60 mph
	Percent (%)	Percent (%)
Driver Characteristics		
Gender (1 if female driver are at fault in a crash, 0 otherwise)	20.41	14.60
Younger drivers (1 if age of at-fault driver are group below 35 years, 0 otherwise)	30.82	31.22
Middle-aged drivers (1 if age of at-fault driver are between 35 and 50 years, 0 otherwise)	27.06	30.26
Older drivers (1 if age of at-fault driver are group above 50 years, 0 otherwise)	41.97	38.47
Driving too fast (1 if the contributing factor of crash is driving too fast, 0 otherwise)	24.75	43.31
Distracted (1 if the contributing factor of crash is distracted, 0 otherwise)	8.83	0.80
Failed (1 if the contributing factor of crash is failed to yield right of way, 0 otherwise)	12.45	4.21
Improper usage (1 if the contributing factor of crash is improper lane usage, 0 otherwise)	12.88	37.99
Under influence (1 if the contributing factor of crash is under the influence, 0 otherwise)	2.32	2.02
Crash Characteristics		
1 vehicle (1 if the number of vehicles involved in a crash is 1 or more, 0 otherwise)	6.80	6.13
2-vehicles (1 if the number of vehicles involved in a crash is 2, 0 otherwise)	84.66	78.48
3+ vehicles (1 if the number of vehicles involved in a crash is 3 or more, 0 otherwise)	8.39	15.34
Rear End (1 if manner of collision is rear end, 0 otherwise)	32.13	39.80
Sideswipe (1 if manner of collision is sideswipe, 0 otherwise)	24.02	35.59
Crash Characteristics		
Angle (1 if manner of collision is Angle, 0 otherwise)	21.85	10.18
Fixed object (1 if 1st harmful event is fixed object, 0 otherwise)	5.93	8.10
Not fixed object (1 if 1st harmful event is Not fixed object, 0 otherwise)	92.19	90.09
No collision (1 if 1st harmful event is no collision, 0 otherwise)	1.74	1.76
Roadway Characteristics		

Variables	Speed < 60 mph	Speed ≥ 60 mph
	Percent (%)	Percent (%)
SC, US Primary (1 if crash occurred in SC or US Primary, 0 otherwise)	58.60	0.61
Interstate (1 if crash occurred in interstate, 0 otherwise)	3.01	98.01
County/Secondary/Ramp (1 if crash occurred in County, Secondary or Ramp, 0 otherwise)	38.49	1.45
Straight on grade (1 if crash occurred in a straight on grade, 0 otherwise)	7.96	12.79
Straight level (1 if crash occurred in a straight level, 0 otherwise)	84.66	83.22
Roadway (1 if first harmful event occurred on roadway, 0 otherwise)	90.88	89.24
Two-way divided (1 if traffic-way is two-way undivided, 0 otherwise)	33.00	98.08
Environmental Characteristics		
Dark (1 if crash occurred in a dark lighting condition, 0 otherwise)	11.87	31.43
Dawn or Dusk (1 if crash occurred in a dawn or dusk lighting condition, 0 otherwise)	1.30	3.68
Daylight (1 if crash occurred in a daylight lighting condition, 0 otherwise)	86.69	64.84
Clear (1 if crash occurred in a clear weather condition, 0 otherwise)	90.45	84.50
Dry (1 if crash occurred in a dry surface condition, 0 otherwise)	93.63	88.07
Work Zone Characteristics		
Shoulder/Median (1 if work zone type is Shoulder or Median, 0 otherwise)	30.97	51.04
Lane closure (1 if work zone type is Lane Closure, 0 otherwise)	36.90	31.22
Lane shift/crossover (1 if work zone type is lane shift or crossover, 0 otherwise)	7.67	8.95
Work Zone Characteristics		
Activity area (1 if crash location is in the work zone activity area, 0 otherwise)	69.03	66.38
Before first sign (1 if crash location is before the first sign, 0 otherwise)	2.75	4.26
Advanced warning (1 if crash location is in the work zone advanced warning area, 0 otherwise)	11.43	9.54
Workers present (1 if workers present, 0 otherwise)	72.07	50.93
Temporal Characteristics		
Weekday (1 if crash happens on weekday, 0 otherwise)	94.07	87.37

For the second analysis, the dataset was filtered to include only rear-end collisions with collision speeds greater than or equal to 35 mph. It was suspected that rear-end crashes with higher collision speeds would likely increase the risk of injury. This hypothesis led the project team to the work of Jurewicz et al. (2016). who analyzed the relationship between collision speed and the probability of fatal and serious injuries in rear-end crashes for a range of common crash scenarios. They found rear-end collision speeds of 55 km/h (~35 mph) are more likely to produce an injury probability of approximately 10%, considered a critical threshold in Safe Systems or Vision Zero. When we plotted the cumulative distribution function for rear-end crashes (see Figure 3), we found that 10% of the fatal and serious injury crashes have collision speeds less than 32.20 mph. For these reasons, 35 mph was chosen as the collision speed threshold. No other speed threshold was considered because the lone study found on this topic Jurewicz et al. (2016), and our data suggest 35 mph to be the most appropriate value. As mentioned, the collision speeds are approximated by the investigating officers through drivers' testimonies and evidence gathered from the crash site, such as skid marks, airbag deployment, and the extent of damage sustained by the vehicles. To ensure adequate data representation across various injury severity levels, the five distinct injury levels were consolidated into two: injuries sustained and property damage only (PDO). The resultant dataset comprised 3,648 collisions, among which 1,144 led to injuries, while 2,504 resulted solely in PDO. Table 3 shows the descriptive statistics for all variables in this dataset.

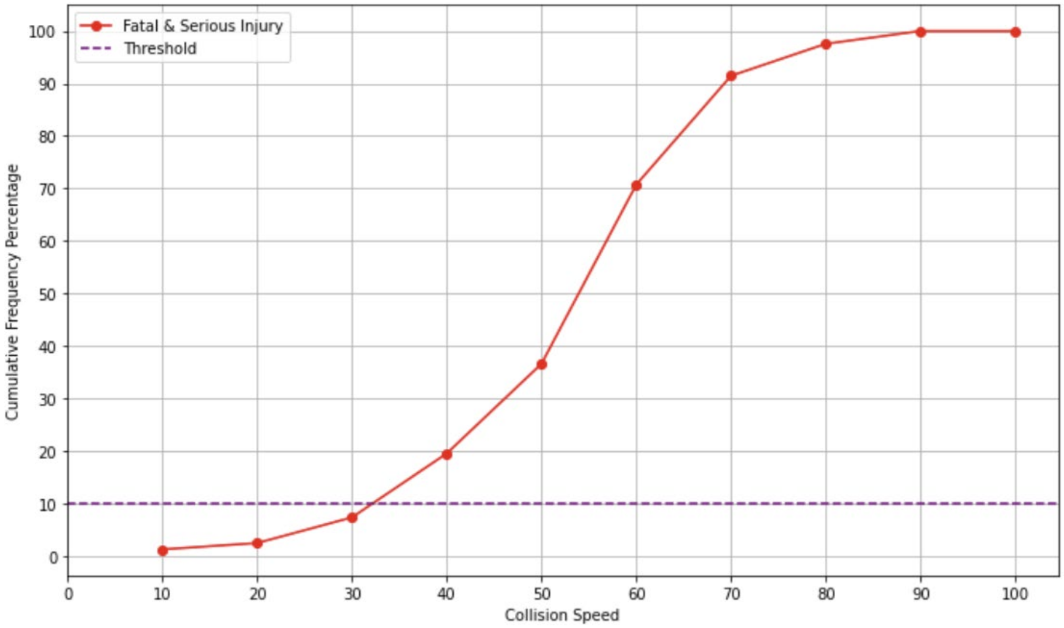


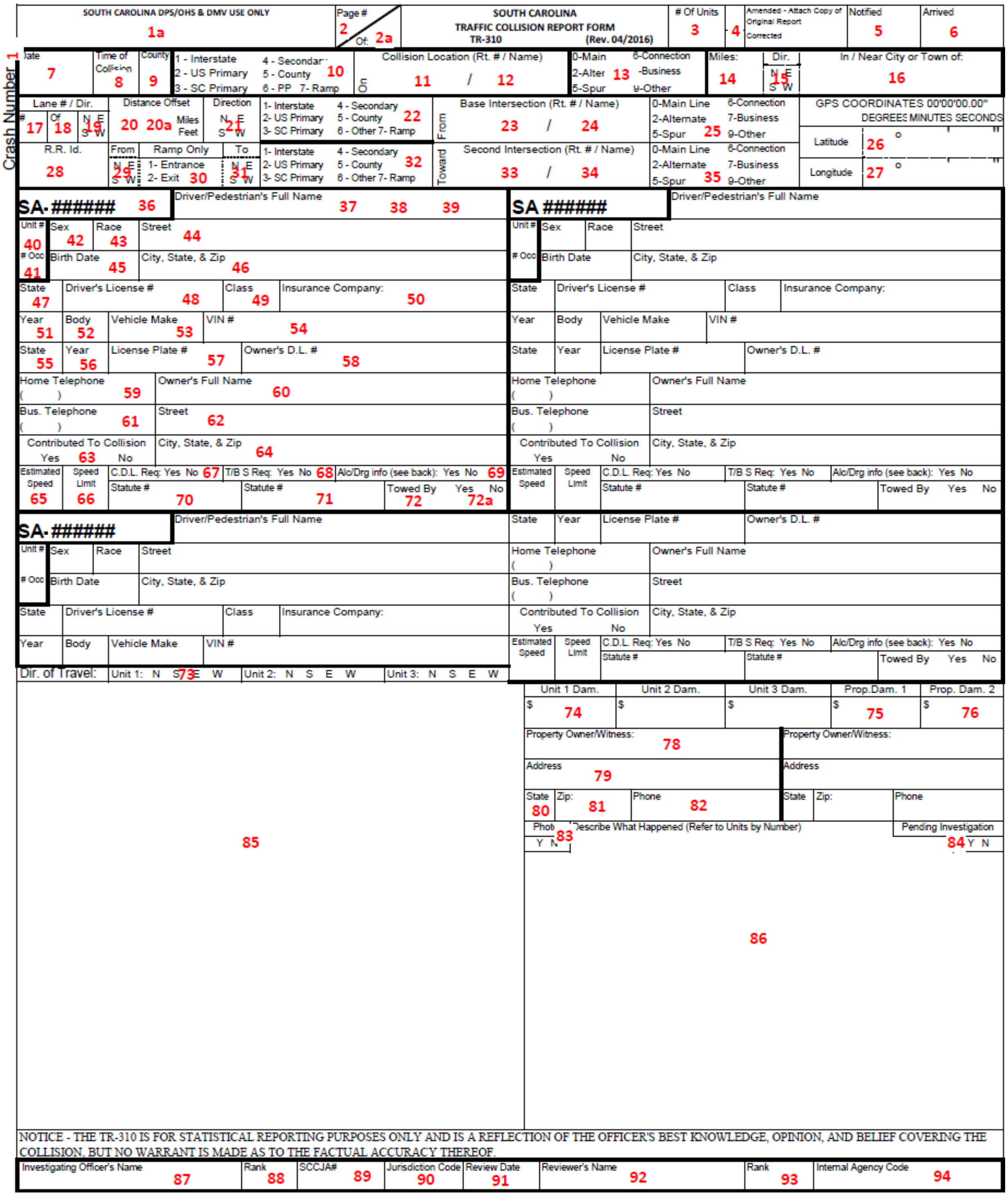
Figure 3. Cumulative frequency percentage vs. collision speeds for rear end crashes.

Table 3. Descriptive statistics of variables for Rear-end crashes with high collision speed.

Variables	**Percent (%)** (when variable = 1)
Driver Characteristics	
Gender (1 if at-fault driver in a crash is female, 0 otherwise)	34.34
Younger drivers (1 if age of at-fault driver is below 35, 0 otherwise)	55.93
Middle-aged drivers (1 if age of at-fault driver is between 35 and 50 years, 0 otherwise)	20.96
Older drivers (1 if age of at-fault driver is above 50, 0 otherwise)	23.16
Driving too fast (1 if marked as contributing factor by investigation officer, 0 otherwise)	84.02
Distracted (1 if marked as contributing factor by investigation officer, 0 otherwise)	2.19
Failed to yield right of way (1 if marked as contributing factor by investigation officer, 0 otherwise)	0.66
Under the influence (1 if marked as contributing factor by investigation officer, 0 otherwise)	3.73
Crash Characteristics	
2-vehicles (1 if the number of vehicles involved in a crash is 2, 0 otherwise)	71.72
3+ vehicles (1 if the number of vehicles involved in a crash is 3 or more, 0 otherwise)	28.31
Truck involved (1 if a truck is involved in the crash, 0 otherwise)	16.88
Vehicle Characteristics	
Airbag (1 if airbag is deployed, 0 otherwise)	26.50
Roadway Characteristics	
Interstate (1 if crash occurred on an interstate, 0 otherwise)	71.47
Curve - level (1 if crash occurred on a horizontal curve with level grade, 0 otherwise)	1.92
Straight - on grade (1 if crash occurred on a straight section on a grade, 0 otherwise)	11.37
Straight - level (1 if crash occurred on a straight section on level grade, 0 otherwise)	85.26
Roadway (1 if first harmful event occurred on roadway, 0 otherwise)	98.44
Two-way divided (1 if roadway is divided, 0 otherwise)	80.27
Environmental Characteristics	
Dark (1 if crash occurred in dark lighting condition, 0 otherwise)	22.20
Dawn or Dusk (1 if crash occurred in dawn or dusk lighting condition, 0 otherwise)	3.26
Daylight (1 if crash occurred in daylight lighting condition, 0 otherwise)	74.60
Clear (1 if crash occurred in a clear weather condition, 0 otherwise)	86.49
Posted speed limit (1 if posted speed limit is above 60, 0 otherwise)	45.79

3.1.2. Data for Crash Report Narrative Discrepancies

Traffic collision forms (TR-310 forms) of fatal crashes occurring within work zones from 2014 to 2020 were provided by the SCDOT in PDF format as shown in Figures 4 and 5. Fields containing personal information were removed from the reports by the SCDOT. The information in the collision forms has been digitized by the SCDPS, 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. The reason for not reviewing all 300 forms is that the process of analyzing the information on the form and documenting the discrepancies can take up to two hours for each form. A sample size of 200 or two-thirds of the population is often considered sufficient.



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 | 4 Amended - Attach Copy of Original Report Corrected | Notified 5 | Arrived 6

Crash Number 1

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 ° ' " Longitude 27 ° ' "

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

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 | Estimated Speed | Speed Limit | C.D.L. Req: Yes No | T/B S Req: Yes No | Alc/Drg info (see back): Yes No

Statute # 70 | Statute # 71 | Towed By 72 Yes No 72a | 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

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

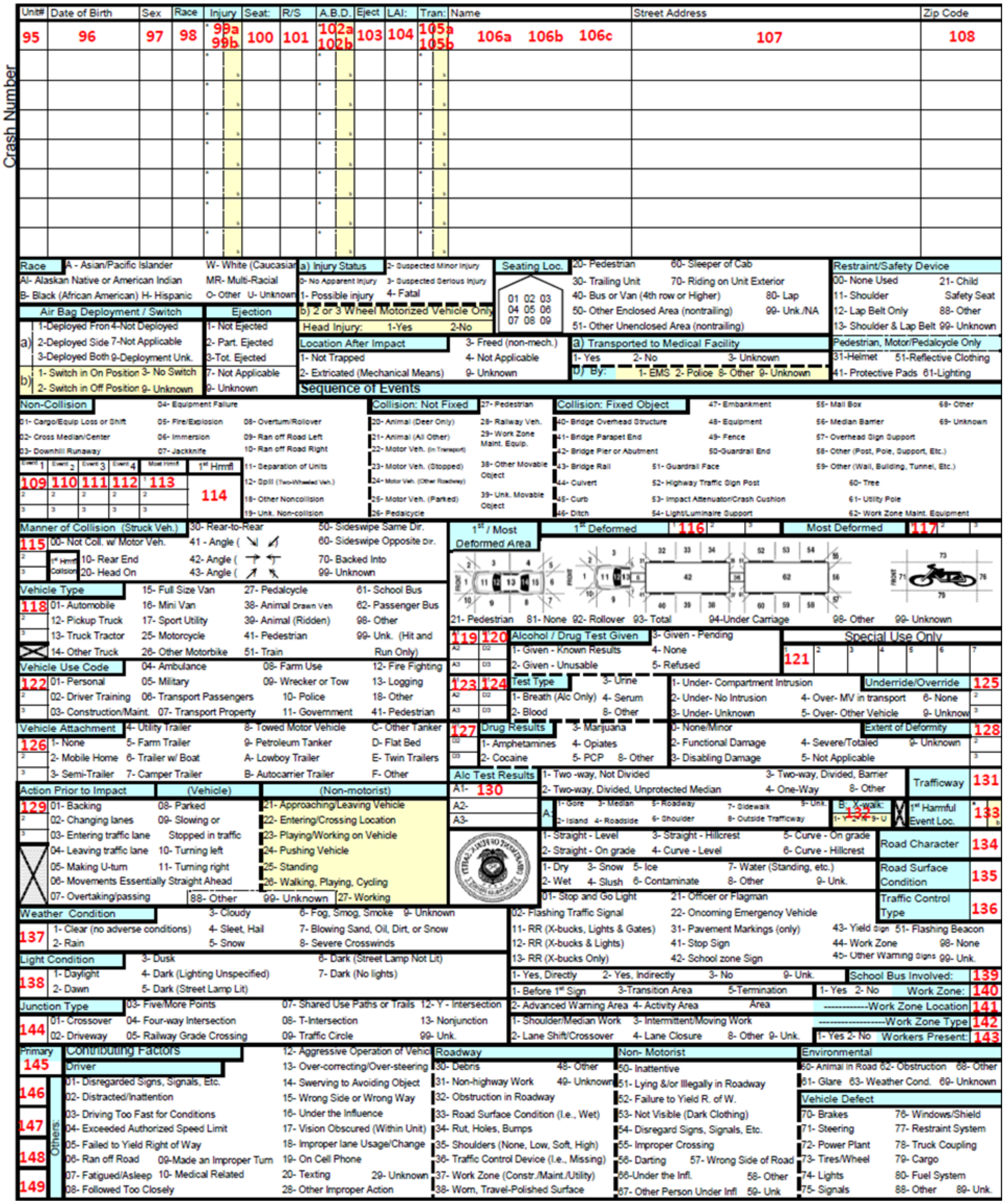
85

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 4. South Carolina traffic collision form (TR-310), front side.



Crash Number

Unit#	Date of Birth	Sex	Race	Injury	Seat:	R/S	A.B.D.	Eject	LAI:	Tran:	Name	Street Address	Zip Code
95	96	97	98	99a 99b	100	101	102a 102b	103	104	105a 105b	106a 106b 106c	107	108

Race
A - Asian/Pacific Islander
AI- Alaskan Native or American Indian
B- Black (African American)
H- Hispanic
W- White (Caucasian
MR- Multi-Racial
O- Other
U- Unknown

a) Injury Status
0- No Apparent Injury
1- Possible injury
2- Suspected Minor Injury
3- Suspected Serious Injury
4- Fatal

b) 2 or 3 Wheel Motorized Vehicle Only
Head Injury: 1-Yes 2-No

Seating Loc.
01 02 03
04 05 06
07 08 09
20- Pedestrian
30- Trailing Unit
40- Bus or Van (4th row or Higher)
50- Other Enclosed Area (nontrailing)
51- Other Unenclosed Area (nontrailing)
60- Sleeper of Cab
70- Riding on Unit Exterior
80- Lap
99- Unk./NA

Restraint/Safety Device
00- None Used
11- Shoulder
12- Lap Belt Only
13- Shoulder & Lap Belt
21- Child Safety Seat
88- Other
99- Unknown

Pedestrian, Motor/Pedalcycle Only
31-Helmet
41- Protective Pads
51-Reflective Clothing
61-Lighting

Air Bag Deployment / Switch
a) 1-Deployed Fron 2-Deployed Side 3-Deployed Both 4-Not Deployed 7-Not Applicable 9-Deployment Unk.
b) 1- Switch in On Position 2- Switch in Off Position 3- No Switch 9- Unknown

Ejection
1- Not Ejected
2- Part. Ejected
3-Tot. Ejected
7- Not Applicable
9- Unknown

Location After Impact
1- Not Trapped
2- Extricated (Mechanical Means)
3- Freed (non-mech.)
4- Not Applicable
9- Unknown

a) Transported to Medical Facility
1- Yes 2- No 3- Unknown
b) By: 1- EMS 2- Police 8- Other 9- Unknown

Sequence of Events

Non-Collision
01- Cargo/Equip Loss or Shift
02- Cross Median/Center
03- Downhill Runaway
04- Equipment Failure
05- Fire/Explosion
06- Immersion
07- Jackknife
08- Overturn/Rollover
09- Ran off Road Left
10- Ran off Road Right
11- Separation of Units
12- Spill (Two-Wheeled Veh.)
18- Other Noncollision
19- Unk. Non-collision

Event 1	Event 2	Event 3	Event 4	Most Hrmfl	1st Hrmfl
109	110	111	112	1 113	114
2	2	2	2	2	
3	3	3	3	3	

Collision: Not Fixed
20- Animal (Deer Only)
21- Animal (All Other)
22- Motor Veh. (In Transport)
23- Motor Veh. (Stopped)
24- Motor Veh. (Other Roadway)
25- Motor Veh. (Parked)
26- Pedalcycle
27- Pedestrian
28- Railway Veh.
29- Work Zone Maint. Equip.
38- Other Movable Object
39- Unk. Movable Object

Collision: Fixed Object
40- Bridge Overhead Structure
41- Bridge Parapet End
42- Bridge Pier or Abutment
43- Bridge Rail
44- Culvert
45- Curb
46- Ditch
47- Embankment
48- Equipment
49- Fence
50-Guardrail End
51- Guardrail Face
52- Highway Traffic Sign Post
53- Impact Attenuator/Crash Cushion
54- Light/Luminaire Support
55- Mail Box
56- Median Barrier
57- Overhead Sign Support
58- Other (Post, Pole, Support, Etc.)
59- Other (Wall, Building, Tunnel, Etc.)
60- Tree
61- Utility Pole
62- Work Zone Maint. Equipment
68- Other
69- Unknown

Manner of Collision (Struck Veh.)
115 / 2 / 3 — 1st Hrmfl Collision
00- Not Coll. w/ Motor Veh.
10- Rear End
20- Head On
30- Rear-to-Rear
41 - Angle (↘ ↙
42- Angle (→ ←
43- Angle (↗ ↖
50- Sideswipe Same Dir.
60- Sideswipe Opposite Dir.
70- Backed Into
99- Unknown

1st / Most Deformed Area
1st Deformed: 1 116 | 2 | 3
Most Deformed: 1 117 | 2 | 3
21- Pedestrian 81- None 92- Rollover 93- Total 94-Under Carriage 98- Other 99- Unknown

Vehicle Type
118 / 2 / 3
01- Automobile
12- Pickup Truck
13- Truck Tractor
14- Other Truck
15- Full Size Van
16- Mini Van
17- Sport Utility
25- Motorcycle
26- Other Motorbike
27- Pedalcycle
38- Animal Drawn Veh
39- Animal (Ridden)
41- Pedestrian
51- Train
61- School Bus
62- Passenger Bus
98- Other
99- Unk. (Hit and Run Only)

Alcohol / Drug Test Given
119 120 (A2 D2, A3 D3)
1- Given - Known Results
2- Given - Unusable
3- Given - Pending
4- None
5- Refused

Special Use Only
1 121	2	3	4	5	6	7

Vehicle Use Code
122 / 2 / 3
01- Personal
02- Driver Training
03- Construction/Maint.
04- Ambulance
05- Military
06- Transport Passengers
07- Transport Property
08- Farm Use
09- Wrecker or Tow
10- Police
11- Government
12- Fire Fighting
13- Logging
18- Other
41- Pedestrian

Test Type
123 124 (A2 D2, A3 D3)
1- Breath (Alc Only)
2- Blood
3- Urine
4- Serum
8- Other

Underride/Override 125 / 2 / 3
1- Under- Compartment Intrusion
2- Under- No Intrusion
3- Under- Unknown
4- Over- MV in transport
5- Over- Other Vehicle
6- None
9- Unknow

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

Drug Results
127 (D2, D3)
1- Amphetamines
2- Cocaine
3- Marijuana
4- Opiates
5- PCP
8- Other

Extent of Deformity 128 / 2 / 3
0- None/Minor
2- Functional Damage
3- Disabling Damage
4- Severe/Totaled
5- Not Applicable
9- Unknown

Alc Test Results
A1- 130
A2-
A3-

Trafficway 131
1- Two -way, Not Divided
2- Two-way, Divided, Unprotected Median
3- Two-way, Divided, Barrier
4- One-Way
8- Other

1st Harmful Event Loc. 133
A: 1- Gore 2- Island 3- Median 4- Roadside 5- Roadway 6- Shoulder 7- Sidewalk 8- Outside Trafficway 9- Unk.
B: X-walk: 132 1- Y 2- N 9- U

Action Prior to Impact
129 / 2 / 3
(Vehicle)
01- Backing
02- Changing lanes
03- Entering traffic lane
04- Leaving traffic lane
05- Making U-turn
06- Movements Essentially Straight Ahead
07- Overtaking/passing
08- Parked
09- Slowing or Stopped in traffic
10- Turning left
11- Turning right
88- Other
99- Unknown
(Non-motorist)
21- Approaching/Leaving Vehicle
22- Entering/Crossing Location
23- Playing/Working on Vehicle
24- Pushing Vehicle
25- Standing
26- Walking, Playing, Cycling
27- Working

Road Character 134
1- Straight - Level
2- Straight - On grade
3- Straight - Hillcrest
4- Curve - Level
5- Curve - On grade
6- Curve - Hillcrest

Road Surface Condition 135
1- Dry
2- Wet
3- Snow
4- Slush
5- Ice
6- Contaminate
7- Water (Standing, etc.)
8- Other
9- Unk.

Traffic Control Type 136
01- Stop and Go Light
02- Flashing Traffic Signal
11- RR (X-bucks, Lights & Gates)
12- RR (X-bucks & Lights)
13- RR (X-bucks Only)
21- Officer or Flagman
22- Oncoming Emergency Vehicle
31- Pavement Markings (only)
41- Stop Sign
42- School zone Sign
43- Yield Sign
44- Work Zone
45- Other Warning Signs
51- Flashing Beacon
98- None
99- Unk.

Weather Condition 137
1- Clear (no adverse conditions)
2- Rain
3- Cloudy
4- Sleet, Hail
5- Snow
6- Fog, Smog, Smoke
7- Blowing Sand, Oil, Dirt, or Snow
8- Severe Crosswinds
9- Unknown

School Bus Involved: 139
1- Yes, Directly 2- Yes, Indirectly 3- No 9- Unk.

Light Condition 138
1- Daylight
2- Dawn
3- Dusk
4- Dark (Lighting Unspecified)
5- Dark (Street Lamp Lit)
6- Dark (Street Lamp Not Lit)
7- Dark (No lights)

Work Zone: 140
1- Yes 2- No

Work Zone Location 141
1- Before 1st Sign
2- Advanced Warning Area
3-Transition Area
4- Activity Area
5-Termination Area

Work Zone Type 142
1- Shoulder/Median Work
2- Lane Shift/Crossover
3- Intermittent/Moving Work
4- Lane Closure
8- Other
9- Unk.

Workers Present: 143
1- Yes 2- No

Junction Type 144
01- Crossover
02- Driveway
03- Five/More Points
04- Four-way Intersection
05- Railway Grade Crossing
07- Shared Use Paths or Trails
08- T-Intersection
09- Traffic Circle
12- Y - Intersection
13- Nonjunction
99- Unk.

Contributing Factors
Primary 145
Others: 146 147 148 149

Driver
01- Disregarded Signs, Signals, Etc.
02- Distracted/Inattention
03- Driving Too Fast for Conditions
04- Exceeded Authorized Speed Limit
05- Failed to Yield Right of Way
06- Ran off Road
07- Fatigued/Asleep
08- Followed Too Closely
09-Made an Improper Turn
10- Medical Related
12- Aggressive Operation of Vehicl
13- Over-correcting/Over-steering
14- Swerving to Avoiding Object
15- Wrong Side or Wrong Way
16- Under the Influence
17- Vision Obscured (Within Unit)
18- Improper lane Usage/Change
19- On Cell Phone
20- Texting
28- Other Improper Action
29- Unknown

Roadway
30- Debris
31- Non-highway Work
32- Obstruction in Roadway
33- Road Surface Condition (I.e., Wet)
34- Rut, Holes, Bumps
35- Shoulders (None, Low, Soft, High)
36- Traffic Control Device (I.e., Missing)
37- Work Zone (Constr./Maint./Utility)
38- Worn, Travel-Polished Surface
48- Other
49- Unknown

Non- Motorist
50- Inattentive
51- Lying &/or Illegally in Roadway
52- Failure to Yield R. of W.
53- Not Visible (Dark Clothing)
54- Disregard Signs, Signals, Etc.
55- Improper Crossing
56- Darting
57- Wrong Side of Road
66-Under the Infl.
67- Other Person Under Infl
58- Other
59- Unk

Environmental
60- Animal in Road
61- Glare
62- Obstruction
63- Weather Cond.
68- Other
69- Unknown

Vehicle Defect
70- Brakes
71- Steering
72- Power Plant
73- Tires/Wheel
74- Lights
75- Signals
76- Windows/Shield
77- Restraint System
78- Truck Coupling
79- Cargo
80- Fuel System
88- Other
89- Unk.

Figure 5. South Carolina traffic collision form (TR-310), back side.

When information in a form field does not match the written narrative, the entire traffic collision form is classified as having a discrepancy. An example of a discrepancy is shown in Figures 6 and 7. The narrative describes Unit 2 as moving and Unit 3 as stopped in traffic, but the relevant form field has this information backward.

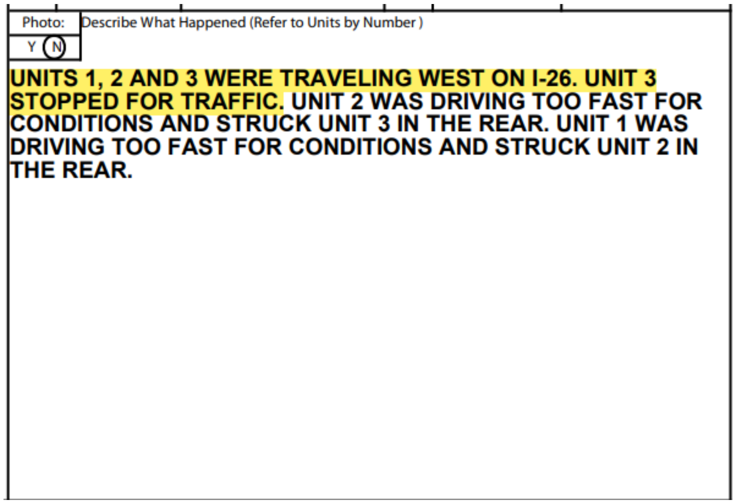
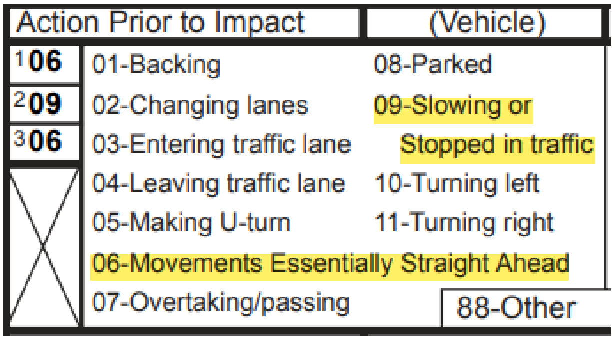


Photo: Y [✓ N]

Describe What Happened (Refer to Units by Number)

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 6. Discrepancy example, written narrative (Field 86). Highlights added for clarity.



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 7. Discrepancy example, action prior to impact (Field 129). Highlights added for clarity.

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 are treated as a single discrepancy. For example, Fields 109 to 112 in Figure 5 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. A total of 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.

A summary of the frequency of discrepancies at the form level is shown in Table 4. It can be seen that 63.5%, 31%, and 5.5% of the forms contained 0, 1, and 2 discrepancies, respectively. The discrepancies by form field are shown in Table 5. The fields with the most discrepancies are 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 4. Number of forms with discrepancies between form fields and narrative.

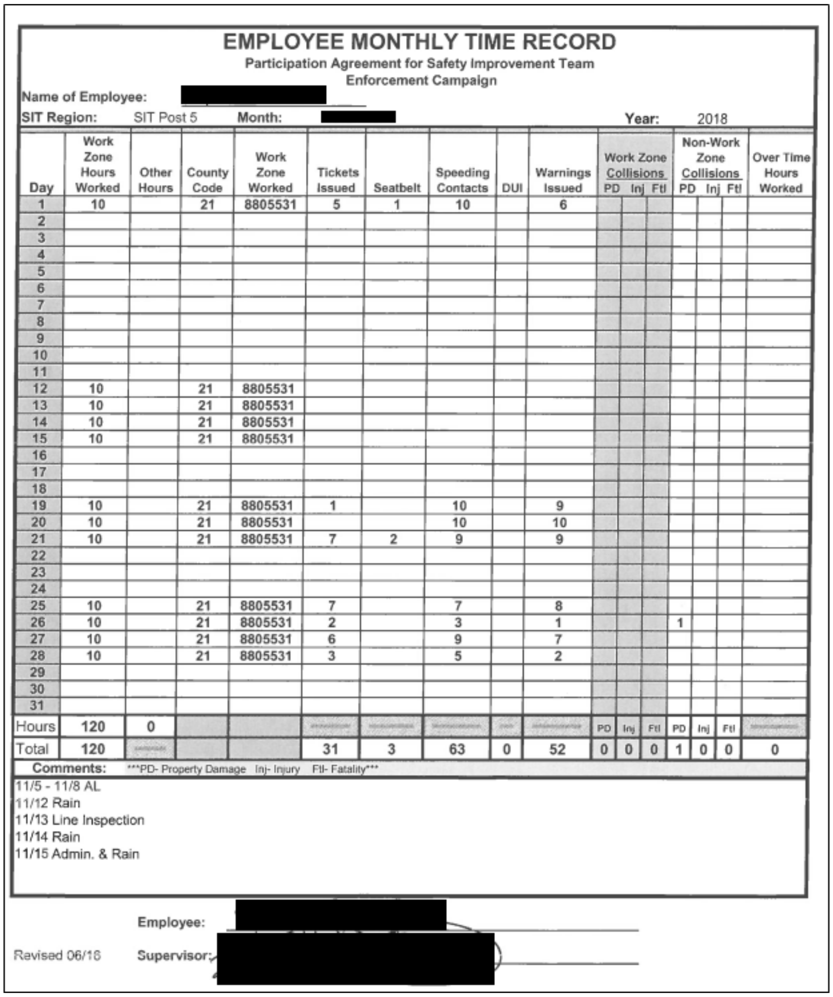
Number of Discrepancies	Traffic Collision Form Count
0	127
1	62
2	11

Table 5. 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

3.1.3. Data for Law Enforcement Effectiveness

To determine the effectiveness of law enforcement, data on police presence was compared to average vehicle speed at five selected work zones. Data on police presence was acquired from the monthly invoices from the Safety Improvement Team (SIT) provided by SCDOT in PDF format for all months between January 2018 and December 2020. Each PDF file contained pages for individual officers displaying the number of hours spent at each work zone identifier. One such page is shown in Figure 8. For the months between January 2018 and December 2019, the officer name, hours worked, and work zone identifier were manually entered into a spreadsheet format. Months within 2020 were excluded at the recommendation of the project steering and implementation committee (PSIC) due to the potential impact of the COVID-19 pandemic on the data. Additionally, it was assumed that law enforcement is “present” if the hours spent at the site are 10 or more, and there is no law enforcement if the hours spent at the site are less than 10 hours. The reason for this is that the assignment is intended to be full-time. Anything less than 10 hours is assumed to not have the intended effect. The percentage of days when troopers spent less than 10 hours on site is 5.6% among the five projects analyzed in this study. Therefore, this assumption does not have a significant bearing on the results.



EMPLOYEE MONTHLY TIME RECORD

Participation Agreement for Safety Improvement Team
Enforcement Campaign

Name of Employee: [redacted]

SIT Region: SIT Post 5 Month: [redacted] Year: 2018

Day	Work Zone Hours Worked	Other Hours	County Code	Work Zone Worked	Tickets Issued	Seatbelt	Speeding Contacts	DUI	Warnings Issued	Work Zone Collisions			Non-Work Zone Collisions			Over Time Hours Worked
										PD	Inj	Ftl	PD	Inj	Ftl	
1	10		21	8805531	5	1	10		6							
2																
3																
4																
5																
6																
7																
8																
9																
10																
11																
12	10		21	8805531												
13	10		21	8805531												
14	10		21	8805531												
15	10		21	8805531												
16																
17																
18																
19	10		21	8805531	1		10		9							
20	10		21	8805531			10		10							
21	10		21	8805531	7	2	9		9							
22																
23																
24																
25	10		21	8805531	7		7		8							
26	10		21	8805531	2		3		1				1			
27	10		21	8805531	6		9		7							
28	10		21	8805531	3		5		2							
29																
30																
31																
Hours	120	0								PD	Inj	Ftl	PD	Inj	Ftl	
Total	120				31	3	63	0	52	0	0	0	1	0	0	0

Comments: ***PD- Property Damage Inj- Injury Ftl- Fatality***

11/5 - 11/8 AL
11/12 Rain
11/13 Line Inspection
11/14 Rain
11/15 Admin. & Rain

Employee: [redacted]

Revised 06/18 Supervisor: [redacted]

Figure 8. One SIT Invoice for November 2018.

From this spreadsheet, a list of all identifiers used by the officers was acquired. Because the type of identifier (contract or project) was not stated in the SIT invoices, researchers searched for each identifier in the P2S database. Five work zones with substantial police presence and P2S information available were selected to gather data on average vehicle speed, as shown in Table

6. Analysis was limited to these five work zones due to the time-consuming nature of acquiring speed data for each work zone.

Table 6. Contract and project IDs for the five work zones analyzed for this project.

Contract ID	Project ID
1091731	24011
4208281	P029074
4210170	P029699
5384210	38111
3288840	P027003

To determine average vehicle speed, each work zone location was identified in GIS. The locations were cross-referenced with a shapefile provided by SCDOT containing link IDs associated with roadway segments. SCDOT provided vehicle speed data for each link ID that the researchers requested. For each work zone, two groupings of link IDs were determined: one for all link IDs within the mile markers of the project, and one for the link ID representing the beginning of the project. There are no records regarding the exact location where troopers positioned themselves in the work zone area, but the SCDOT generally asks troopers to be positioned in the transition area (assumed to be the beginning mile marker of the project). For work zones affecting more than one road, the link IDs were determined separately for each road, and speed data was combined using an average weighted by the length of each affected road. For a given interstate link ID, SCDOT provided speed data for each hour of each day in 2019. For non-interstate link IDs, a spreadsheet showing the TCD IDs associated with each link ID was provided, and the hourly speed data for each TCD ID was given. As such, for non-interstate roads, link ID lists representing each road were converted to TCD ID lists before speed data was determined.

Because officers only record the number of hours spent at a given work zone and not the times of day in SIT invoices, vehicle speed was averaged over different daily periods. These periods included 0:00 to 11:59, 12:00 to 23:59, 7:00 to 16:59, and 0:00 to 23:59. The final dataset included one value (in miles per hour) for the average speed over the entire work zone and one value for the average speed at the start of the work zone for each work zone on each day of 2019.

During one of the monthly meetings, the project committee expressed concern that comparing average vehicle speed to police presence from the SIT invoices would not accurately reflect real-life conditions because this comparison did not account for lane closures. It was suggested that highway patrol troopers are most often present during lane closures, and thus, lane closures could potentially account for slowdowns in vehicle speed. The only source of information available for lane closures was daily work reports (DWRs) provided by SCDOT. Given the volume of reports, keywords were used to designate days with lane closures rather than individually studying each report. Lane closures were determined present at a work zone on a given day if one of the selected keywords (“closure”, case-sensitive “LC”, or “traffic control”) was mentioned in at least one daily work report for that day. The results of this query were compared to the number of days known to have police presence from the SIT invoices as well as the number of days each project was active in 2019. The results are shown in Table 7. The PSIC concluded that information from the daily work reports was insufficient to accurately capture the effect of lane closures. As such, results from the models developed to measure law enforcement effectiveness should be taken with the disclaimer that they cannot account for the impact of lane closures on average vehicle speed.

Table 7. Comparison of project duration, daily work report findings, and SIT invoice data for 2019.

Project ID	Duration	DWR Keyword Hits	Police Presence	Overlapping Days
24011	365	31	136	18
P027003	365	215	183	138
P029074	326	-	79	-
P029699	121	-	77	-
38111	365	277	167	143

3.1.4. Data for Negative Binomial Crash Prediction Model

The crash prediction model was designed to provide an expected number of crashes for a work zone given the work zone's length, duration, and AADT. Length and duration were to be acquired from the P2S data provided by SCDOT at the beginning of the research project; however, it was discovered when comparing against data from the SIT invoices previously mentioned that the P2S data did not seem to provide a comprehensive list of work zones. To replace the original dataset, the PSIC recommended that the researchers manually extract all P2S data for four work zone types: widening, rehabilitation, reconstruction, and preservation. This amended P2S dataset included information regarding length, duration, and project type. For work zone length, project descriptions listed the roads affected by the project and their beginning and ending mile points. The highest mile point for each road was subtracted from the lowest mile point for each road mentioned in the project description, and all separate road lengths were added together if the work zone project included more than one road. For work zone duration, project IDs included a notice to proceed (NTP) date and a substantial work completion (SWC) date. The NTP date was subtracted from the SWC date to find the duration. Because crash data was only provided for 2014 to 2020, in cases where the SWC date was empty (implying the project was ongoing) or after 12/31/2020, a date of 12/31/2020 was used instead, and NTP dates were replaced with 01/01/2014 if they were earlier than 2014.

To find AADT, traffic count station locations were extracted from SCDOT's GIS resources. Using the work zone location information from P2S, a list of all traffic count stations within the work zone's boundaries was determined for each work zone. For all years during which the work zone was active, traffic counts from historical data provided by SCDOT were averaged together to determine a single AADT value for a work zone over the project duration. For work zones affecting more than one road, the AADT for each segment was found separately and combined using an average weighted by each segment's length. The descriptive statistics of the variables used in the crash prediction model are shown in Table 8: length (miles), duration (days), and AADT (veh/day). The statistics provided for each project type are 25th percentile, 75th percentile, minimum, maximum, mean, and standard deviation.

Table 8. Descriptive statistics of crash prediction model variables.

Variable	Project Type	25th pct	75 pct	Min	Max	Mean	Std. Dev.
Length	Widening	1.25	5.6	0.06	21.23	4.14	4.48
Length	Rehabilitation	1.5	4.38	0.01	36.82	3.66	4.02
Length	Preservation	1.6	6.26	0.02	61.66	5.55	7.15
Length	Reconstruction	1.68	4.27	0.05	22.05	3.39	2.9
Duration	Widening	484.0	1416.75	91.0	2037.0	920.48	531.38
Duration	Rehabilitation	270.0	537.0	14.0	1535.0	414.56	233.29
Duration	Preservation	163.0	331.0	23.0	1091.0	265.23	164.57
Duration	Reconstruction	213.0	466.0	16.0	794.0	350.28	199.63
AADT	Widening	4843.75	14968.75	683.0	69400.0	14772.92	17003.52
AADT	Rehabilitation	1680.25	9141.5	100.0	95975.0	8047.96	12935.95
AADT	Preservation	1312.5	10750.0	50.0	126508.0	9561.56	14433.04
AADT	Reconstruction	1050.0	6220.0	50.0	61533.0	5558.82	7708.49

The researchers and PSIC discussed accounting for traffic control measures in the zero-inflated negative binomial model, but this was not done due to the limited availability of traffic control data in South Carolina. Researchers attempted to gather traffic control data from two separate sources: traffic control plans from ProjectWise and daily work reports provided by SCDOT. The traffic control plans provided information regarding traffic control practices at each work zone but did not include any dates or times during which these measures were implemented. Daily work reports were consulted as an attempt to fill this gap in information. Researchers found that the volume of daily work reports made it difficult to filter useful from irrelevant information. Any mentions of traffic control were not substantial enough for use in conjunction with the traffic control plans; as such, it was determined that the available data was insufficient to determine how effective a particular traffic control was at the work zone.

To count how many crashes occurred in each work zone, the P2S dataset was filtered by crash date, road type, route number, and mile marker where the crashes occurred. This helped us identify where and when each crash took place within the work zones. The process of filtering the P2S dataset is shown in Figure 9.

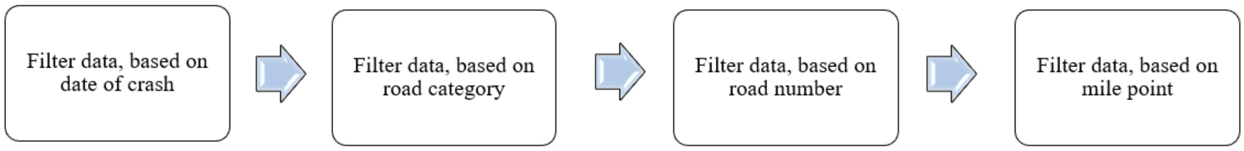

Filter data, based on date of crash
Filter data, based on road category
Filter data, based on road number
Filter data, based on mile point

Figure 9. The process of filtering the P2S dataset to find the number of crashes in each work zone

3.2. Methods to Identify Contributing Factors

Based on the previous studies, mixed logit models have been considered as an efficient method to overcome unobserved heterogeneity due to their ability to account for observation-specific variation for explanatory variables Anastasopoulos and Mannering (2011); Anderson and Hernandez (2017); Chen et al. (2019). Thus, two separate analyses were conducted to identify contributing factors using mixed logit models which are explained below:

3.2.1. Factors Affecting Injury in Interstate and Non-Interstate Work Zone Crashes

Sub-sections 3.2.1.1, 3.2.1.2, and 3.2.1.3 present the mathematical details of the mixed logit model, marginal effect of factors and likelihood ratio test Madarshahian et al. (2023).

3.2.1.1. Mixed Logit Model

The utility function for mixed logit model is derived by establishing the linear relation between injury severity level i for observation crash n which is demonstrated in Eq. (1) Madarshahian et al. (2023)

$$y_{in} = \beta_i x_{in} + \varepsilon_{in} \tag{1}$$

where y_{in} is defined as a variable explaining each injury severity level i (i in I representing injury and no injury severity level) for driver n. β_i is considered as a vector of estimated parameters, x_{in} is a vector of explanatory variables affecting work zone driver-injury severity level i and ε_{in} is the error term or extreme value to capture unobserved heterogeneity distributed independent and identically over time, individual and alternatives. If the error term is generalized extreme value distributed, then the choice probability can be determined using the standard multinomial logit shown in Eq. (2).

$$p_{n(i)} = \frac{\exp\,[\beta_i x_{in}]}{\sum_{i\epsilon I} \exp\,[\beta_i x_{in}]} \tag{2}$$

$p_{n(i)}$ is regarded as the probability of injury severity level i caused by driver n. Mixed logit models allow the vector of estimated parameters to vary across different crashes. Each element of β_i

may be either fixed or randomly distributed with fixed means, allowing for heterogeneity within the observed crash dataset. Extending the above multinomial logit model, Eq. (3) can be rewritten as:

$$p_{n(i|\emptyset)} = \frac{\exp[\beta_i x_{in}]}{\sum_{i\epsilon I} \exp[\beta_i x_{in}]} f(\beta_i|\emptyset) d\beta_i \tag{3}$$

where $p_{n(i|\emptyset)}$ is the weighted average of the multinomial logit probabilities called mixed logit. The weight used to estimate the probability is calculated by $f(\beta_i|\emptyset)d\beta_i$ which is the density function of β_i and $\emptyset$ is the parameter vector. The density function uses a distribution of parameter $\emptyset$, where both a mean and variance are estimated. For the current work, normal distribution is used. It should be noted that elements of β_i are fixed and randomly distributed with specific statistical distributions. If the estimated variance is statistically significant then the modeled injury severity levels vary with respect to x across observations and account for crash-specific variation due to unobservables[59]. To overcome the computation complexity of estimating the parameters β_i maximum likelihood estimation is implemented using simulation-based procedure and Halton draws ("Statistical and Econometric Methods for Transportation Data Analysis"). The pseudo R-squared (ρ^2) value is used to assess the overall model fit; it is computed using Eq. (4).

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)} \tag{4}$$

In above equation $LL(0)$ is defined as the log-likelihood at zero and $LL(\beta)$ calculates log-likelihood at convergence.

3.2.1.2. Marginal Effect

Marginal effect is used to determine how the probability of injury severity levels would be changed considering one unit change in the explanatory variables illustrated in. Eq. (5).

$$M_{X_{ink}}^{p_{in}} = p_{in}[givenX_{ink} = 1] - p_{in}[givenX_{ink} = 0] \tag{5}$$

In the above equation, p_{in} states the probability of injury severity level i for driver n and X_{ink} is the k-th independent variable affecting injury severity level i for driver n.

3.2.1.3. Likelihood Ratio Test

To determine whether the data should be modeled using two different speed categories, the log-likelihood ratio (LR) test between the full model using the entire dataset and speed models using separate datasets shown in Eq. (6) was performed ("Statistical and Econometric Methods for Transportation Data Analysis," n.d.).

$$LR_{full} = -2[LL(\beta^{full}) - LL(\beta^{speed<60}) - LL(\beta^{speed\geq60})] \tag{6}$$

The model's log likelihood at convergence for the full model on the entire dataset is defined as $LL(\beta^{full})$ while $LL(\beta^{speed<60})$ and $LL(\beta^{speed\geq60})$ are the model's log-likelihood at the convergence on the separated data sets for speed limit less than 60 mph and speed limit greater than or equal to 60 mph respectively. It should be noted that to calculate the log-likelihood values for two separate speed limits, the variables identified from the full model should be tested on the two categorized speed limit datasets. LR statistic has χ^2 distribution with the degree of freedom computed by the difference among the summation of the number of estimated variables in two models and the number of estimated variables in the full model.

Parameter transferability is another test ("Statistical and Econometric Methods for Transportation Data Analysis"). often used to ascertain whether two different speed limits should be modeled separately; it is calculated using Eq. (7).

$$LR_{a_b} = -2[LL(\beta^{a_b}) - LL(\beta^{a})] \tag{7}$$

The log likelihood at convergence for speed model a on the data from model b is defined as $LL(\beta^{a_b})$ and the log likelihood at convergence for speed model a is defined as $LL(\beta^{a})$. The degrees of freedom of this test are equal to the number of estimated variables in β^{a_b}.

To estimate contributing factors affecting injury severity levels and to test the need to estimate separate models, the NLOGIT software (version 6) was used. The process used to produce model estimates is shown in Figure 10. As shown, this study used three datasets provided by the South Carolina DOT: (1) unit crash dataset which contains information of all vehicles involved in crashes, (2) location dataset which provides environmental and temporal characteristics of the crashes, and (3) occupant dataset which includes details about the occupants of all vehicles involved in the crashes. It's important to note that each dataset has a different number of observations. To create the final dataset for modeling in Nlogit software, the three datasets were merged using a common index, and this was accomplished using Python programming. The dataset was then filtered to include only truck-involved crashes and relevant variables. Subsequently, the variables were categorized based on previous research and SCDOT practices. The last data preparation step involved creating binary variables for modeling. The forward and backward stepwise selection method was used to arrive at the final model specification. A variable was retained in the model if it is significant at the 90% confidence interval. Models were compared against one another using the log-likelihood at convergence and the McFadden Pseudo R-squared Madarshahian et al. (2023).

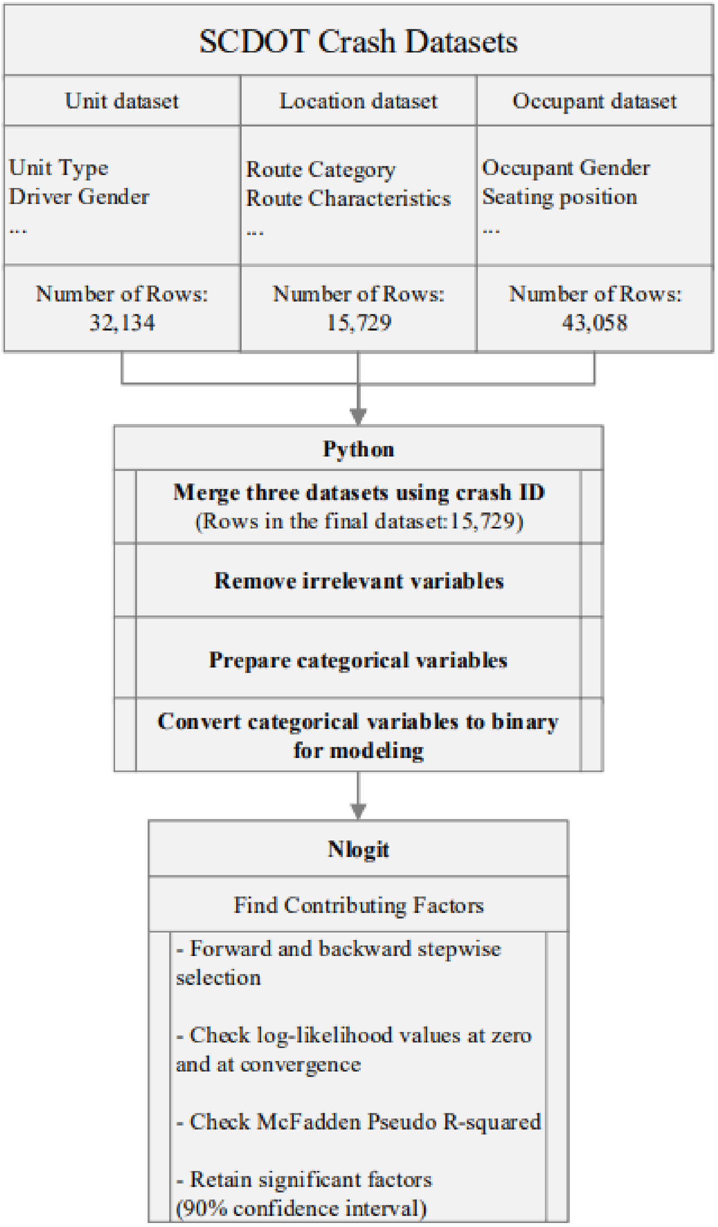

SCDOT Crash Datasets
Unit dataset
Location dataset
Occupant dataset
Unit Type
Driver Gender
...
Route Category
Route Characteristics
...
Occupant Gender
Seating position
...
Number of Rows: 32,134
Number of Rows: 15,729
Number of Rows: 43,058
Python
Merge three datasets using crash ID
(Rows in the final dataset:15,729)
Remove irrelevant variables
Prepare categorical variables
Convert categorical variables to binary for modeling
Nlogit
Find Contributing Factors
- Forward and backward stepwise selection
- Check log-likelihood values at zero and at convergence
- Check McFadden Pseudo R-squared
- Retain significant factors (90% confidence interval)

Figure 10. Model estimation process.

3.2.2. Factors Affecting Injury in Work Zone Rear End Crashes where collision speed ≥ 35 mph

The method used for this analysis is the same as the one described in Section 3.2.1, except that it accounts for heterogeneity in both mean and variance. In such a model, the vector of estimable parameters is permitted to vary across crash observations Mannering et al. (2016); Seraneeprakarn et al. (2017) as shown in Eq. (8).

$$\beta_{in} = \beta_i + \Phi_{in}\mathrm{Z}_{in} + \sigma_{in}EXP(\Psi_{in}W_{in})\nu_{in} \tag{8}$$

The parameter estimate β_i represents the average value calculated for all crashes. The vector Z_{in} comprises explanatory variables specific to crash, which account for heterogeneity in the mean impacting injury severity level i, the vector Φ_{in} consists of coefficients assigned to estimable parameters, W_{in} is the vector of crash-specific explanatory variables that address the heterogeneity in the standard deviation σ_{in} having an associated parameter vector Ψ_{in}, and ν_{in} is considered as a disturbance term. To address the computational complexity associated with estimating the parameters β_i, a simulation-based method and Halton draws O'Donnell and Connor (1996) are utilized in implementing maximum likelihood estimation.

3.3. Analyze Crash Report Narratives and Identify Discrepancies

This project determined discrepancies within individual traffic collision forms by comparing the narrative text (Field 86) to information recorded in the form fields. In this project, the text in the narrative field was 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 affect the officer's cognitive ability to recall information and record it in a consistent manner. To this end, this project sought to determine the level of discrepancies in South Carolina traffic collision forms and to identify factors that may have contributed to the discrepancies. The researchers postulated 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 Craik and Lockhart (1972). The psychology-based approach to understanding discrepancies in traffic collision forms is unique in the study of misclassification. Both structural equation modeling (SEM) and multiple linear regression (MLR) were used to identify factors that may have contributed to the discrepancies. Specifically, SEM was used to investigate the relationships between latent variables and level of processing, and MLR was used to investigate factors that affect the frequency of discrepancies in form fields.

3.3.1. Structural Equation Model

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 the license class of the at-fault driver. The level of processing was 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 (LCD) 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 9. 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 9. 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

SEM allows the relationship between different latent variables to be modeled. In this project, 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 project, the observed variables are those shown in Table 9. 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 6 and the latent variables were developed. The weather conditions factor was operationalized by wind speed, temperature, humidity, and precipitation. The crash characteristics factor was 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 was 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 project, the endogenous latent variable is the 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.

$$x = \mathit{\Lambda}_x \xi + \delta \tag{9}$$

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. $\mathit{\Lambda}_x$ 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.

$$y = \mathit{\Lambda}_y \eta + \varepsilon \tag{10}$$

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. $\mathit{\Lambda}_y$ 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.

$$\eta = \beta\eta + \mathit{\Gamma}\xi + \zeta \tag{11}$$

where β is an $(m \times m)$ matrix of coefficients for the effects between latent endogenous variables. Since this project 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 . RMSEA is computed using the following equation:

$$RMSEA = \sqrt{\frac{\chi_M^2 - df_M}{df_M(N-1)}} \tag{12}$$

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 a good model fit (Boonyoo and Champahom (2022); Champahom et al. (2020); Hair et al. (2006); Mw (1993); Shi et al. (2011); Steiger (2007); Wang and Qin (2014)). 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 . The equation for TLI is shown below.

$$TLI = \frac{\chi_B^2/df_B - \chi_M^2/df_M}{\chi_B^2/df_B - 1} \tag{13}$$

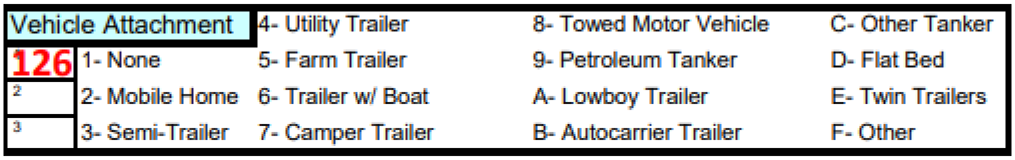
The equation for CFI is shown below.

$$CFI = 1 - \frac{max\ (\chi_M^2 - df_M, 0)}{max\ (\chi_B^2 - df_B, 0)} \tag{14}$$

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 has good relative fit .

3.3.2. Multiple Linear Regression

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 11, form field 126 (Vehicle Attachment) has an input count of one for each unit, with three boxes provided. If there are more than three units, a second page is required. For each input/box, the officer can select from 15 options. 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 10.



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 11. Form field 126 (Vehicle Attachment). The red number is for labeling each field and does not appear on the actual form.

Table 10. 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

For MLR, the following assumptions are made: (1) the residuals are normally distributed, (2) there is a linear relationship between the dependent and independent variables, (3) the variance of errors is consistent across independent variables (homoskedasticity), and (4) the independent variables are independent . The data set used for the MLR model was assessed and found to satisfy the assumption criteria Osborne and Waters (2019); Uyanık and Güler (2013); Williams et al. (2019). An MLR model was created to assess the effect of the 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.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon \qquad (15)$$

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 the 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 Eberly (2007); Favero et al. (2023); Jobson (2012).

3.4. Law Enforcement Effectiveness

To determine the effectiveness of law enforcement, a split-plot design with a blocking factor was used and AADT was used as the covariate. The split-plot design structure is a hierarchical or multi-level design consisting of experimental units with two different sizes with separate randomization steps Stroup et al. (2018). It is a useful design when it is difficult to have the same size of experimental units and by using it we can remove some variability due to the larger experimental units. That is, we can gain extra precision for some comparisons compared to a factorial treatment design. The main factor in the design was the season, which had four levels, while the subplot factor was the presence of state troopers. We also accounted for the type of work zone (treated as a blocking factor). Overall, we assessed and compared eight different models as depicted in Table 11. Models 1 through 4 were based on the average speed of the entire work zone as the response variable, whereas Models 5 through 8 focused on the average speed specifically within the transition area. For each response type, we examined two variations. The first involved subtracting the temporary posted speed limit from the average speed. This approach aimed to shed light on the degree of speeding or adherence to the temporary posted speed limit. Another variation entailed the inclusion of a covariate. Traffic volume was introduced as a covariate, and relative efficiency was computed to gauge whether adding the covariate enhanced the model's performance. The eight evaluated models are listed below. Models 1, 3, 5, and 7, excluding the covariate, can be represented as:

$$y_{ijk} = \mu + \alpha_i + w_{ij} + \tau_k + (\alpha\tau)_{ik} + b_j + \varepsilon_{ijk} \qquad (16)$$

Where y_{ijk} represents the response variable, which is the average speed in the work zone for Models 1-4, and the average speed where troopers are stationed for Models 5-8, corresponding to season i $(i = 1, 2, 3, 4)$, where $i = 1$ represents 'Fall', $i = 2$ represents 'Winter', $i = 3$ represents 'Spring', and $i = 4$ represents 'Summer'. The blocks are denoted by j $(j = 1, 2, 3, 4, 5)$, and police presence is denoted by k $(k = 1, 2)$, where $k = 1$ indicates no police presence in the work zone and $k = 2$ indicates police presence in the work zone. The block term denotes the different types of work zones, with $j = 1$ corresponding to 'Widening', $j = 2$ to 'Bridge Replacement', $j = 3$ to 'Resurfacing', $j = 4$ to 'Interchange Improvement', and $j = 5$ to 'Rehabilitation'. The overall mean is denoted by μ. α_i stands for the effect of season i, which is the main plot effect. w_{ij} represents the main plot error term, also interpreted as the interaction between the main plot effect and the block effect (the interaction between the season effect and the type of work zone). τ_k signifies the subplot effect, reflecting the effect of police presence. $(\alpha\tau)_{ik}$ indicates the interaction between the main plot and subplot effects (the interaction between the season and police presence effects). b_j denotes the block effect, representing the effect of the type of work zone. Lastly, ε_{ijk} represents the subplot. The main plot error and subplot error terms are assumed to follow identical and independent normal distributions.

$(w_{ij} \sim N(0, \sigma_w^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma^2))$.

Models 2, 4, 6, and 8, incorporating the covariate, can be expressed as:

$$y_{ijk} = \mu + \alpha_i + w_{ij} + \tau_k + (\alpha\tau)_{ik} + b_j + \beta(x_{ijk} - \bar{x})\varepsilon_{ijk} \qquad (17)$$

In Eq. (16), the terms mentioned, apart from β, x_{ijk}, and $\bar{x}$, have been previously defined. In Eq. (17), we introduce the covariate into the model using the term $\beta(x_{ijk} - \bar{x})$. Here, β represents the coefficient signifying the relationship between the response y_{ijk} and the specific covariate x_{ijk}, which in this case is traffic volume. The data were centered by subtracting the overall mean $\bar{x}$ from the actual covariate. It was determined that a common slope β across all treatment combinations was adequate.

Table 11. Splot-plot Models Analyzed

Model	Response Variable	Subtract Temporary Posted Speed Limit from Observed Speed	Add Traffic Volume as a Covariate
1	Average speed throughout work-zone	No	No
2	Average speed throughout work-zone	No	Yes
3	Average speed throughout work-zone	Yes	No
4	Average speed throughout work-zone	Yes	Yes
5	Average speed where troopers stationed	No	No
6	Average speed where troopers stationed	No	Yes
7	Average speed where troopers stationed	Yes	No

Model	Response Variable	Subtract Temporary Posted Speed Limit from Observed Speed	Add Traffic Volume as a Covariate
8	Average speed where troopers stationed	Yes	Yes

3.4.1. Model Efficiency with Traffic Volume as a Covariate

To evaluate the efficacy of incorporating a covariate into the model for error control, one could assess the difference in error variances when comparing models with and without the covariate adjusted by the treatment, as denoted in Eq. (18).

$$E = \frac{MSE_{(COV)}[1 + \frac{T_{xx}}{(t-1)E_{XX}}]}{MSE} \tag{18}$$

The efficiency (denoted by E) of including a covariate in the analysis can be quantified. Mean Squared Error (MSE) is a metric that gauges the average squared deviation between observed and predicted values. It is computed for both models, with and without the covariate. In this context, E_{XX} represents the sum of squared discrepancies between each observed value of the covariate and its mean, while T_{xx} signifies the sum of squared discrepancies between the predicted values of the covariate from the ANCOVA model and its mean. In Eq. (18), "t" corresponds to the number of treatment groups being compared.

3.5. Inflated Zero Negative Binomial Crash Prediction Model

There were a number of work zones with zero crashes as shown in Table 12. It can be seen that 39% of widening projects had zero crashes, 82% for rehabilitation, 86% for reconstruction, and 89% for preservation. For this reason, the zero-inflated Negative Binomial (ZINB) model was used for it is designed to handle overdispersion and excessive zeros simultaneously.

Table 12. Percentage of zero crashes for each work zone type

Percentage of zero crashes			
Widening	Rehabilitation	Reconstruction	Preservation
39%	82%	86%	89%

To further explain Table 12, it should be noted that a 39% widening indicates that among all crashes occurring in all widening projects, 39% of these projects experienced no crashes. This trend is also consistent for rehabilitation, reconstruction, and preservation projects.

The ZINB model accounts for both the frequency of non-zero counts and the presence of excess zeros in the data as shown in Eq. (19):

$$E(crashes) = (1 - \pi_0) \times e^{x_i \beta} \tag{19}$$

In the formula shown above, the value of π_0 is determined by the following equation:

$$\pi_0 = \frac{1}{1 + e^{-\acute{x_\imath}\beta}} \tag{20}$$

In the ZINB model, π_0 represents the likelihood of observing excess zeros, indicating the probability that a given observation arises from the zero-inflated component of the model rather than the count component. The variable $\acute{x_\imath}$ in Eq. (20) signifies the predictor or explanatory variable (work zone length, work zone duration, and AADT) to estimate the probability of excess zeros for each observation in the dataset. Additionally, β represents the coefficients linked to the predictor variables $\acute{x_\imath}$, illustrating their impact on the probability of excess zeros; positive coefficients imply an increase in the likelihood of observing excess zeros, while negative coefficients suggest the opposite effect. On the contrary, $x_i\beta$ in Eq. (19) arises from the count model, representing the linear combination of predictor variables x_i (work zone length, work zone duration, and AADT) and their corresponding coefficients β. This term encapsulates the relationship between the predictors and the expected count of non-zero observations, providing insights into how changes in the predictors influence the count outcome. In essence, $x_i\beta$ quantifies the impact of the predictor variables on the expected count, facilitating the prediction of non-zero observations in the dataset.

It should be noted that we used the natural logarithm of AADT as the independent variable for predicting the crash counts. The reason for using log(AADT) instead of just AADT is that its variance is significantly different from that of other variables, resulting in parameter estimation errors. By using the logarithmic transformation, we successfully mitigated this issue. The ZINB model was estimated using the R statistical software. A copy of the code is provided in Appendix A.

4. Findings

4.1. Contributing Factors

There are many angles from which factors that contribute to work zone-related crashes in SC could be analyzed. In this project, two different angles are taken. The first is to determine if there is any difference in the contributing factors between work zones on roads with speed limits of 60 mph or higher and work zones on roads with speed limits less than 60 mph. The motivation for this analysis is to determine if the stringent work zone guidelines required for interstates lower injury risk. If so, the SCDOT could consider increasing traffic control standards for work zones on lower-speed roads. The other angle is to determine if rear-end crashes with collision speeds greater than or equal to 35 mph increase injury risk. If so, the SCDOT could consider putting in countermeasures to reduce the traffic speed through the work zones.

4.1.1. Factors Affecting Injury in Interstate and Non-Interstate Work Zone Crashes

The log-likelihood ratio test yielded a value of 20.92 with 10 degrees of freedom ($\rho\_value<$ 0.022); the log-likelihood value for the full model is -1329.43, the log-likelihood value for the posted speed limit less than 60 mph is -744.83, and the log-likelihood value for the posted speed limit greater than or equal to 60 mph equals is -574.14. To find the log-likelihood values for the different speed categories, the full model is needed, and its estimation results are shown in Table 13.

Table 13. Parameters estimate and marginal effects for full model.

Variable	Coefficient	t-statistic	p-value	Marginal Effects	
				Injury	PDO
Defined for injury					
Rear End (standard deviation of parameter distribution)	0.86 (1.037)	4.02 (1.68)	0.000 (0.09)	0.064	-0.064
Constant	-0.49	-2.31	0.020		
Two vehicles	-1.24	-8.67	0.000	-0.109	0.109
Interstate	-0.42	-3.36	0.000	-0.039	0.039
Dark	0.42	3.52	0.000	0.017	-0.017
Female	0.53	3.54	0.000	0.011	-0.011
Weekday	-0.40	-2.62	0.009	-0.441	0.441
Lane shift/Crossover	-0.49	-2.25	0.025	-0.004	0.004
Under Influence	-1.04	-2.76	0.006	0.004	-0.004
Model Statistics					
Number of observations	3064				
Log-likelihood at zero, $LL(0)$	-2123.8				
Log-likelihood at convergence, $LL(\beta)$	-1329.4				
$\rho^2 = 1 - LL(\beta)/LL(0)$	0.37				

The result of the log-likelihood test suggests that the two different speed limit groups should be modeled separately with over 95% confidence. It follows from the parameter transferability tests that two separate mixed logit models are to be estimated, one for a posted speed limit less than 60 mph (representing non-interstates) and one for a posted speed limit greater than or equal to 60 mph (representing interstates). Table 14 shows the results of the parameter transferability test Madarshahian et al. (2023).

Table 14. Results of parameter transferability tests for two speed categories.

Speed limit category	Speed limit category	
	$< 60\ mph$	$\geq to\ 60\ mph$
$< 60\ mph$	-	32.69 (p<0.001)
$\geq to\ 60\ mph$	28.28 (12) (p=0.005)	-

Each model predicts two levels of injury severity: injury and PDO. A simulation-based maximum likelihood method was utilized to estimate the parameters β for the mixed logit models. To estimate random parameters, the normal distribution was considered, and 500 Halton draws were used. The normal distribution was adopted because it was found to be statistically significant in several previous studies Uddin and Huynh2020, 2017). During the model development process, variables were retained in the specification if they had t-statistics corresponding to the 90% confidence level or higher on a two-tailed t-test. The random parameters were retained if their standard deviations had t-statistics corresponding to the 90% confidence level or higher. Model estimation results are shown in Tables 14 and 15 along with marginal effects for all the variables included in the final specifications. It should be noted that other speed grouops such as < 50 mph and ≥ 50 mph were not evaluated. Thus, the following results and their implications apply only to the selected speed groups, < 60 mph and ≥ 60 mph.

Table 15. Parameter estimates and marginal effects for the model with a speed limit < 60 mph.

Variable	Coefficient	t-statistic	p-value	Marginal Effects	
				Injury	PDO
Defined for injury					
Two vehicles (standard deviation of parameter distribution)	-2.37 (2.72)	-3.13 (3.12)	0.002 (0.002)	-0.0044	0.0044
Constant	-2.40	-5.78	0.000		
SC, US primary	1.10	3.85	0.000	0.2880	-0.2880
Dark	0.67	2.78	0.005	0.0176	-0.0176
Female	0.71	2.25	0.024	0.0096	-0.0096
Age less than 35	0.51	2.32	0.020	0.0133	-0.0133
Activity area	0.49	-2.12	0.034	0.0304	-0.0304
Driving too fast	1.09	-4.48	0.000	0.0404	-0.0404
Sideswipe	-0.86	2.81	0.005	-0.0171	0.0171
Workers present	0.45	-2.01	0.004	0.0249	-0.0249
Fixed Object	-1.28	3.53	0.000	-0.0097	0.0097

Model Statistics	
Number of observations	1748
Log-likelihood at zero, $LL(0)$	-1211.62
Log-likelihood at convergence, $LL(\beta)$	-730.77
$\rho^2 = 1 - LL(\beta)/LL(0)$	0.397

Table 16. Parameter estimates and marginal effects for the model with a speed limit ≥ 60 mph.

Variable	Coefficient	t-statistic	p-value	Marginal Effects	
				Injury	PDO
Defined for injury					
Constant	-2.42	-7.40	0.000		
Shoulder median (standard deviation of parameter distribution)	-1.1 (2.62)	2.13 (3.74)	0.033 (0.000)	0.0325	-0.0325
Multi vehicles	1.82	7.20	0.000	0.0484	-0.0484
Driving too fast	0.61	2.52	0.012	0.0330	-0.0330
Rear end	0.96	3.86	0.000	0.0526	-0.0526
Weekday	-0.71	-2.68	0.007	-0.0607	0.0607
Before first sign	0.64	-1.80	0.072	0.0051	-0.0051
Dark	0.95	-4.16	0.000	0.0308	-0.0308
Female	0.65	-2.35	0.019	0.0094	-0.0094
Model Statistics					
Number of observations	1305				
Log-likelihood at zero, $LL(0)$	-904.56				
Log-likelihood at convergence, $LL(\beta)$	-567.27				
$\rho^2 = 1 - LL(\beta)/LL(0)$	0.37				

Table 15 shows the parameter estimates for the model corresponding to work zone crashes where the posted speed limit of the roadway is less than 60 mph. A positive coefficient implies that the variable is positively associated with the likelihood of that specific injury severity level. In other words, an increase in an independent variable with a positive coefficient results in a higher probability of occurrence of the specific injury severity level. In this model, one indicator variable, *Two vehicles*, has a statistically significant standard deviation (random parameter). This result suggests that the effect of the *Two vehicles'* variable on injury severity varied significantly across crashes. This coefficient is normally distributed with a mean of -2.37 and a standard deviation of 2.72, indicating that this variable has a positive impact on 19.18% of observations (increases the likelihood of an injury crash) and a negative impact on 80.82% of observations (decreases the likelihood of an injury crash). This finding suggests that for a majority of truck-involved crashes at work zones where the roadway posted speed limit is below 60 mph (representing non-interstates), the involvement of two vehicles (as opposed to three or greater) reduces the likelihood of an injury crash.

Table 16 shows the parameter estimates for the model corresponding to work zone crashes where the posted speed limit of the roadway is 60 mph or greater. In this model, one indicator variable, *Shoulder/Median* (from field 142 shown in Figure 4), has a statistically significant standard deviation (random parameter). This result suggests that the effect of the *Shoulder/Median* variable on injury severity varied significantly across crashes. This coefficient is normally distributed with a mean of 1.1 and a standard deviation of 2.62, indicating that this variable has a positive impact on 66.27% of observations (increases the likelihood of an injury crash) and a negative impact on 33.73% of observations (decreases the likelihood of an injury crash). This finding suggests that for a majority of truck-involved crashes at work zones where the roadway posted speed limit is 60 mph or greater (representing interstates), crash occurrence on a shoulder or median increases the likelihood of injury. A possible explanation for this is the use of concrete barriers on interstates in South Carolina and the smaller clear zones in some areas. According to the Federal Highway Administration, "By creating Clear Zones, roadway agencies can increase the likelihood that a roadway departure results in a safe recovery rather than a crash, and mitigate the severity of crashes that do occur."

Building separate injury severity models based on posted speed limits allows for a deeper understanding of how contributing factors vary across different speed limit ranges. The two models presented in this section show that there are considerable differences in terms of the combination of factors affecting injury severity, and the magnitude of the impact of these factors. These results highlight the fact that the posted speed limit of the roadway where the work zone is located interacts greatly with other factors impacting injury severity. Table 17 provides a summary of the variables that are statistically significant for the two speed-limit groups. The random parameters are not included in this table because they have varying impacts across observations.

Table 17. Models Comparison.

Variable	**Speed < 60 mph**		**Speed ≥ 60 mph**	
	Injury	PDO	Injury	PDO
SC, US Primary	↑	↓		
Dark	↑	↓	↑	↓
female	↑	↓	↑	↓
Younger Driver	↑	↓		
Activity area	↑	↓		
Driving too fast	↑	↓	↑	↓
Sideswipe	↓	↑		
Workers present	↑	↓		
Fixed object	↓	↑		
3+ vehicles				↓
Rear End			↑	↓
Before 1st sign			↑	↓
Weekday			↓	↑

4.1.2. Factors Affecting Injury in Work Zone Rear End Crashes where collision speed ≥ 35 mph

The mixed-logit model which takes into account variations in both mean and variance was estimated using the NLOGIT (version 6) software. The estimation of parameters β_i was conducted through a simulation-based maximum likelihood method with 1000 Halton draws. In analyzing random parameters, a normal distribution was assumed, following its statistical significance as noted by previous studies Uddin and Huynh (2020, 2017). Variables were included in the model if their t-statistics met or surpassed the 90% confidence level, with random parameters retained if their variance showed significance at the same confidence level. Table 18 shows the final model coefficients and corresponding t-statistics, p-values, marginal effects, and base level. Interpretation of the mixed logit model is the same as that of the multinomial logit model, wherein a positive coefficient indicates a positive association with injury probability. Notably, the model demonstrates a favorable statistical fit, evidenced by an ρ^2 value of 0.2. The random parameter linked to the "Interstate" variable reveals a statistically significant standard deviation, suggesting variability in its impact on injury severity across different crashes. Further analysis reveals that the "Interstate" variable predominantly influences injury severity positively for approximately 83.52% of the cases, with a minority (16.48%) showing a negative impact. This implies a higher likelihood of injury in rear-end crashes at work zones on interstates in South Carolina compared to non-interstates when collision speed exceeds 35 mph. Moreover, the estimation highlights heterogeneity in both mean and variance of the "Interstate" random parameter, with variations noted based on crash time and lighting conditions, as well as the presence of drivers under the influence contributing to increased variance.

Table 18. Model estimation results.

Variable	Coefficient	t-statistic	p-value	Marginal Effects		Base Level
				Injury	PDO	
Interstate (Standard deviation of parameter distribution)	-0.78 (0.80)	-5.02 (1.83)	0.000 (0.067)	-0.051	0.051	N.A
Heterogeneity in the mean of random parameter						
Interstate: Late night (1 if crash occurs between 12-6 a.m., 0 otherwise)	0.6	2.60	0.009			
Interstate: Dawn or Dusk (1 if crash occurred in a dawn or dusk lighting condition, 0 otherwise)	1.70	3.03	0.002			
Heterogeneity in the variance of random parameter						
Interstate: Under influence (1 if the contributing factor of crash is under the influence, 0 otherwise)	1.35	1.77	0.076			
Constant	-0.35	-1.70	0.089			
3+ Vehicles	0.76	7.29	0.000	0.042	-0.042	2 Vehicles

Variable	Coefficient	t-statistic	p-value	Marginal Effects		Base Level
				Injury	PDO	
Airbag deployed	1.32	10.31	0.000	0.073	-0.073	Airbag not deployed
Termination/transition	-0.39	-1.95	0.050	-0.012	0.012	Before first sign
Advanced warning area	-0.48	-2.33	0.020	-0.012	0.012	Before first sign
Activity area	-0.39	-2.14	0.033	-0.043	0.043	Before first sign
Lane shift/crossover	-0.39	-2.41	0.016	-0.006	0.006	Lane closure
Shoulder/Median	-0.29	-2.85	0.004	-0.028	0.028	Old driver
Middle-aged drivers	-0.38	-2.96	0.003	-0.013	0.013	Old driver
Dawn or Dusk	-1.10	-2.26	0.024	-0.006	0.006	Day light
Dark	0.29	2.69	0.007	0.012	-0.012	Day light
Truck involved	0.51	4.29	0.000	0.016	-0.016	No-truck involved
Model statistics						
Number of observations	3648					
Log-likelihood at zero, $LL(0)$	-2528.60					
Log-likelihood at convergence, $LL(\beta)$	-2037.27					
$\rho^2 = 1 - \; LL(\beta)/LL(0)$	0.2					

Table 19 summarizes the impact of statistically significant variables concerning rear-end crashes at collision speeds greater than or equal to 35 mph. The random parameter is not included since its effects vary across observations. These variables exhibit a positive effect on injury: involvement of 3 or more vehicles, deployment of airbags, dark conditions, and involvement of one or more trucks. All other variables demonstrate a negative effect.

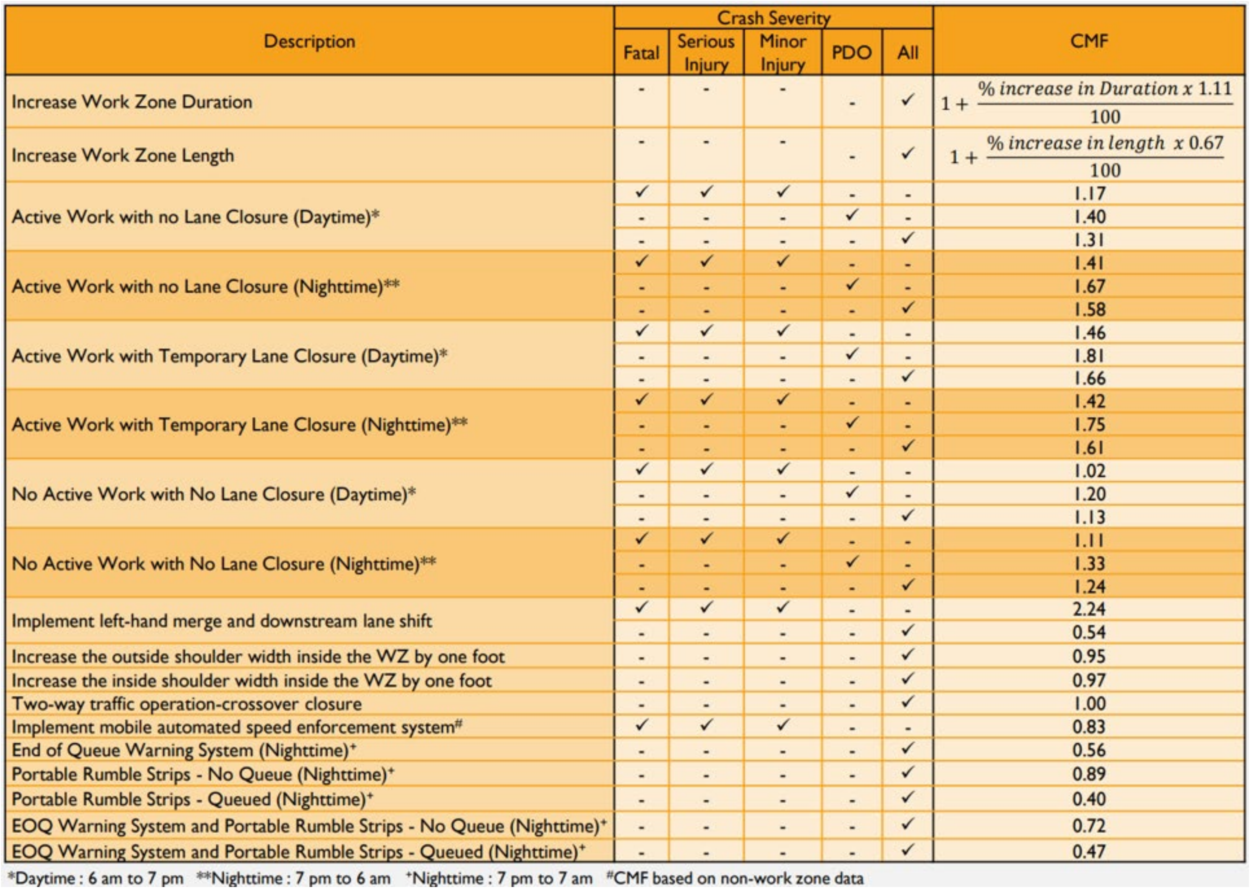
Table 19. effect of variables.

Variable	Base Level	Injury
3+ Vehicles	2 Vehicles	↑
Airbag deployed	Airbag not deployed	↑
Termination/transition	Before first sign	↓
Advanced warning area	Before first sign	↓
Activity area	Before first sign	↓
Lane shift/crossover	Lane closure	↓
Shoulder/Median	Lane closure	↓
Young drivers	Old drivers	↓
Middle-aged drivers	Old drivers	↓
Dawn or Dusk	Day light	↓
Dark	Day light	↑
Truck involved	No-truck involved	↑

4.2. Countermeasures

As detailed in Section 3.1.4, efforts to extract pertinent work zone data from traffic control plans and daily work reports proved unsuccessful. Given the absence of suitable SC data to assess the effectiveness of specific traffic controls or countermeasures at work zones, it is recommended that the SCDOT consider utilizing work zone countermeasures developed specifically for work zones by the University of Missouri. These countermeasures were established through research supported by the U.S. Department of Transportation under cooperative agreement numbers DTFH6113RA00019 and 693JJ31750003. Their corresponding Crash Modification Factors (CMFs) are shown in Figure 12 which were used in the work zone risk assessment tool to calculate the reduction in the number of crashes. For the "Active Work with no Lane Closure" countermeasure, the baseline is "no work zone." For the "Implement left-hand merge and downstream lane shift," the baseline is the conventional right-lane closure. When more than one countermeasure is used, the combined effect can be determined using Eq. (21):

$$CMF_T = CMF_1 \, X \, CMF_2 \, X \, \ldots \tag{21}$$

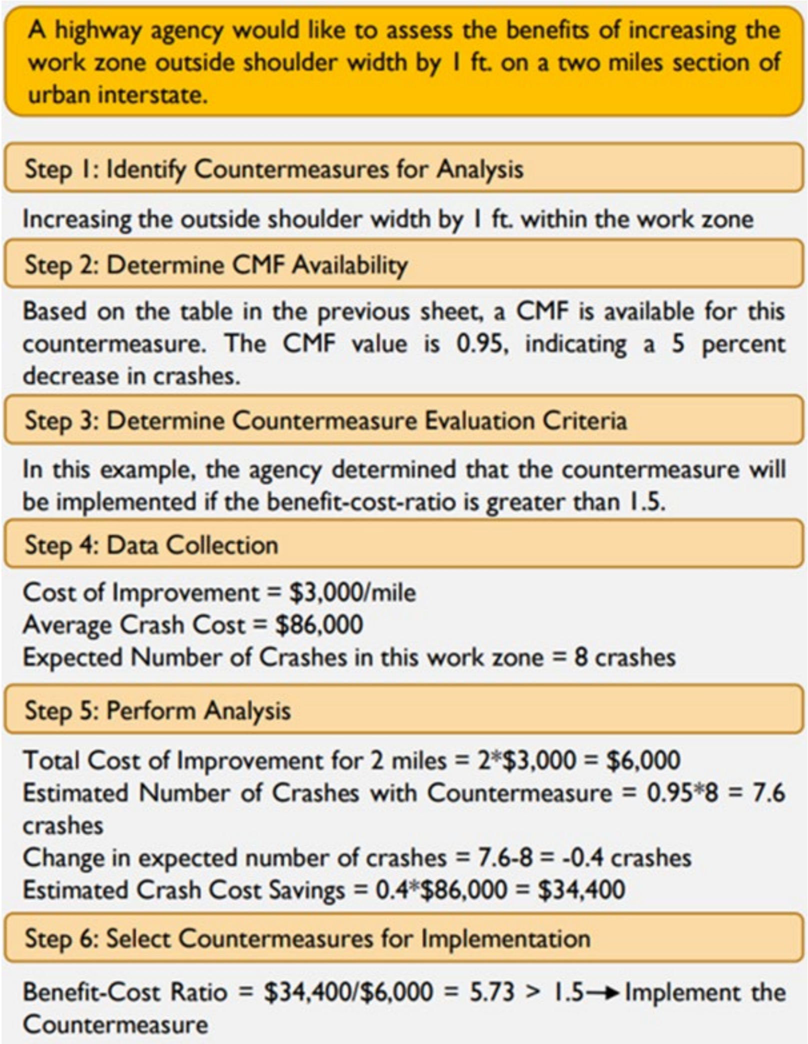


Description	Crash Severity					CMF
	Fatal	Serious Injury	Minor Injury	PDO	All	
Increase Work Zone Duration	-	-	-	-	✓	$1+\frac{\%\ increase\ in\ Duration\ x\ 1.11}{100}$
Increase Work Zone Length	-	-	-	-	✓	$1+\frac{\%\ increase\ in\ length\ \ x\ 0.67}{100}$
Active Work with no Lane Closure (Daytime)*	✓	✓	✓	-	-	1.17
	-	-	-	✓	-	1.40
	-	-	-	-	✓	1.31
Active Work with no Lane Closure (Nighttime)**	✓	✓	✓	-	-	1.41
	-	-	-	✓	-	1.67
	-	-	-	-	✓	1.58
Active Work with Temporary Lane Closure (Daytime)*	✓	✓	✓	-	-	1.46
	-	-	-	✓	-	1.81
	-	-	-	-	✓	1.66
Active Work with Temporary Lane Closure (Nighttime)**	✓	✓	✓	-	-	1.42
	-	-	-	✓	-	1.75
	-	-	-	-	✓	1.61
No Active Work with No Lane Closure (Daytime)*	✓	✓	✓	-	-	1.02
	-	-	-	✓	-	1.20
	-	-	-	-	✓	1.13
No Active Work with No Lane Closure (Nighttime)**	✓	✓	✓	-	-	1.11
	-	-	-	✓	-	1.33
	-	-	-	-	✓	1.24
Implement left-hand merge and downstream lane shift	✓	✓	✓	-	-	2.24
	-	-	-	-	✓	0.54
Increase the outside shoulder width inside the WZ by one foot	-	-	-	-	✓	0.95
Increase the inside shoulder width inside the WZ by one foot	-	-	-	-	✓	0.97
Two-way traffic operation-crossover closure	-	-	-	-	✓	1.00
Implement mobile automated speed enforcement system#	✓	✓	✓	-	-	0.83
End of Queue Warning System (Nighttime)+	-	-	-	-	✓	0.56
Portable Rumble Strips - No Queue (Nighttime)+	-	-	-	-	✓	0.89
Portable Rumble Strips - Queued (Nighttime)+	-	-	-	-	✓	0.40
EOQ Warning System and Portable Rumble Strips - No Queue (Nighttime)+	-	-	-	-	✓	0.72
EOQ Warning System and Portable Rumble Strips - Queued (Nighttime)+	-	-	-	-	✓	0.47

*Daytime : 6 am to 7 pm **Nighttime : 7 pm to 6 am +Nighttime : 7 pm to 7 am #CMF based on non-work zone data

Figure 12. Work zone crash modification factors (“Development of Work Zone Crash Modification Factors (CMFs),” n.d.).

Figure 13 illustrates the process incorporated into the risk assessment tool for determining whether countermeasures should be implemented. The first step is to select suitable countermeasures for the identified contributing factors. The second step is to calculate the combined effect using Eq. (21) if more than one countermeasure is selected. The third step is to determine the threshold for the benefit-cost ratio (BCR). Generally, if a project has a BCR greater than 1.0, then it is expected to deliver a positive net present value to the agency. However, some agencies may prefer to have a higher threshold than 1.0. The fourth step is to determine the cost of implementing the countermeasures per mile, the average crash cost, and the expected number of crashes in the work zone. The fifth step is to determine the estimated crash cost savings and total cost of improvement. The sixth and last step is to calculate their quotient (i.e., BCR) and if it is greater than the threshold determined in the third step, then the implementation of the countermeasures is justified.



A highway agency would like to assess the benefits of increasing the work zone outside shoulder width by 1 ft. on a two miles section of urban interstate.

Step 1: Identify Countermeasures for Analysis

Increasing the outside shoulder width by 1 ft. within the work zone

Step 2: Determine CMF Availability

Based on the table in the previous sheet, a CMF is available for this countermeasure. The CMF value is 0.95, indicating a 5 percent decrease in crashes.

Step 3: Determine Countermeasure Evaluation Criteria

In this example, the agency determined that the countermeasure will be implemented if the benefit-cost-ratio is greater than 1.5.

Step 4: Data Collection

Cost of Improvement = $3,000/mile
Average Crash Cost = $86,000
Expected Number of Crashes in this work zone = 8 crashes

Step 5: Perform Analysis

Total Cost of Improvement for 2 miles = 2*$3,000 = $6,000
Estimated Number of Crashes with Countermeasure = 0.95*8 = 7.6 crashes
Change in expected number of crashes = 7.6-8 = -0.4 crashes
Estimated Crash Cost Savings = 0.4*$86,000 = $34,400

Step 6: Select Countermeasures for Implementation

Benefit-Cost Ratio = $34,400/$6,000 = 5.73 > 1.5→ Implement the Countermeasure

Figure 13. Procedure for Determining benefit-cost of contemplated Countermeasures.

4.3. Crash Report Narratives and Discrepancies

4.3.1. Discrepancies by Form

First, the Confirmatory Factor Analysis or CFA was conducted to assess the fit of the proposed model. The results indicated a good model fit, so the SEM model was developed. Both CFA and SEM analysis were performed using SPSS Amos. Figure 14 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 the 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 the level of processing is explained by crash characteristics and weather conditions.

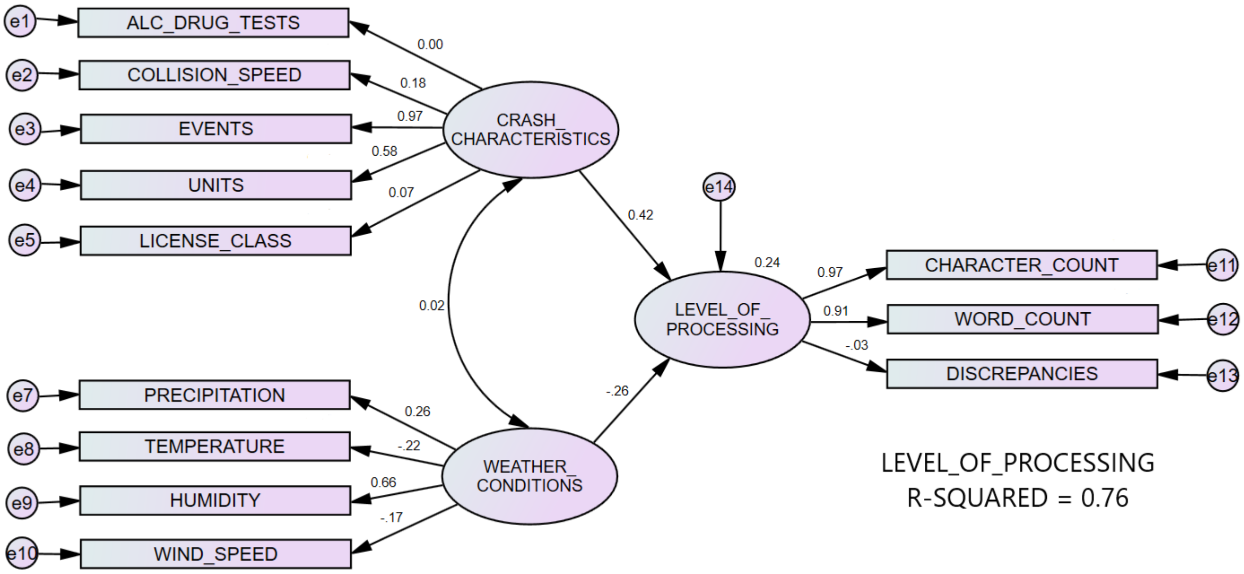


Figure 14. SEM Results. Regression estimates have been standardized.

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 the level of processing. Humidity and precipitation have positive effects on weather conditions (0.26 and 0.66, respectively), and thus will lower the level of processing with an increase in value due to the negative relationship between weather conditions and the 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 the level of processing; an increase in temperature will increase the level of processing. Multiplying their coefficients shows a positive impact of temperature on the 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 the 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.

4.3.2. Discrepancies by Type

The MLR model estimation results are shown in Table 20. The model's R-squared and adjusted R-squared are 0.752 and 0.695, respectively, indicating a 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 20. 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

4.4. Law Enforcement Effectiveness

The findings from eight models discussed in Section 3.4 are examined using the PROC MIXED2 method in SAS OnDemand for Academics. Table 21 displays the outcomes for all potential models.

Table 21. Test of model's fixed effects.

Model 1[1]		
Effect	**Degree of freedom**	**P-value**
Seasons	3	0.6641
Trooper presence	1	0.0001
Seasons* Trooper presence	3	0.2484
Model 2		
Effect	**Degree of freedom**	**P-value**
Seasons	3	0.4557
Trooper presence	1	0.0001
Seasons* Trooper presence	3	0.0891
Model 5[2]		
Effect	**Degree of freedom**	**P-value**
Seasons	3	0.1145
Trooper presence	1	0.0001
Seasons* Trooper presence	3	0.0062
Model 6		
Effect	**Degree of freedom**	**P-value**
Seasons	3	0.0600
Trooper presence	1	0.0001
Seasons* Trooper presence	3	0.0012

1: Models 3 and 4 yield identical results to Models 1 and 2, respectively.
2: Models 7 and 8 yield identical results to Models 5 and 6, respectively.

In Model 1, the presence of troopers significantly affected the average speed throughout the work zone, indicating effective law enforcement in reducing speed. However, seasons and the interaction between seasons and police presence didn't show significant effects. Model 2, which included traffic volume as a covariate, showed similar results, with trooper presence significantly impacting average speed in the work zone. There were no significant effects for seasons or their interaction with police presence. Models 3 and 4 focused on excess speed rather than average speed. Their results reaffirmed the effectiveness of law enforcement in reducing speed across the work zone, with no significant seasonal effects observed. Model 5 considered only the average speed in the area where troopers were stationed. Trooper presence significantly impacted speed, with winter showing the most significant reduction compared to fall and summer. Model 6, incorporating traffic volume as a covariate, also showed significant effects of trooper presence on speed, with winter exhibiting the most significant reduction compared to fall and summer. The results of Models 7 and 8, which focused on excess speed, mirror that of

Models 5 and 6, highlighting law enforcement's effectiveness in reducing excessive speed in the transition area and the presence of a significant interaction between seasons and trooper presence.

4.4.1. Efficiency of ANCOVA

After analyzing how traffic volume affects the models, we found that adding it as a factor in Model 1 (resulting in Model 2) and Model 3 (resulting in Model 4) gave almost the same results, with an efficiency of 1.016. The same goes for adding traffic volume to Model 5 (resulting in Model 6) and Model 7 (resulting in Model 8), efficiency is 1.021. This similarity is due to these models sharing the same criteria for evaluation. If the efficiency is over 1, including traffic volume will likely improve the result’s accuracy. However, since the efficiency is close to 1, keeping traffic volume as a covariate in Models 2, 4, 6, and 8 will not make much of a difference.

4.4.2. Evaluating Model Fitness: A Comparative Analysis

In Table 22, the goodness-of-fit (GOF) statistics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Corrected Akaike Information Criterion (AICC) for all eight models are presented. Model 3 has the lowest GOF statistics, indicating its superiority over Models 1, 2, and 4, which used average speed across the entire work zone as their response variable. On the other hand, for models using average speed in the transition area as their response variable, Model 7 has the lowest GOF statistics. However, considering the efficiency gained by adding traffic volume as a covariate, Model 4 is preferred over Model 3, and Model 8 is favored over Model 7.

Table 22. Comparison of Models’ Goodness-of-Fit statistics.

Model	AIC	BIC	AICC
1	9129.2	9128.0	9129.2
2	9256.4	9255.2	9256.4
3	9123.7	9123.7	9122.5
4	9251.1	9250.0	9251.2
5	9880.4	9880.5	9879.3
6	9996.6	9995.5	9996.6
7	9875.7	9874.5	9875.7
8	9990.9	9989.8	9991.0

4.5. Inflated Zero Negative Binomial Crash Prediction Model

The zero-inflated negative binomial models were estimated using the R statistical software. The estimation results for each work zone type are presented in subsequent subsections.

4.5.1. Work Zones for Widening Projects

Table 23 shows the estimation results of the ZINB model for widening projects. As shown, the results consist of two parts, a count model with negative binomial distribution and a zero-inflation model with binomial distribution. In the count model, significant predictors include Length, Duration, and log(AADT), indicating their positive association with crash counts. Although these variables are not significant in the zero-inflation model, it is best to retain them due to their importance based on previous research. All predictors have a positive coefficient which implies that as their values increase so will the crash count. The interpretation of the negative binomial regression coefficient is as follows: for a one unit change in the predictor variable, the log of expected counts of the response variable changes by the respective regression coefficient, given the other predictor variables in the model are held constant. A simpler way to interpret the coefficients is to use the Incidence Rate Ratio (IRR) which is shown in Table 22. The IRR for Length indicates that with each additional unit increase in Length, the crash count increases by 14.46%. Similarly, for each additional unit increase in Duration, the crash count increases by about 0.19%, and for each additional unit increase in the logarithm of AADT, the crash rate increases by approximately 96.05%.

Table 23. Estimation Results of ZINB model for Widening Projects.

	Count model coefficients					
	IRR	Estimate	Std. Error	z value	Pr(>|z|)	Significant
Intercept	0.0018	-6.3334	1.808551	-3.502	0.000462	***
Length	1.1446	0.135035	0.037864	3.566	0.000362	***
Duration	1.0019	0.001923	0.000391	4.921	8.60E-07	***
log(AADT)	1.9605	0.673193	0.201265	3.345	0.000823	***
Log(theta)	0.9005	-0.10486	0.252087	-0.416	0.677424	
	Zero-inflation model coefficients					
	Estimate		Std. Error	z value	Pr(>|z|)	Significant
Intercept	392.80902		545.98777	0.719	0.472	
Length	-7.02164		10.10864	-0.695	0.487	
Duration	-0.07483		0.10102	-0.741	0.459	
Log(AADT)	-37.28779		51.65359	-0.722	0.470	
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.5.2. Work Zones for Rehabilitation Projects

Table 24 shows the estimation results of the ZINB model for rehabilitation projects. The results are similar to that of widening projects. From the count model, the IRR for Length indicates that increasing it by one unit will increase the crash count by 11.2%. Similarly, increasing Duration by one unit will increase the crash count by 0.17%, and increasing the logarithm of AADT by one unit will increase the crash count by 235.35%.

Table 24. Estimation Results of ZINB model for Rehabilitation Projects.

	Count model coefficients					
	IRR	Estimate	Std. Error	z value	Pr(>|z|)	Significant
Intercept	0	-1.252e+01	1.257e+00	-9.962	< 2e-16	***
Length	1.112	1.062e-01	2.969e-02	3.577	0.000347	***
Duration	1.0017	1.743e-03	4.116e-04	4.235	2.29e-05	***
log(AADT)	3.3535	1.210e+00	1.383e-01	8.752	< 2e-16	***
Log(theta)	0.5911	-5.257e-01	2.304e-01	-2.281	0.022531	*
	Zero-inflation model coefficients					
	Estimate		Std. Error	z value	Pr(>|z|)	Significant
Intercept	14.301665		5.622745	2.544	0.0110	*
Length	-1.878046		0.755548	-2.486	0.0129	*
Duration	0.005086		0.003188	1.595	0.1107	
Log(AADT)	-1.427746		0.626858	-2.278	0.0227	*
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.5.3. Work Zones for Preservation Projects

Table 25 shows the estimation results of the ZINB model for preservation projects. The results are similar to that of widening and rehabilitation projects. From the count model, the IRR for Length indicates that increasing it by one unit will increase the crash count by 3.4%. Similarly, increasing Duration by one unit will increase the crash count by 0.3%, and increasing the logarithm of AADT by one unit will increase the crash count by 258.59%.

Table 25. Estimation Results of ZINB model for Preservation Projects.

	Count model coefficients					
	IRR	Estimate	Std. Error	z value	Pr(>|z|)	Significant
Intercept	0	-1.407e+01	1.412e+00	-9.968	< 2e-16	***
Length	1.0304	2.999e-02	1.352e-02	2.219	0.0265	*
Duration	1.003	2.994e-03	4.854e-04	6.169	6.88e-10	***
log(AADT)	3.5859	1.277e+00	1.410e-01	9.060	< 2e-16	***
Log(theta)	0.8091	-2.118e-01	2.502e-01	-0.847	0.3971	
	Zero-inflation model coefficients					
	Estimate		Std. Error	z value	Pr(>|z|)	Significant
Intercept	5.085280		3.851104	1.320	0.18668	
Length	-0.658410		0.239424	-2.750	0.00596	**
Duration	-0.000152		0.001516	-0.100	0.92014	
Log(AADT)	-0.331898		0.399613	-0.831	0.40623	
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.5.4. Work Zones for Reconstruction Projects

Table 26 shows the estimation results of the ZINB model for reconstruction projects. For this model, the Duration predictor is not significant. From the count model, the IRR for Length indicates that increasing it by one unit will increase the crash count by 12.1%. Similarly, increasing the logarithm of AADT by one unit will increase the crash count by 257.46%.

Table 26. Estimation Results of ZINB model for Reconstruction Projects.

	Count model coefficients					
	IRR	Estimate	Std. Error	z value	Pr(>|z|)	Significant
Intercept	0	-13.324255	1.763071	-7.557	4.11e-14	***
Length	1.121	0.114232	0.037731	3.028	0.00247	**
Duration	1.0011	0.001071	0.001207	0.888	0.37480	
log(AADT)	3.5746	1.273861	0.207155	6.149	7.78e-10	***
Log(theta)	1.7199	0.542267	0.577903	0.938	0.34807	
	Zero-inflation model coefficients					
	Estimate		Std. Error	z value	Pr(>|z|)	Significant
Intercept	-771.4504		601.2049	-1.283	0.199	
Length	4.0232		3.0292	1.328	0.184	
Duration	-0.8338		0.6506	-1.282	0.200	
Log(AADT)	103.4465		80.7976	1.280	0.200	
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.6 Summary

Based on the findings of the SEM and MLR models assessing discrepancies in police crash reports, 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. This study 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 study. 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 study 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 study cannot be generalized to the entire nation.

In terms of contributing factors on interstates, multiple vehicles, driving too fast for conditions, rear-end collisions, area before the first work zone sign, weekdays, dark light conditions, and female drivers are significant, whereas on non-interstates, factors such as specific road types, dark light conditions, female drivers, work zone activity areas, younger at-fault drivers, sideswipe collisions, worker presence, and collisions with fixed objects are significant. Marginal effects analysis suggests higher impacts of certain factors like driving too fast for conditions and female at-fault drivers on interstates. Additionally, darker lighting conditions have higher marginal effects on interstates, highlighting the need for brighter lighting in interstate work zones compared to non-interstates to mitigate the heightened risks associated with higher traveling speeds.

In terms of contributing factors to rear-end crashes with high collision speed (≥ 35 mph), compared to drivers aged 50 or older, those below 35 exhibit lower injury probabilities by 0.028, while those between 35 and 50 show a decrease by 0.013. Work zone configurations such as (termination/transition, advanced warning,) and activity areas demonstrate reduced injury probabilities by 0.012 and 0.043, respectively, compared to crashes occurring before the first sign. Conversely, crashes in lane closure configurations serve as the base level, indicating higher injury probabilities. Additionally, crashes involving trucks increase injury likelihood by 0.016 compared to those without trucks, while rear-end crashes involving three or more vehicles elevate injury probabilities by 0.042. Deployed airbags raise injury probabilities by 0.073 compared to crashes without airbag deployment, signifying higher collision severity. Moreover, crashes in dark conditions exhibit increased injury probabilities by 0.012, while those during dawn and dusk show decreases by 0.006, contradicting daylight crashes. These findings provide valuable insights for designing safer work zones and implementing targeted safety measures.

A split-plot design with blocking showed that there was a decrease in average speed across the work zone when troopers were present, and similar speed reductions were observed in the transition area. Additionally, trooper presence increases the likelihood of compliance with the posted temporary speed limits. Most models found no significant variation in average speed reduction between seasons. When the average speed of the entire work zone was considered, there is no interaction between seasons. However, when the average traffic speed around where troopers were stationed was considered, there was interaction between trooper presence and fall and summer seasons. Moreover, the average speed in the transition area was the lowest in the winter regardless of trooper presence.

5. Conclusions, Recommendations, and Implementation

5.1. Conclusions

The SEM and MLR models created to predict discrepancies in traffic collision suggest that site-specific factors affect written narrative length but not the frequency of discrepancies occurring. However, the frequency of discrepancies in a form field will increase with additional inputs or if it has a higher level of difficulty.

The mixed logit models for two separate speed limit models provided better insights than a single aggregate model, supported by statistical tests. Two mixed logit models developed for speed limits under 60 mph and 60 mph or greater showed factors contributing to injury included dark lighting conditions, female drivers at fault, and driving too fast for conditions. Additional significant factors varied by the speed limit group, such as roadway type, work zone activity area, and crash characteristics.

Developing a mixed logit model with heterogeneity in mean and variance for rear-end crashes with high collision speed revealed significant factors, including the presence of three or more vehicles, airbag deployment, specific work zone areas, lane configurations, at-fault driver's age, lighting conditions, and truck involvement. Countermeasures to mitigate injury include enhancing driver education, improving lighting and warning systems, installing rumble strips, educating the trucking industry and public about work zones, and improving traffic flow in congested areas. These findings emphasize the need for considering heterogeneity in future studies and implementing targeted safety measures in work zones.

Findings of split-plot design with blocking showed the efficacy of trooper presence in reducing speeds across the entire length of work zones and within transition areas, as well as in ensuring compliance with temporary speed limits. Seasonal influences were statistically insignificant. From a statistical standpoint, incorporating traffic volume as a covariate was found to enhance model precision.

5.2. Recommendations

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 to minimize the frequency of discrepancies. Future work in discrepancies could compare discrepancy rates across different states to assess South Carolina's officer training quality.

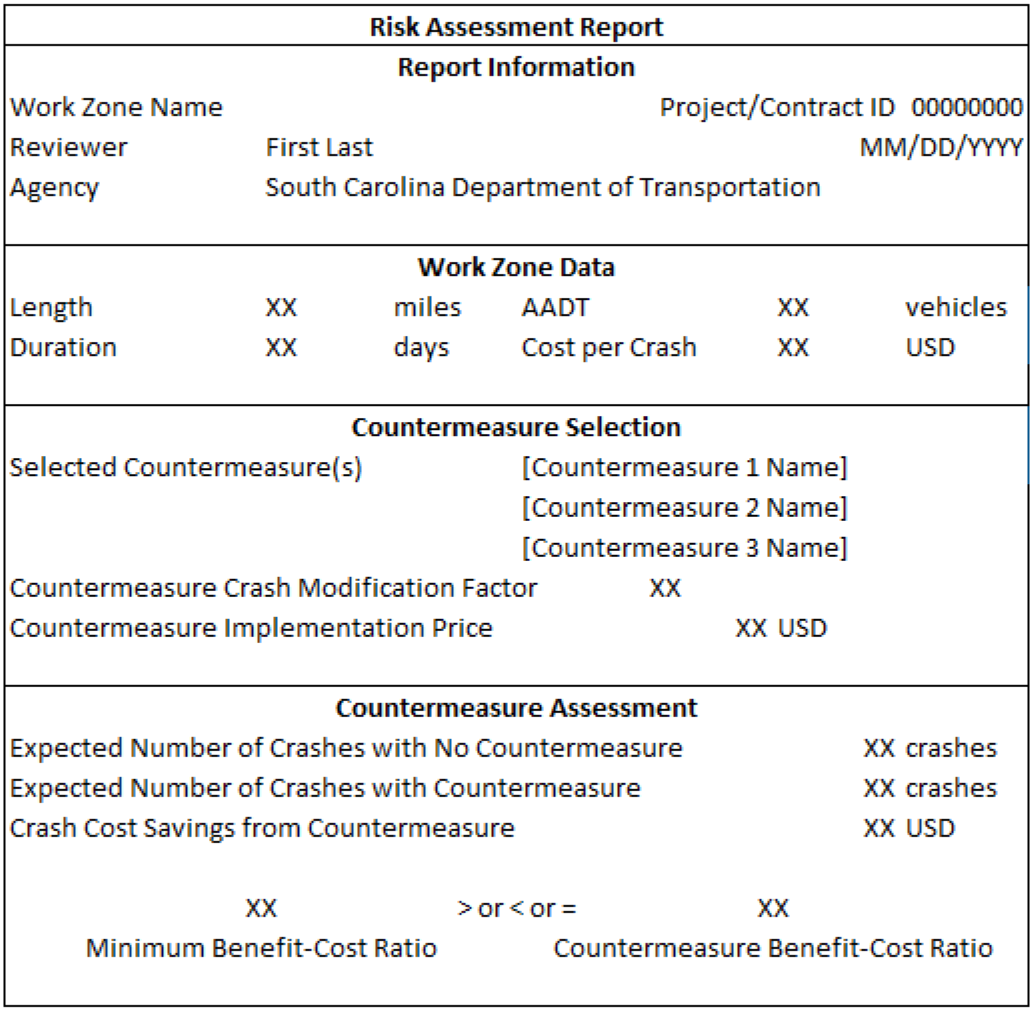
It is recommended that the SCDOT consider improving lighting conditions at work zones at night, as dark lighting conditions were identified as a significant contributing factor to injury in truck-involved crashes. Given that driving too fast for roadway conditions was identified as a significant factor contributing to injury, the SCDOT should consider prioritizing speed management measures in work zones; it should be noted that in practice, officers may use "driving too fast for conditions" as a contributing factor when they are unable to pinpoint the exact reason such as distracted driving. This includes implementing speed enforcement strategies, enhancing signage, and utilizing traffic calming measures to encourage drivers to adhere to posted speed limits and adjust their speed according to roadway conditions. Lastly, educational campaigns to improve safety such as promoting the use of seat belts and avoiding distracted driving should target both male and female drivers.

Several countermeasures are recommended to address factors contributing to injury in work zone rear-end crashes with collision speeds of 35 mph or higher. When promoting work zone safety during the National Work Zone Awareness Week held each spring in South Carolina, the SCDOT should consider getting the message to older drivers. The use of lighting and advanced warning systems in work zones improves visibility and safety, particularly during nighttime operations. Thus, these countermeasures should be considered when applicable. Other countermeasures to consider include educating both the trucking industry and the public about the danger of truck-involved crashes in work zones, improving traffic flow and reducing congestion in the proximity of the first work zone sign.

This study found the presence of law enforcement to be effective in reducing traffic speed through the work zones. The SCDOT could consider expanding the Safety Improvement Team Program in partnership with the South Carolina Department of Public Safety (SCDPS). Specific strategies include having troopers stationed near the first work zone sign and in the activity area. To reduce cost, SCDOT and SCDPS could consider nuanced approaches to variable time spent at the work zones without sacrificing effectiveness. Additionally, it may be helpful for the two agencies to evaluate the levels of active enforcement (e.g., the number of citations issued by troopers) on compliance with the work zone posted speed limit.

5.3. Implementation Plan

The work zone risk assessment tool provided to the SCDOT in this study can be utilized to assess the crash risk of a work zone and the benefit-cost of implementing countermeasures to mitigate those risks. The SCDOT should consider distributing this tool to SCDOT engineers or contractors designing work zone traffic control plans. This tool was developed in Microsoft Excel using VBA and the estimation results of the zero-inflated negative binomial model discussed in Section 4.5. The tool returns a predicted number of crashes given the work zone type, length in miles, duration in days, and AADT. Additionally, it determines the benefit-cost ratio following the procedure explained in Section 4.2. The input required from the user are comtemplated countermeasure(s), cost to implement the countermeasure(s), average cost of a crash, and the minimum benefit-cost ratio to justify implementing the countermeasure(s). Figure 15 shows the report generated by the tool.



Risk Assessment Report

Report Information

Work Zone Name		Project/Contract ID	00000000
Reviewer	First Last		MM/DD/YYYY
Agency	South Carolina Department of Transportation		

Work Zone Data

Length	XX	miles	AADT	XX	vehicles
Duration	XX	days	Cost per Crash	XX	USD

Countermeasure Selection

Selected Countermeasure(s)	[Countermeasure 1 Name] [Countermeasure 2 Name] [Countermeasure 3 Name]
Countermeasure Crash Modification Factor	XX
Countermeasure Implementation Price	XX USD

Countermeasure Assessment

Expected Number of Crashes with No Countermeasure	XX crashes
Expected Number of Crashes with Countermeasure	XX crashes
Crash Cost Savings from Countermeasure	XX USD

XX	> or < or =	XX
Minimum Benefit-Cost Ratio		Countermeasure Benefit-Cost Ratio

Figure 15. Report generated by Work Zone Risk Assessment Tool.

References

Al-Bdairi, N.S.S., 2020. Does time of day matter at highway work zone crashes? J. Safety Res. 73, 47–56. https://doi.org/10.1016/j.jsr.2020.02.013

Amoros, E., Martin, J.-L., Chiron, M., Laumon, B., 2007. Road Crash Casualties: Characteristics of Police Injury Severity Misclassification. J. Trauma Acute Care Surg. 62, 482. https://doi.org/10.1097/01.ta.0000202546.49273.f9

Anastasopoulos, P.Ch., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. Accid. Anal. Prev. 43, 1140–1147. https://doi.org/10.1016/j.aap.2010.12.024

Anderson, J., Hernandez, S., 2017. Roadway classifications and the accident injury severities of heavy-vehicle drivers. Anal. Methods Accid. Res. 15, 17–28. https://doi.org/10.1016/j.amar.2017.04.002

Arditi, D., Lee, D.-E., Polat, G., 2007. Fatal accidents in nighttime vs. daytime highway construction work zones. J. Safety Res. 38, 399–405. https://doi.org/10.1016/j.jsr.2007.04.001

Ashqar, H.I., Shaheen, Q.H.Q., Ashur, S.A., Rakha, H.A., 2021. Impact of risk factors on work zone crashes using logistic models and Random Forest, in: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). Presented at the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp. 1815–1820. https://doi.org/10.1109/ITSC48978.2021.9564405

Azimi, M., Oyelade, I., Aremu, A.M., Balal, E., Cheu, R.L., Qi, Y., 2021. Selection and Implementation of Intelligent Transportation Systems for Work Zone Construction Projects. Future Transp. 1, 169–187. https://doi.org/10.3390/futuretransp1020011

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychol. Bull. 107, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Boonyoo, T., Champahom, T., 2022. ANALYSIS OF FACTORS AFFECTING REAR-END CRASH SEVERITY USING STRUCTURAL EQUATION MODELING 29.

Champahom, T., Jomnonkwao, S., Karoonsoontawong, A., Hantanong, N., Beeharry, R., Ratanavaraha, V., 2020. Modeling user perception of bus service quality: A case study in Mauritius.

Chen, F., Song, M., Ma, X., 2019. Investigation on the Injury Severity of Drivers in Rear-End Collisions Between Cars Using a Random Parameters Bivariate Ordered Probit Model. Int. J. Environ. Res. Public. Health 16, 2632. https://doi.org/10.3390/ijerph16142632

Craik, F.I.M., Lockhart, R.S., 1972. Levels of processing: A framework for memory research. J. Verbal Learn. Verbal Behav. 11, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Daniel, J., Dixon, K., Jared, D., 2000. Analysis of Fatal Crashes in Georgia Work Zones. Transp. Res. Rec. 1715, 18–23. https://doi.org/10.3141/1715-03

Debnath, A.K., Blackman, R., Haworth, N., 2015. Common hazards and their mitigating measures in work zones: A qualitative study of worker perceptions. Saf. Sci. 72, 293–301. https://doi.org/10.1016/j.ssci.2014.09.022

Department of Transportation & Infrastructure Studies, Morgan State University, Baltimore, MD, USA, Banerjee, S., Jeihani, M., Moghaddam, Z.R., 2018. Impact of Mobile Work Zone Barriers on Driving Behavior on Arterial Roads. J. Traffic Logist. Eng. 37–42. https://doi.org/10.18178/jtle.6.2.37-42

Development of Work Zone Crash Modification Factors (CMFs) [WWW Document], n.d. . Work Zone Saf. Inf. Clgh. URL https://workzonesafety.org/training/development-of-work-zone-crash-modification-factors-cmfs/ (accessed 6.10.24).

Dias, I., Dissanayake, S., 2016. Comparison of Factors Affecting Work Zone Crash Severity Between Nighttime and Daytime. Presented at the Transportation Research Board 95th Annual MeetingTransportation Research Board.

Dias, I.M., n.d. Work zone crash analysis and modeling to identify the factors affecting crash severity and frequency (Ph.D.). Kansas State University, United States -- Kansas.

Dong, X., Xie, K., Yang, H., 2022. How did COVID-19 impact driving behaviors and crash Severity? A multigroup structural equation modeling. Accid. Anal. Prev. 172, 106687. https://doi.org/10.1016/j.aap.2022.106687

Eberly, L.E., 2007. Multiple Linear Regression, in: Ambrosius, W.T. (Ed.), Topics in Biostatistics. Humana Press, Totowa, NJ, pp. 165–187. https://doi.org/10.1007/978-1-59745-530-5_9

Favero, L.P., Belfiore, P., Souza, R. de F., 2023. Data Science, Analytics and Machine Learning with R. Academic Press.

Fotios, S., Robbins, C., 2024. Incorrect categorisation of ambient light level at the time of a road traffic collision. Light. Res. Technol. 56, 87–101. https://doi.org/10.1177/14771535211069028

Garber, N.J., Zhao, M., 2002. Distribution and Characteristics of Crashes at Different Work Zone Locations in Virginia. Transp. Res. Rec. 1794, 19–25. https://doi.org/10.3141/1794-03

Ghasemzadeh, A., Ahmed, M.M., 2019. Complementary parametric probit regression and nonparametric classification tree modeling approaches to analyze factors affecting severity of work zone weather-related crashes. J. Mod. Transp. 27, 129–140. https://doi.org/10.1007/s40534-018-0178-6

Gupta, R., Asgari, H., Azimi, G., Rahimi, A., Jin, X., 2021. Analysis of Fatal Truck-Involved Work Zone Crashes in Florida: Application of Tree-Based Models. Transp. Res. Rec. 2675, 1272–1290. https://doi.org/10.1177/03611981211033278

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L., 2006. Multivariate data analysis 6th Edition.

Hamdar, S.H., Schorr, J., 2013. Interrupted versus uninterrupted flow: A safety propensity index for driver behavior. Accid. Anal. Prev. 55, 22–33. https://doi.org/10.1016/j.aap.2013.01.017

Hassan, H.M., Abdel-Aty, M.A., 2013. Exploring the safety implications of young drivers' behavior, attitudes and perceptions. Accid. Anal. Prev. 50, 361–370. https://doi.org/10.1016/j.aap.2012.05.003

Hausman, J.A., Abrevaya, J., Scott-Morton, F.M., 1998. Misclassification of the dependent variable in a discrete-response setting. J. Econom. 87, 239–269. https://doi.org/10.1016/S0304-4076(98)00015-3

Hosseini, P., Jalayer, M., Das, S., 2021. A Multiple Correspondence Approach to Identify Contributing Factors Related to Work Zone Crashes. Presented at the Transportation Research Board 100th Annual MeetingTransportation Research BoardTransportation Research Board.

Hou, G., Chen, S., 2020. Study of work zone traffic safety under adverse driving conditions with a microscopic traffic simulation approach. Accid. Anal. Prev. 145, 105698. https://doi.org/10.1016/j.aap.2020.105698

Islam, M., 2022. An analysis of motorcyclists' injury severities in work-zone crashes with unobserved heterogeneity. IATSS Res. 46, 281–289. https://doi.org/10.1016/j.iatssr.2022.01.003

Islam, M., Alnawmasi, N., Mannering, F., 2020. Unobserved heterogeneity and temporal instability in the analysis of work-zone crash-injury severities. Anal. Methods Accid. Res. 28, 100130. https://doi.org/10.1016/j.amar.2020.100130

Jobson, J.D., 2012. Applied Multivariate Data Analysis: Regression and Experimental Design. Springer Science & Business Media.

Jurewicz, C., Sobhani, A., Woolley, J., Dutschke, J., Corben, B., 2016. Exploration of Vehicle Impact Speed – Injury Severity Relationships for Application in Safer Road Design. Transp. Res. Procedia, Transport Research Arena TRA2016 14, 4247–4256. https://doi.org/10.1016/j.trpro.2016.05.396

Khattak, A.J., Targa, F., 2004. Injury Severity and Total Harm in Truck-Involved Work Zone Crashes. Transp. Res. Rec. 1877, 106–116. https://doi.org/10.3141/1877-12

Khattak, Asad J, Khattak, Aemal J, Council, F.M., 2002. Effects of work zone presence on injury and non-injury crashes. Accid. Anal. Prev. 34, 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3

Koilada, K., Mane, A.S., Pulugurtha, S.S., 2020. Odds of work zone crash occurrence and getting involved in advance warning, transition, and activity areas by injury severity. IATSS Res. 44, 75–83. https://doi.org/10.1016/j.iatssr.2019.07.003

Li, Y., Bai, Y., 2009. Effectiveness of temporary traffic control measures in highway work zones. Saf. Sci. 47, 453–458. https://doi.org/10.1016/j.ssci.2008.06.006

Li, Y., Bai, Y., 2008a. Development of crash-severity-index models for the measurement of work zone risk levels. Accid. Anal. Prev. 40, 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Li, Y., Bai, Y., 2008b. Development of crash-severity-index models for the measurement of work zone risk levels. Accid. Anal. Prev. 40, 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Liu, J., Khattak, A., Zhang, M., 2016. What Role Do Precrash Driver Actions Play in Work Zone Crashes?:Application of Hierarchical Models to Crash Data. Transp. Res. Rec. 2555, 1–11. https://doi.org/10.3141/2555-01

Madarshahian, M., Balaram, A., Ahmed, F., Huynh, N., Siddiqui, C.K., Ferguson, M., 2023. Analysis of injury severity of work zone truck-involved crashes in South Carolina for interstates and non-interstates. Sustainability 15, 7188.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13. https://doi.org/10.1016/j.amar.2017.10.002

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16. https://doi.org/10.1016/j.amar.2016.04.001

Meng, Q., Weng, J., 2011. Evaluation of rear-end crash risk at work zone using work zone traffic data. Accid. Anal. Prev. 43, 1291–1300. https://doi.org/10.1016/j.aap.2011.01.011

Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis. Transp. Res. Rec. 2673, 680–692. https://doi.org/10.1177/0361198119845899

Mokhtarimousavi, S., Anderson, J.C., Hadi, M., Azizinamini, A., 2021. A temporal investigation of crash severity factors in worker-involved work zone crashes: Random parameters and machine learning approaches. Transp. Res. Interdiscip. Perspect. 10, 100378. https://doi.org/10.1016/j.trip.2021.100378

Mw, B., 1993. Alternative ways of assessing model fit. Test. Struct. Equ. Models.

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accid. Anal. Prev. 28, 739–753. https://doi.org/10.1016/S0001-4575(96)00050-4

Osborne, J., Waters, E., 2019. Four assumptions of multiple regression that researchers should always test. Pract. Assess. Res. Eval. 8. https://doi.org/10.7275/r222-hv23

Osman, M., Mishra, S., Paleti, R., Golias, M., 2019. Impacts of Work Zone Component Areas on Driver Injury Severity. J. Transp. Eng. Part Syst. 145, 04019032. https://doi.org/10.1061/JTEPBS.0000253

Osman, M., Paleti, R., Mishra, S., 2018. Analysis of passenger-car crash injury severity in different work zone configurations. Accid. Anal. Prev. 111, 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., Mishra, S., Golias, M.M., 2016. Analysis of injury severity of large truck crashes in work zones. Accid. Anal. Prev. 97, 261–273. https://doi.org/10.1016/j.aap.2016.10.020

Rista, E., Barrette, T., Hamzeie, R., Savolainen, P., Gates, T.J., 2017. Work Zone Safety Performance: Comparison of Alternative Traffic Control Strategies. Transp. Res. Rec. 2617, 87–93. https://doi.org/10.3141/2617-11

Santos, B., Trindade, V., Polónia, C., Picado-Santos, L., 2021. Detecting Risk Factors of Road Work Zone Crashes from the Information Provided in Police Crash Reports: The Case Study of Portugal. Safety 7, 12. https://doi.org/10.3390/safety7010012

Sayed, M.A., Qin, X., Kate, R.J., Anisuzzaman, D.M., Yu, Z., 2021. Identification and analysis of misclassified work-zone crashes using text mining techniques. Accid. Anal. Prev. 159, 106211. https://doi.org/10.1016/j.aap.2021.106211

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: A random parameters approach with heterogeneity in means and variances. Anal. Methods Accid. Res. 15, 41–55. https://doi.org/10.1016/j.amar.2017.05.003

Shi, J., Bai, Y., Tao, L., Atchley, P., 2011. A model of Beijing drivers' scrambling behaviors. Accid. Anal. Prev. 43, 1540–1546. https://doi.org/10.1016/j.aap.2011.03.008

Shinar, D., Treat, J.R., McDonald, S.T., 1983. The validity of police reported accident data. Accid. Anal. Prev. 15, 175–191. https://doi.org/10.1016/0001-4575(83)90018-0

Significance tests and goodness of fit in the analysis of covariance structures - ProQuest [WWW Document], n.d. URL https://www.proquest.com/openview/ad7ec3686dd1bbe533b628ecccaf741b/1?cbl=60977&pq-origsite=gscholar&parentSessionId=6%2BtLlr3XzCiBDPz%2Fi69Kyplpzn0CY2h%2BTlHHRNaSn8Y%3D (accessed 6.10.24).

Statistical and Econometric Methods for Transportation Data Analysis | [WWW Document], n.d. URL https://www.taylorfrancis.com/books/mono/10.1201/9780429244018/statistical-econometric-methods-transportation-data-analysis-simon-washington-fred-mannering-panagiotis-anastasopoulos-matthew-karlaftis (accessed 6.10.24).

Steiger, J.H., 2007. Understanding the limitations of global fit assessment in structural equation modeling. Personal. Individ. Differ., Special issue on Structural Equation Modeling 42, 893–898. https://doi.org/10.1016/j.paid.2006.09.017

Stroup, W.W., Milliken, G.A., Claassen, E.A., Wolfinger, R.D., 2018. SAS for Mixed Models: Introduction and Basic Applications. SAS Institute.

Sze, N.N., Song, Z., 2019. Factors contributing to injury severity in work zone related crashes in New Zealand. Int. J. Sustain. Transp. 13, 148–154. https://doi.org/10.1080/15568318.2018.1452083

Thapa, D., Mishra, S., 2021. Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. Accid. Anal. Prev. 156, 106125. https://doi.org/10.1016/j.aap.2021.106125

Theofilatos, A., Ziakopoulos, A., Papadimitriou, E., Yannis, G., Diamandouros, K., 2017. Meta-analysis of the effect of road work zones on crash occurrence. Accid. Anal. Prev. 108, 1–8. https://doi.org/10.1016/j.aap.2017.07.024

Uddin, M., Huynh, N., 2020. Injury severity analysis of truck-involved crashes under different weather conditions. Accid. Anal. Prev. 141, 105529. https://doi.org/10.1016/j.aap.2020.105529

Uddin, M., Huynh, N., 2017. Truck-involved crashes injury severity analysis for different lighting conditions on rural and urban roadways. Accid. Anal. Prev. 108, 44–55. https://doi.org/10.1016/j.aap.2017.08.009

Ullman, G.L., Scriba, T.A., 2004. Revisiting the Influence of Crash Report Forms on Work Zone Crash Data. Transp. Res. Rec. 1897, 180–182. https://doi.org/10.3141/1897-23

Uyanık, G.K., Güler, N., 2013. A Study on Multiple Linear Regression Analysis. Procedia - Soc. Behav. Sci., 4th International Conference on New Horizons in Education 106, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Wang, K., Qin, X., 2014. Use of Structural Equation Modeling to Measure Severity of Single-Vehicle Crashes. Transp. Res. Rec. 2432, 17–25. https://doi.org/10.3141/2432-03

Wei, X., Shu, X., Huang, B., Taylor, E.L., Chen, H., 2017. Analyzing Traffic Crash Severity in Work Zones under Different Light Conditions. J. Adv. Transp. 2017, 5783696. https://doi.org/10.1155/2017/5783696

Weng, J., Du, G., Li, D., Yu, Y., 2018. Time-varying mixed logit model for vehicle merging behavior in work zone merging areas. Accid. Anal. Prev. 117, 328–339. https://doi.org/10.1016/j.aap.2018.05.005

Weng, J., Meng, Q., 2011. Analysis of driver casualty risk for different work zone types. Accid. Anal. Prev. 43, 1811–1817. https://doi.org/10.1016/j.aap.2011.04.016

Weng, J., Meng, Q., Yan, X., 2014. Analysis of work zone rear-end crash risk for different vehicle-following patterns. Accid. Anal. Prev. 72, 449–457. https://doi.org/10.1016/j.aap.2014.08.003

Weng, J., Xue, S., Yang, Y., Yan, X., Qu, X., 2015. In-depth analysis of drivers' merging behavior and rear-end crash risks in work zone merging areas. Accid. Anal. Prev. 77, 51–61. https://doi.org/10.1016/j.aap.2015.02.002

Weng, J., Zhu, J.-Z., Yan, X., Liu, Z., 2016. Investigation of work zone crash casualty patterns using association rules. Accid. Anal. Prev. 92, 43–52. https://doi.org/10.1016/j.aap.2016.03.017

Williams, M., Grajales, C., Kurkiewicz, D., 2019. Assumptions of Multiple Regression: Correcting Two Misconceptions. Pract. Assess. Res. Eval. 18. https://doi.org/10.7275/55hn-wk47

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., Jiang, X., 2020. A machine-learning method for improving crash injury severity analysis: a case study of work zone crashes in Cairo, Egypt. Int. J. Inj. Contr. Saf. Promot. 27, 266–275. https://doi.org/10.1080/17457300.2020.1746814

Yang, H., Ozbay, K., Ozturk, O., Xie, K., 2015. Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. Traffic Inj. Prev. 16, 387–396. https://doi.org/10.1080/15389588.2014.948615

Yang, H., Ozbay, K., Ozturk, O., Yildirimoglu, M., 2013. Modeling work zone crash frequency by quantifying measurement errors in work zone length. Accid. Anal. Prev. 55, 192–201. https://doi.org/10.1016/j.aap.2013.02.031

Yu, C.-Y., n.d. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Ph.D.). University of California, Los Angeles, United States -- California.

Yu, M., Zheng, C., Ma, C., 2020. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. Anal. Methods Accid. Res. 27, 100126. https://doi.org/10.1016/j.amar.2020.100126

Zhang, K., Hassan, M., 2019a. Identifying the Factors Contributing to Injury Severity in Work Zone Rear-End Crashes. J. Adv. Transp. 2019, 4126102. https://doi.org/10.1155/2019/4126102

Zhang, K., Hassan, M., 2019b. Crash severity analysis of nighttime and daytime highway work zone crashes. PLOS ONE 14, e0221128. https://doi.org/10.1371/journal.pone.0221128

Zhang, K., Hassan, M., Yahaya, M., Yang, S., 2018. Analysis of Work-Zone Crashes Using the Ordered Probit Model with Factor Analysis in Egypt. J. Adv. Transp. 2018, 8570207. https://doi.org/10.1155/2018/8570207

Appendix A

```
library(readr)

library(dplyr)

library(MASS)

library(pscl)

#install.packages("pscl")

*********************************************************************************************

# Whole Data

Data <- read.csv("Path\\Neg_Bin_Dat.csv")

*********************************************************************************************

# Subset data for "Widening"

Data_Widening <- Data[Data$TYPE == "Widening", ]

# Fit zero-inflated negative binomial model

zeroinfl_model <- zeroinfl(CRASHES ~ LENGTH + DURATION + log(AADT) | LENGTH + DURATION + log(AADT), data = Data_Widening, dist = "negbin", maxit = 500)

summary(zeroinfl_model)

predicted_crashes <- predict(zeroinfl_model, newdata = Data_Widening, type = "response")

observed_counts <- Data_Widening$CRASHES

result_data <- data.frame(Observed = observed_counts, Predicted = round(predicted_crashes,0))

head(result_data)

*********************************************************************************************

# Rehabilitation dataset

Data_Rehabilitation <- Data[Data$TYPE == "Rehabilitation", ]

inflated_all_Rehabilitation <- zeroinfl(CRASHES ~ LENGTH + DURATION + log(AADT) | LENGTH + DURATION + log(AADT), data = Data_Rehabilitation,link = "logit", dist = "negbin")

summary(inflated_all_Rehabilitation)
```

```
round(inflated_all_Rehabilitation$fitted.values,0)

predicted_crashes <- predict(inflated_all_Rehabilitation, newdata = Data_Rehabilitation, type =
"response")

observed_counts <- Data_Rehabilitation$CRASHES

result_data <- data.frame(Observed = observed_counts, Predicted = round(predicted_crashes,0))

head(result_data)

*************************************************************************************

# Subset data for "Preservation"

Data_Preservation <- Data[Data$TYPE == "Preservation", ]

# Fit zero-inflated negative binomial model

zeroinfl_model_Preservation <- zeroinfl(CRASHES ~ LENGTH + DURATION + log(AADT)| LENGTH +
DURATION + log(AADT), data = Data_Preservation, dist = "negbin", maxit = 500)

# Display summary statistics

summary(zeroinfl_model_Preservation)

predicted_crashes <- predict(zeroinfl_model_Preservation, newdata = Data_Preservation, type =
"response")

observed_counts <- Data_Preservation$CRASHES

result_data <- data.frame(Observed = observed_counts, Predicted = round(predicted_crashes,0))

head(result_data)

*************************************************************************************

Data_Reconstruction <- Data[Data$TYPE == "Reconstruction", ]

# Fit zero-inflated negative binomial model

zeroinfl_model_Reconstruction <- zeroinfl(CRASHES ~ LENGTH + DURATION + log(AADT) |LENGTH +
DURATION + log(AADT), data = Data_Reconstruction, dist = "negbin", maxit = 1000)

# Display summary statistics

summary(zeroinfl_model_Reconstruction)
```

```
predicted_crashes <- predict(zeroinfl_model_Reconstruction, newdata = Data_Reconstruction, type = "response")

observed_counts <- Data_Reconstruction$CRASHES

result_data <- data.frame(Observed = observed_counts, Predicted = round(predicted_crashes,0))

head(result_data)
```