



February 2014

FEDERAL MOTOR CARRIER SAFETY

Modifying the
Compliance, Safety,
Accountability
Program Would
Improve the Ability to
Identify High Risk
Carriers

Why GAO Did This Study

From 2009 to 2012, large commercial trucks and buses have averaged about 125,000 crashes per year, with about 78,000 injuries and over 4,100 fatalities. In 2010, FMCSA replaced its tool for identifying the riskiest carriers—SafeStat—with the CSA program. CSA is intended to reduce the number of motor carrier crashes by better targeting the highest risk carriers using information from roadside inspections and crash investigations. CSA includes SMS, a data-driven approach for identifying motor carriers at risk of causing a crash.

GAO was directed by the Consolidated Appropriations Act of 2012 to monitor the implementation of CSA. This report examines the effectiveness of the CSA program in assessing safety risk for motor carriers. GAO spoke with FMCSA officials and stakeholders to understand SMS. Using FMCSA's data, GAO replicated FMCSA's method for calculating SMS scores and assessed the effect of changes—such as stronger data-sufficiency standards—on the scores. GAO also evaluated SMS's ability to predict crashes.

What GAO Recommends

GAO recommends that FMCSA revise the SMS methodology to better account for limitations in drawing comparisons of safety performance information across carriers. In addition, determination of a carrier's fitness to operate should account for limitations in available performance information. In response to comments from the Department of Transportation (USDOT), GAO clarified one of the recommendations. USDOT agreed to consider the recommendations.

View [GAO-14-114](#). For more information, contact Susan Fleming at (202) 512-2834 or flemings@gao.gov.

FEDERAL MOTOR CARRIER SAFETY

Modifying the Compliance, Safety, Accountability Program Would Improve the Ability to Identify High Risk Carriers

What GAO Found

The Federal Motor Carrier Safety Administration's (FMCSA) Compliance, Safety, Accountability (CSA) program has helped the agency contact or investigate more motor carrier companies that own commercial trucks and buses and has provided a range of safety benefits to safety officials, law enforcement, and the industry than the previous approach, SafeStat. Specifically, from fiscal year 2007 to fiscal year 2012, FMCSA more than doubled its number of annual interventions, largely by sending warning letters to riskier carriers.

A key component of CSA—the Safety Measurement System (SMS)—uses carrier performance data collected from roadside inspections or crash investigations to identify high risk carriers for intervention by analyzing relative safety scores in various categories, including Unsafe Driving and Vehicle Maintenance. FMCSA faces at least two challenges in reliably assessing safety risk for the majority of carriers. First, for SMS to be effective in identifying carriers more likely to crash, the violations that FMCSA uses to calculate SMS scores should have a strong predictive relationship with crashes. However, based on GAO's analysis of available information, most regulations used to calculate SMS scores are not violated often enough to strongly associate them with crash risk for individual carriers. Second, most carriers lack sufficient safety performance data to ensure that FMCSA can reliably compare them with other carriers. To produce an SMS score, FMCSA calculates violation rates for each carrier and then compares these rates to other carriers. Most carriers operate few vehicles and are inspected infrequently, providing insufficient information to produce reliable SMS scores. FMCSA acknowledges that violation rates are less precise for carriers with little information, but its methods do not fully address this limitation. For example, FMCSA requires a minimum level of information for a carrier to receive an SMS score; however, this requirement is not strong enough to produce sufficiently reliable scores. As a result, GAO found that FMCSA identified many carriers as high risk that were not later involved in a crash, potentially causing FMCSA to miss opportunities to intervene with carriers that were involved in crashes.

FMCSA's methodology is limited because of insufficient information, which reduces the precision of SMS scores. GAO found that by scoring only carriers with more information, FMCSA could better identify high risk carriers likely to be involved in crashes. This illustrative approach involves trade-offs; it would assign SMS scores to fewer carriers, but these scores would generally be more reliable and thus more useful in targeting FMCSA's scarce resources.

In addition to using SMS scores to prioritize carriers for intervention, FMCSA reports these scores publicly and is considering using a carrier's performance information to determine its fitness to operate. Given the limitations with safety performance information, determining the appropriate amount of information needed to assess a carrier requires consideration of how reliable and precise the scores need to be for the purposes for which they are used. Ultimately, the mission of FMCSA is to reduce crashes, injuries, and fatalities. GAO continues to believe a data-driven, risk-based approach holds promise; however, revising the SMS methodology would help FMCSA better focus intervention resources where they can have the greatest impact on achieving this goal.

Contents

Letter		1
	Background	5
	CSA Program Increases Carrier Interventions, but FMCSA Faces Challenges in Identifying High Risk Carriers	13
	Conclusions	31
	Recommendations for Executive Action	31
	Agency Comments	32
Appendix I	Scope and Methodology	34
Appendix II	Estimating Rates of Regulatory Violations in the Safety Measurement System	39
Appendix III	Evaluating the Statistical Validity of the Safety Measurement System	46
Appendix IV	Prior Evaluations of SMS Scores as Measures of Safety for Specific Carriers and Risk Groups	52
Appendix V	Analysis of Regulatory Violations and Crash Risk	58
Appendix VI	Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis	73
Appendix VII	GAO Contact and Staff Acknowledgments	103
Tables		
	Table 1: FMCSA's Carrier Safety Measurement System Categories	7
	Table 2: CSA Interventions Conducted in Fiscal Year 2012	11

Table 3: Number of FMCSA Interventions, Fiscal Years 2007 to 2012/14	
Table 4: FMCSA's Existing Method of Identifying High Risk Carriers Compared with an Illustrative Alternative	25
Table 5: Crash Rates per 100 Vehicles for Carriers with an SMS Score above and below FMCSA's Intervention Thresholds Using FMCSA's Method and Illustrative Alternative	27
Table 6: Comparison of FMCSA's Method and Illustrative Alternative to Identify Carriers with an SMS Score in at Least One BASIC	28
Table 7: Model Groups Based on Crash Status Measure, Violation Rate Measure, and Carrier Size Restrictions	60
Table 8: A list of Sub-Model Descriptions according to Data Restrictions (Restricted to Data for Carriers with Greater Than 20 Vehicles versus Full Data with All Carriers), Violation Rates (Observed versus Bayesian), and Sample (Model Building versus Validation)	61
Table 9: Logistic Regression Results for Sub-Models Simple, Stepwise, and Full-of-Outcome Crash Status (Yes/No); Note That the Simple Model Is Redundant for Model Groups 2 and 4 Since No Violation Rates Are Included in the Simple Model	63
Table 10: Classification of Predicted Values from Models for the Crash-Status (Yes/No) Using the Average Observed Predicted Rate as the Cut-Point, Based on the Model-Building Sample	65
Table 11: Linear Regression Model Results for a Bayesian Crash-Rate Model, Using the Model Developed for the Crash Status (Yes/No) Outcome, Estimated with the Model-Building Sample	66
Table 12: Numbers of Models for which Violations Were Significant and Stable Predictors, for Violations That Were Significant in 5 or More Models	68
Table 13: Fit Statistics Based on the Validation Sample, for Crash Status (Yes/No)	70
Table 14: Distribution of Crashes, Power Units, Inspections, and High Risk Status by Carrier Size (GAO Analysis Population)	74
Table 15: Comparison of Crash Involvement for Carriers above and below Intervention Threshold Using FMCSA's Methodology (Compare to Illustrative Alternative Analysis in Following Table)	75

Table 16: Comparison of Crash Involvement for Carriers above and below Intervention Threshold using Illustrative Alternative (Compare to FMCSA's Methodology in Previous Table)	78
Table 17: SMS Outcomes as Reported by FMCSA Compared to Outcomes from GAO Analysis	80

Figures

Figure 1: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Hours-of-Service Compliance BASIC	18
Figure 2: Average and Range (between the 1st and 99th Percentiles) of Violation Rates for Carriers in the Unsafe Driving BASIC	19
Figure 3: Percentage of FMCSA Scored Carriers in the Hours-of-Service BASIC above the Intervention Threshold by Number of Inspections	22
Figure 4: Distribution of FMCSA Scored Carriers above the Unsafe Driving BASIC Threshold by Carrier Size	23
Figure 5: Percentage of Carriers Identified as above FMCSA's Intervention Threshold, or High Risk, That Crashed during the Evaluation Period, Comparing FMCSA's Existing Method and Illustrative Alternative	26
Figure 6: Example of the Relationship between Exposure and the Precision of Rate Estimates	42
Figure 7: Relationships between Exposure and Rate Estimates for a Population of Motor Carriers Active from December 2007 through June 2011	44
Figure 8: Examples of Empirical Bayes Rate Estimates for a Sample of Carriers Active from December 2007 through June 2011	45
Figure 9: SMS as a Measurement Model	49
Figure 10: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Unsafe Driving BASIC	81
Figure 11: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Hours-of-Service Compliance BASIC	82
Figure 12: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Driver Fitness BASIC	83

Figure 13: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Controlled Substances and Alcohol BASIC	84
Figure 14: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Vehicle Maintenance BASIC	85
Figure 15: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Hazardous Materials BASIC	86
Figure 16: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Crash Indicator BASIC	87
Figure 17: Percentage of FMCSA Scored Carriers in the Unsafe Driving (Straight Segment) BASIC above the Intervention Threshold by Number of Inspections	88
Figure 18: Percentage of FMCSA Scored Carriers in the Unsafe Driving (Combo Segment) BASIC above the Intervention Threshold by Number of Inspections	89
Figure 19: Percentage of FMCSA Scored Carriers in the Hours-of-Service Compliance BASIC above the Intervention Threshold by Number of Inspections	90
Figure 20: Percentage of FMCSA Scored Carriers in the Driver Fitness BASIC above the Intervention Threshold by Number of Inspections	91
Figure 21: Percentage of FMCSA-Scored Carriers in the Controlled Substances and Alcohol BASIC above the Intervention Threshold by Number of Inspections	92
Figure 22: Percentage of FMCSA-Scored Carriers in the Vehicle Maintenance BASIC above the Intervention Threshold by Number of Inspections	93
Figure 23: Percentage of FMCSA-Scored Carriers in the Hazardous Materials BASIC above the Intervention Threshold by Number of Inspections	94
Figure 24: Percentage of FMCSA-Scored Carriers on the Crash Indicator (Straight Segment) above the Intervention Threshold by Number of Inspections	95
Figure 25: Percentage of FMCSA-Scored Carriers on the Crash Indicator (Combo Segment) above the Intervention Threshold by Number of Inspections	96
Figure 26: Distribution of FMCSA-Scored Carriers above the Unsafe Driving BASIC Threshold by Carrier Size	97

Figure 27: Distribution of FMCSA-Scored Carriers above the Hours-of-Service Compliance BASIC Threshold by Carrier Size	97
Figure 28: Distribution of FMCSA-Scored Carriers above the Driver Fitness BASIC Threshold by Carrier Size	98
Figure 29: Distribution of FMCSA-Scored Carriers above the Controlled Substance and Alcohol BASIC Threshold by Carrier Size	99
Figure 30: Distribution of FMCSA-Scored Carriers above the Vehicle Maintenance BASIC Threshold by Carrier Size	100
Figure 31: Distribution of FMCSA-Scored Carriers above the Hazardous Materials BASIC Threshold by Carrier Size	101
Figure 32: Distribution of FMCSA-Scored Carriers above the Crash Indicator Threshold by Carrier Size	102

Abbreviations

ATRI	American Transportation Research Institute
BASIC	Behavioral Analysis and Safety Improvement Categories
CDC	Centers for Disease Control and Prevention
CMV	commercial motor vehicle
CSA	Compliance, Safety, Accountability
FMCSA	Federal Motor Carrier Safety Administration
MCMIS	Motor Carrier Management Information System
SMS	Safety Measurement System
UMTRI	University of Michigan Transportation Research Institute
USDOT	U.S. Department of Transportation
VMT	vehicle miles traveled

This is a work of the U.S. government and is not subject to copyright protection in the United States. The published product may be reproduced and distributed in its entirety without further permission from GAO. However, because this work may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.



February 3, 2014

Congressional Committees

Large commercial trucks and buses are vital for the movement of goods and people across America. According to the American Trucking Associations, the trucking industry moved 9.4 billion tons of freight in 2012, and according to the American Bus Association, the “motor-coach” industry provided about 694 million passenger trips in 2010. However, this activity comes with a cost. From 2009 to 2012, crashes involving large commercial trucks and buses averaged around 125,000 per year, resulting in about 78,000 injuries and about 4,100 fatalities.

The primary mission of the U.S. Department of Transportation’s (USDOT) Federal Motor Carrier Safety Administration (FMCSA) is to reduce crashes, injuries, and fatalities involving large trucks and buses. FMCSA partners with states to conduct roadside inspections and uses inspection or crash information to assess and prioritize the riskiest motor carriers for further intervention. From 1997 through 2010, FMCSA used a program known as SafeStat to track how well motor carriers—the companies that own commercial trucks and buses—complied with safety standards. Under SafeStat, FMCSA reviewed only a small percentage of the more than 500,000 motor carriers operating in the United States in a given year. In an attempt to increase the number of motor carriers that FMCSA can evaluate each year and, ultimately, to improve large commercial truck and bus safety, FMCSA began to develop the Compliance, Safety, Accountability (CSA) program in 2004.¹ One component of the CSA program is the Safety Measurement System (SMS), a data-driven approach for identifying motor carriers at risk of presenting a safety hazard or causing a crash. SMS uses information collected during roadside inspections and from reported crashes to calculate scores across seven categories that quantify a carrier’s safety performance relative to other carriers.

¹ FMCSA was required under section 4138 of the Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFTEA-LU) to “ensure that compliance reviews are completed on motor carriers that have demonstrated through performance data that they pose the highest safety risk.” Pub. L. No.109-59, § 4138, 119 Stat. 1144, 1745 (2005).

Since 2008, when CSA was first piloted, law enforcement and industry stakeholders have been generally supportive of FMCSA's overall CSA approach. Nonetheless, several evaluations of CSA conducted by a range of outside groups concluded that some SMS safety scores inaccurately assess a carrier's relative crash risk. The precision and accuracy of these scores is vital because FMCSA investigators and their state partners use SMS results to focus their resources to help reduce the number of motor carrier crashes, injuries, and fatalities. In addition, FMCSA currently posts most of the scores publicly on its website for use by industry stakeholders and the public² and has indicated that a future rulemaking will include similar information to help determine whether a carrier is fit to operate motor vehicles.³

We were directed in a Senate Appropriations Committee report to continue monitoring FMCSA's implementation of the CSA program.⁴ This report examines the effectiveness of the CSA program in assessing safety risk for motor carriers.

To examine the effectiveness of the CSA program, we obtained documentation and spoke with FMCSA officials about the CSA program. To examine the SMS methodology and scores, we collected carrier data from FMCSA's Motor Carrier Management Information System (MCMIS) and historical scores from SMS.⁵ We then replicated the methods FMCSA uses to calculate SMS scores (SMS Methodology 3.0) and assessed how changes to key steps and assumptions affected SMS scores and identification of the highest risk carriers. Given FMCSA's use of these scores as quantitative determinations of a carrier's safety performance, we assessed the reliability of SMS scores as defined by the precision, accuracy, and confidence of these scores when calculated for carriers

² See <http://ai.fmcsa.dot.gov/sms/>

³ 79 Fed. Reg. 896, 1038 (Jan. 7, 2014), Department of Transportation, Semiannual Regulatory Agenda (proposed rule anticipated May 2014).

⁴ This direction is contained in the Senate Appropriations Committee Report, S. Rep. No. 112-83, at 52, accompanying the Transportation, and Housing and Urban Development, and Related Agencies Appropriations Bill, 2012, which was eventually included in the Consolidated and Further Continuing Appropriations Act, 2012, Pub. L. No. 112-55, 125 Stat. 552 (2011).

⁵ FMCSA provided us historical carrier data for several time periods, including December 2008, December 2010, June 2012, and December 2012.

with varying levels of carrier exposure—measured by FMCSA as either inspections or an adjusted number of vehicles.⁶ We assessed changes in FMCSA’s requirements for carriers to receive SMS scores, changes in SMS score calculation, and adjustments to the scoring weights. We also evaluated the potential of FMCSA’s general approach to predict future crashes by using data on violations of FMCSA regulations and crashes to examine the relationships, if any, between violations of specific regulations and subsequent crashes. Due to ongoing litigation related to CSA and the publication of SMS scores, we did not assess the potential effects or tradeoffs resulting from the display or any public use of these scores.⁷

Our analysis included nearly 315,000 U.S.-based carriers that were under FMCSA’s jurisdiction and, with reasonable certainty, were active during the period from December 2007 through June 2011. We considered a carrier active during this period if it received a state or federal inspection, was involved in a crash, or reported the number of vehicles it operates to FMCSA. Information on inspections, violations, and crashes from December 2007 through December 2009, our observation period, was used to calculate SMS scores. We used crash information from the remaining 18 month period—from December 2009 through June 2011—referred to as our evaluation period, to determine these carriers’ subsequent crash rates and involvement in crashes.⁸ Carriers in our analysis population accounted for approximately 120,000 reported

⁶ GAO, *Assessing the Reliability of Computer-Processed Data*, [GAO-09-680G](#) (Washington, D.C.: July 2009).

⁷ See *Alliance for Safe, Efficient and Competitive Truck Transportation v. FMCSA*, No. 12-1305, D.C. Cir. (filed July 16, 2012; oral argument Sept. 10, 2013). The litigation has been brought against FMCSA by a number of motor carrier trade associations and challenges, among other things, the agency’s public disclosure of the SMS scores and its encouragement of the use of these public data to help make sound business judgments. The carriers have requested the court to order that the SMS scores not be publicly available until alleged flaws in the methodology are addressed in the context of the planned rulemaking. Under GAO’s policy to avoid addressing the merits of matters pending in litigation, we did not assess these matters.

⁸ On behalf of FMCSA, the Volpe Institute uses a “tool” for measuring the effectiveness of the SMS model, which consists of calculating rates of future crash involvement among groups of carriers found to have more or less safety risk. We chose this evaluation period to match the information and dates used by FMCSA to conduct its effectiveness test of changes made for SMS in the version 3.0 methodology. While the snapshot of carrier data GAO used for this analysis was dated December 2008 through June 2012, we were able to extract the relevant data for our specified time period.

crashes during this 18-month period. Throughout this report, our analysis is based on this population, during this time frame, unless otherwise specified.

To identify any modifications to FMCSA's method that could improve effectiveness, we compared the results from our changes to FMCSA's existing methodology and identified an illustrative combination of changes that better distinguished between carriers that later crashed and those that did not. These illustrative changes included a change to the data sufficiency standards for a carrier to receive an SMS score and changes to the calculation method.

We also spoke with 1) FMCSA officials in its headquarters office, Western Service Center in Colorado, and Colorado Division Office about the implementation of CSA and 2) representatives from the Colorado State Patrol and industry and safety interest groups. We selected Colorado because it was one of the initial pilot states for CSA and has been implementing the program since early 2008. We reviewed existing studies and literature on CSA and Congressional testimony from industry and safety interest representatives from a September 2012 hearing for the House Transportation and Infrastructure Committee. Appendix I contains a more detailed explanation of our scope and methodology. Appendix II contains details about estimating rates of regulatory violations in the SMS component of CSA. Appendix III contains details about the statistical validity of the SMS component of CSA. Appendix IV describes prior evaluations of SMS scores as measures of safety. Appendix V describes our analysis of regulatory violations and crash risk. Appendix VI describes the carriers we analyzed and provides the results from our analysis of FMCSA's methodology and our illustrative alternative.

We conducted this performance audit from August 2012 through February 2014 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on the audit objectives.

Background

Motor Carrier Industry Diversity

The commercial motor carrier industry represents a range of businesses, including private and for-hire freight transportation, passenger carriers, and specialized transporters of hazardous materials. As of 2012, FMCSA estimates that there were more than 531,000 active motor carriers, a number that fluctuates over time due to the approximately 75,000 new applications that enter the industry each year combined with thousands of carriers annually leaving the market. Among carriers we assessed for this report, most that operate in the United States are small firms; 93 percent of carriers own or operate 20 or fewer motor vehicles. Nonetheless, a large percentage of vehicles on the road are operated by large carriers. Approximately 270 carriers have more than 1,000 vehicles each and account for about 29 percent of all vehicles that FMCSA oversees.

FMCSA's Role

FMCSA is responsible for overseeing this large and diverse industry. FMCSA establishes safety standards for interstate motor carriers as well as intrastate hazardous material carriers operating in the United States.⁹ To enforce compliance with these standards, FMCSA partners with state agencies to perform roadside inspections of vehicles and investigations of carriers.¹⁰ In fiscal year 2012, FMCSA had a budget of approximately \$550 million and more than 1,000 FMCSA staff members located at headquarters, four regional service centers, and 52 division offices.

In 2008, FMCSA launched an operational model test of CSA in four states and began implementing the CSA program nationwide in 2010.¹¹ CSA is intended to improve safety beyond the prior SafeStat program by

⁹ 49 U.S.C. §§ 31136, 5103.

¹⁰ State agencies include state highway patrols, departments of transportation, and public utility commissions. FMCSA employs full-time vehicle inspectors on the southern border of the United States. In addition, all FMCSA safety investigators, safety auditors, and inspectors must conduct a minimum number and certain types of inspections annually to maintain certification.

¹¹ Originally, the Operational Model Test was conducted in Colorado, Georgia, Missouri, and New Jersey. Carriers in these States were randomly divided into a "test" group that was subject to the provisions of the new CSA Operational Model, and a "control" group that would continue to be monitored by the Agency's current process. For the four original States, the test ran for 29 months from February 2008 through June 2010. Five additional States (Montana, Minnesota, Maryland, Kansas, and Delaware) were phased into the program as test-only States.

identifying safety deficiencies through better use of roadside inspection data, assessing the safety fitness of more motor carriers and drivers,¹² and using less resource-intensive interventions to improve investigative and enforcement actions. From fiscal year 2007 through fiscal year 2013, FMCSA obligated \$59 million to its CSA program, including CSA development and technical support, information technology upgrades, and training. For fiscal year 2014, FMCSA requested \$7.5 million for CSA.¹³

CSA has three main components:

- *Safety Measurement System.* SMS uses data obtained from federal or state roadside inspections and from crash investigations to identify the highest risk carriers. SMS was designed to improve on SafeStat by incorporating all of the safety-related violations recorded during roadside inspections. Carriers potentially receive an SMS score in seven categories based on this information.
- *Intervention.* A set of enforcement tools, such as warning letters, additional investigations, or fines are used to encourage the highest risk carriers to correct safety deficiencies, or place carriers out-of-service.
- *Safety Fitness Determination Rule.* This future rulemaking will amend regulations to allow a determination—based in part on some of the same information used to calculate SMS—as to whether a motor carrier is fit to operate on the nation’s roads.¹⁴

SMS Carrier Performance

SMS, the measurement system component of CSA, uses the data collected from roadside inspections and crash reports to quantify a carrier’s safety performance relative to other carriers. Specific carrier violations recorded during roadside inspections are assigned to one of six Behavioral Analysis and Safety Improvement Categories (BASIC).

¹² In 2011, we found that FMCSA implemented part of the planned CSA program, but key components, including the rulemaking to determine if a carrier is unfit to operate, were still outstanding. See GAO, *Motor Carrier Safety: More Assessment and Transparency Could Enhance Benefits of New Oversight Program*, [GAO-11-858](#) (Washington, D.C.: Sept. 29, 2011).

¹³ The totals do not include full time employees dedicated to the program, which were not available.

¹⁴ 79 Fed. Reg. 896, 1038 (Jan. 7, 2014), Department of Transportation, Semiannual Regulatory Agenda.

According to FMCSA, these BASICS were developed under the premise that motor carrier crashes can be traced to the behavior of motor carriers and their drivers.¹⁵ A seventh category, called the Crash Indicator, measures a carrier's crash involvement history (see table 1). Each SMS score is designed to be a quantitative determination of a carrier's safety performance.

Table 1: FMCSA's Carrier Safety Measurement System Categories

BASIC/Crash indicator categories	Description	Percentage of carriers in our analysis population receiving an SMS score in each BASIC^a
Crash Indicator	Histories or patterns of high crash involvement, including frequency and severity. ^b	4.9%
Controlled Substances and Alcohol	Operation of a commercial motor vehicle (CMV) by a driver who is impaired due to alcohol, illegal drugs, or misuse of prescription or over-the-counter medications, including possession of controlled substances or alcohol.	0.8%
Driver Fitness	Operation of a CMV by a driver who is unfit due to lack of training, experience, medical qualification, or English language proficiency.	2.6%
Hours-of-Service Compliance	Operation of a CMV while ill, fatigued, or in noncompliance with hours-of-service regulations.	16.0%
Hazardous Materials	Unsafe handling or marking of hazardous material on a CMV.	0.6%
Unsafe Driving	Operation of a CMV in a dangerous or careless manner.	10.4%
Vehicle Maintenance	Failure to properly maintain a CMV or prevent shifting loads.	21.1%

Source: GAO and FMCSA.

^aSMS scores were calculated with data from December 2009 using FMCSA's SMS Methodology 3.0 based on our analysis population of approximately 315,000 carriers.

^bSMS evaluates a motor carrier's crash history. Although crash history is not specifically a behavior, it can be a consequence of behavior and may indicate a problem with the carrier that warrants intervention.

For each of the approximately 800 violations that fall under the various BASICS, FMCSA assigns a severity weight that is meant to reflect the violation's association with crash occurrence and crash consequence when compared with other violations within the same BASIC. For example, reckless driving violations, categorized in the Unsafe Driving BASIC, are assigned a severity weight of 10 out of a possible 10 because

¹⁵CSA, Carrier Safety Measurement System (CSMS) Methodology, Version 3.0.1, Revised August 2013.

FMCSA determined that these violations have a stronger relationship to safety risk than some other types of violations. Unlawfully parking, by comparison, is also categorized in the Unsafe Driving BASIC, but is assigned a severity weight of 1 out of 10.

FMCSA calculates SMS scores for carriers every month through a process that has three main steps, each of which is made up of several calculations.

Step 1: Establishing carriers' violation rates.¹⁶ To establish rates at which carriers violate regulations, FMCSA first assigns differing weights to each violation that occurred over the past 2 years, depending on the relative severity of each violation and the amount of time elapsed between the violation's occurrence and the score's calculation. These weighted violations are then summed for each BASIC. To obtain the violation rate, FMCSA divides the weighted total violations by one of two measures that FMCSA uses to adjust for a carrier's exposure to violations.

- For the Controlled Substances and Alcohol, Driver Fitness, Hours-of-Service Compliance, Hazardous Materials, and Vehicle Maintenance BASICs, FMCSA divides the number of weighted violations by the time-weighted number of relevant inspections a carrier received.¹⁷
- For the Unsafe Driving BASIC, and the Crash Indicator, FMCSA divides the number of weighted violations by a number obtained via

¹⁶ We use the term "violation rate" to refer to the calculation of all the time and severity weighted violations a carrier has incurred in a BASIC over a 24-month period relative to the carriers' exposure, measured either by the time-weighted number of inspections or the number of vehicles a carrier operates adjusted by the number of miles it travels. FMCSA refers to this calculation as the SMS "measure."

¹⁷ Relevant inspections are either a driver inspection, in which the inspection focuses on driver-related requirements, such as the driver's record of duty or medical certificate, or a vehicle inspection, which focuses on the condition of the motor vehicle. Driver inspections are the relevant inspection for the Unsafe Driving, Hours-of-Service Compliance, Driver Fitness, and Controlled Substances and Alcohol BASICs. Vehicle inspections are considered relevant inspections for the Vehicle Maintenance BASIC. For the Hazardous Materials BASIC, carriers that transport placardable quantities of hazardous materials are also subject to vehicle inspections as the relevant inspections. Throughout the report, we will refer to relevant inspections as simply inspections.

another calculation—the number of vehicles a carrier operates adjusted by the number of vehicle miles.¹⁸

FMCSA accounts for exposure in order to make the scores comparable across carriers. This approach has tradeoffs; while carriers can be compared without penalizing some for having had more inspections or road activity, exposure itself can be considered an element of risk. All else being equal, carriers with more road activity are involved in more crashes and potentially pose more risk to safety.

Step 2: Data sufficiency. Depending on the BASIC, carriers generally receive SMS scores if they meet minimum thresholds of exposure (i.e., number of vehicles or inspections), or a minimum number of inspections with violations (i.e., “critical mass”).¹⁹ For purposes of display on FMCSA’s public website and identifying the highest risk carriers for directing enforcement resources, FMCSA does not include scores for carriers that do not meet a so-called critical mass of violations. For each BASIC, this typically requires a minimum number of inspections that include violations in that BASIC, a violation in that BASIC in the last 12 months, and, for some BASICs, a violation during the most recent inspection.

Step 3: Dividing carriers into peer groups. After calculating violation rates, FMCSA assigns carriers it determines have sufficient exposure to peer groups with similar levels of on-road activity, or what the agency refers to as safety event groups. According to FMCSA, safety event groups are designed to account for the inherent greater variability in violation rates based on limited levels of exposure and the stronger level of confidence in violation rates based on carriers with higher exposure. FMCSA assigns carriers to safety event groups based on their number of

¹⁸ FMCSA uses an alternate measure of exposure for these BASICs because unsafe driving violations and crashes typically prompt an inspection, while other violations are typically discovered during an inspection.

¹⁹ FMCSA only displays SMS scores publicly, or uses SMS scores for further intervention, for carriers that have a “critical mass” of inspections with violations, which varies by BASIC. For the Hours-of-Service Compliance, Driver Fitness, Vehicle Maintenance, and Hazardous Materials BASICs, “critical mass” is defined as either three or five inspections with a violation in that BASIC. For the Unsafe Driving and Controlled Substances and Alcohol BASICs and Crash Indicator, “critical mass” is defined by the safety event group, which establishes the minimum number of inspections with violations required to be included in a safety event group.

inspections, the number of inspections with violations, or crashes the carriers have accrued in the previous 2 years. Within each safety event group, FMCSA calculates SMS scores by ranking carriers' violation rates (obtained in step 1 above) and assigning each carrier a percentile score ranging from 0 to 100, where 100 indicates the highest violation rate and the highest estimated risk for future crashes. FMCSA displays scores for five of the BASICS on its public website.²⁰

Interventions

Once SMS scores are calculated, FMCSA begins a Safety Evaluation that uses SMS scores to identify carriers with safety performance problems requiring intervention. FMCSA has defined a fixed percentage threshold for each BASIC that identifies those carriers that pose the greatest safety risk. (For example, the threshold for the Unsafe Driving BASIC is 65 for most carriers.) These carriers are then subject to one or more FMCSA actions from a suite of intervention tools that were expanded as part of CSA. Tools such as warning letters and on- and off-site investigations allow FMCSA and state investigators to focus on specific safety behaviors. FMCSA can also use enforcement strategies such as fines or placing a carrier out-of-service.²¹ The range of available enforcement options gives FMCSA investigators flexibility to apply interventions commensurate with a carrier's safety performance (see table 2). Seven of the nine interventions are currently implemented nationwide.²² Prior to CSA, FMCSA investigators' only tool was a labor intensive, comprehensive on-site investigation. With the additional set of interventions, FMCSA aims to reach more carriers with its existing resources.

²⁰ The remaining BASIC, Hazardous Materials Compliance, is restricted for a 1-year introductory period and the Crash Indicator is currently restricted from public view due to limitations with identifying crash fault. See <http://ai.fmcsa.dot.gov/sms/>

²¹ Currently, a carrier can only be declared unfit to operate upon a final unsatisfactory rating following an on-site inspection.

²² FMCSA has suspended plans to implement the remaining two interventions—off-site focused investigations and cooperative safety plans—nationwide until 2014 when implementation of a key piece of technology needed to implement them is scheduled to be completed.

Table 2: CSA Interventions Conducted in Fiscal Year 2012

Intervention	Description	Number in FY 2012^a
New interventions under CSA^b		
Warning letter	SMS automatically generates a warning letter to a carrier when it detects that a carrier has exceeded a specified threshold in one or more BASICS. This letter will describe the safety problem(s), offer suggestions for improvement, and explain how the carrier may challenge the accuracy of FMCSA's findings.	24,126
On-site focused investigation or federal/state focused compliance review	Carriers that (1) continue to exceed BASIC thresholds, (2) are involved in a fatal crash, or (3) are the subject of a complaint will undergo an on-site focused investigation so that FMCSA can attempt to determine the root causes of a specific safety problem and take corrective action.	10,361
Off-site investigation	Carriers that continue to exceed BASIC thresholds will be asked to voluntarily submit documents to help FMCSA evaluate carrier's safety management practices, determine the root causes of the safety problem, and take corrective action. For example, FMCSA may ask a carrier that exceeds the threshold in the Controlled Substances and Alcohol BASIC for records pertaining to its driver drug testing program. If a carrier does not comply with FMCSA's request, the agency may intervene through an on-site investigation.	573
Cooperative safety plan	Following an off-site or on-site investigation, the carrier and FMCSA will collaboratively create a safety plan that addresses the root causes of the problem, which the carrier has the option to implement.	402
Interventions used during and prior to CSA		
Notice of claim	Carriers with regulatory violations that are severe and warrant penalties will receive a legal notification of violation and penalty.	7,064
On-site comprehensive investigation or federal/state full compliance review	In instances of broad or complex safety problems, a carrier will be subject to a comprehensive on-site investigation similar to those conducted by FMCSA prior to CSA.	6,641
Unfit suspension/out-of-service order ^c	Carriers that receive a final unsatisfactory rating based on an on-site investigation will be prevented from operating.	855
Notice of violation	Carriers with regulatory violations that do not warrant fines and can be immediately corrected will receive a formal notice that requires a response. To avoid further intervention, including fines, the carrier must provide evidence of corrective action or initiate a successful challenge to the violation.	206

Source: FMCSA.

^aFMCSA considers data preliminary for 18 months after the fiscal year.

^bCSA also provides roadside inspectors with data that identifies a carrier's specific safety problems, by BASIC, based on SMS scores.

^cCurrently, a carrier can only be declared unfit to operate upon a final unsatisfactory rating following an on-site inspection.

According to FMCSA and state safety officials, an investigation or other intervention can also be initiated based on the results of a crash investigation, a complaint against a carrier, or a consistent pattern of

unsafe behavior by a carrier. FMCSA further designates some carriers that exceed multiple BASIC thresholds as “high risk.” According to FMCSA, many of these carriers are assigned a Safety Investigator, who must complete a comprehensive review within a year regardless of any changes in the carrier’s score. A carrier is considered high risk if it either:

- has an SMS score of 85 or higher in the Unsafe Driving BASIC or Hours-of-Service Compliance BASIC or the Crash Indicator, and one other BASIC at or above the intervention threshold,²³ or
- exceeds the intervention threshold for any four or more BASICs.

Carrier Fitness to Operate

Currently, FMCSA can only declare a carrier as unfit to operate upon a final unsatisfactory rating following an on-site inspection. In addition, FMCSA can order a carrier to cease interstate operations if it determines that the carrier is an imminent hazard. FMCSA can make this determination for several reasons including:

- receiving an “unsatisfactory” safety rating during an on-site comprehensive investigation and failing to improve the rating within 45 or 60 days;
- failing to pay a fine after 90 days;
- failing to meet the standards required for a New Entrant Audit;²⁴ or
- FMCSA determining the carrier to be an imminent hazard.

According to FMCSA, during fiscal year 2012, the agency issued 855 out-of-service orders due to an unsatisfactory rating, 1,557 for failing to pay a fine, and 47 because a carrier was determined to be an imminent hazard.

²³ FMCSA applies different thresholds for passenger carriers and hazardous materials carriers. For all other motor carriers, the threshold is established at 80 for Driver Fitness, Controlled Substances and Alcohol, and Vehicle Maintenance; and 65 for Unsafe Driving, Hours-of-Service Compliance, and the Crash Indicator.

²⁴ After a carrier registers for a USDOT number, FMCSA uses the new entrant safety assurance program to examine all new entrants registered to operate in interstate commerce—including all for-hire and private passenger, household goods, and freight carriers—and intrastate hazardous materials carriers. Under this program, which began in 2003, carriers are required to undergo a safety audit within 18 months of obtaining a USDOT number and beginning interstate operations. The purpose of this audit is to determine whether carriers are knowledgeable about and compliant with applicable safety regulations.

FMCSA has indicated its plans to propose using the same performance data that inform SMS scores to determine whether a carrier is fit to continue to operate. According to FMCSA, the Safety Fitness Determination rulemaking would seek to allow FMCSA to determine if a motor carrier is not fit to operate based on a carrier's performance in five of the BASICs, an investigation, or a combination of roadside and investigative information.²⁵ FMCSA proposes doing this through a public rulemaking process; it currently estimates that it will issue a proposed rule in May 2014.

CSA Program Increases Carrier Interventions, but FMCSA Faces Challenges in Identifying High Risk Carriers

CSA has been successful in raising the profile of safety in the motor carrier industry and providing FMCSA with more tools to increase interventions with carriers. However, FMCSA faces two major challenges in reliably assessing safety risk for the majority of carriers in the industry and prioritizing the riskiest carriers for intervention. First, we found that the majority of regulations used to calculate SMS scores are not violated often enough to strongly associate them with crash risk for individual carriers. Second, for most carriers, FMCSA lacks sufficient safety performance information to ensure that FMCSA can reliably compare them with other carriers. FMCSA mitigates this issue by—among other things—establishing data sufficiency standards. However, we found that these standards are set too low, and by strengthening data sufficiency standards SMS would better identify risky carriers and better prioritize intervention resources to more effectively reduce crashes. Setting a data sufficiency standard involves tradeoffs between scoring more carriers and ensuring that the scores calculated are reliable for the purposes for which they are used.

CSA Expands FMCSA's Reach and Raises the Profile of Safety in the Industry

CSA has helped FMCSA reach more carriers and provided benefits to a range of stakeholders. Since CSA was implemented nationwide in 2010, FMCSA has intervened with more carriers annually than under SafeStat. From fiscal year 2007 to fiscal year 2012, FMCSA increased its number of annual interventions from about 16,000 to about 44,000, largely by sending warning letters to carriers deemed to be above the intervention threshold in one or more BASICs (see table 3). FMCSA and state

²⁵ 79 Fed. Reg. 896, 1038 (Jan. 7, 2014), Department of Transportation, Semiannual Regulatory Agenda.

partners also took advantage of new ways to investigate carriers, such as off-site investigations and on-site focused investigations, to complete 23 percent more investigations in fiscal year 2012 compared to fiscal year 2007 when only compliance reviews were used.

Table 3: Number of FMCSA Interventions, Fiscal Years 2007 to 2012

Intervention	FY2007	FY2008	FY2009	FY2010	FY 2011	FY2012
Investigations ^a	16,385	15,625	16,923	20,155	18,422	20,213
Warning letters ^b	—	—	9,681	15,328	40,944	24,126
Total	16,385	15,625	26,604	35,483	59,366	44,339

Source: GAO analysis of FMCSA data.

^aFor fiscal year 2007 to fiscal year 2009, investigations include all compliance reviews, including hazardous materials reviews, household goods reviews, motor coach reviews, and conditional carrier reviews. For fiscal year 2010 to fiscal year 2012, investigations include all FMCSA reviews including off-site investigations, on-site focused investigations, on-site comprehensive investigations, full and focused compliance reviews (beginning in 2011), hazardous materials reviews, household goods reviews, passenger reviews, and motor coach reviews.

^bAccording to FMCSA, full-scale national deployment of warning letters occurred during fiscal year 2011 resulting in a spike in warning letters issued.

In addition, CSA provides data for law enforcement and industry stakeholders about the safety record of individual carriers. For example, as part of the CSA program, FMCSA publicly provides historical individual carrier data on inspections, violations, crashes, and investigations on its website. According to law enforcement and industry stakeholders we spoke with, CSA organizes violation information for law enforcement and carrier data related to the BASICs help guide the work of state inspectors during inspections.

Law enforcement officials and industry stakeholders generally supported the structure of the CSA program. These stakeholders told us that CSA's greater reach and provision of data have helped raise the profile of safety issues across the industry. According to industry stakeholders, carriers are now more engaged and more frequently consulting with law enforcement for safety briefings. In Colorado, law enforcement officials told us that CSA has improved awareness and engagement within the motor carrier industry there. A state industry representative told us that CSA has improved safety because carriers are in a competitive business and can feel pressure to improve safety scores to gain an advantage over the competition.

Relationship between Violation of Most Regulations and Crash Risk Is Unclear

The relationship between violation of most regulations FMCSA included in the SMS methodology and crash risk is unclear, potentially limiting the effectiveness of SMS in identifying carriers that are likely to crash. According to FMCSA, SMS was designed to improve on its previous approach to identify unsafe motor carriers by incorporating into the BASICS all of the safety-related violations recorded during roadside inspections. For SMS to be effective in identifying carriers that crash, the violation information that is used to calculate SMS scores should have a relationship with crash risk. Carriers that violate a given regulation more often should have a higher chance of a crash or a higher crash rate than carriers that violate the regulation less often. However, we found that FMCSA's safety data do not allow for validations of whether many regulatory violations are associated with higher crash risk for individual carriers. Our analysis found that most of the regulations used in SMS were violated too infrequently over a 2-year period to reliably assess whether they were accurate predictors of an individual carrier's likelihood to crash in the future. We found that 593 of the approximately 750 regulations we examined were violated by less than one percent of carriers.²⁶ Of the remaining regulations with sufficient violation data, we found 13 regulations for which violations consistently had some association with crash risk in at least half the tests we performed, and only two violations had sufficient data to consistently establish a substantial and statistically reliable relationship with crash risk across all of our tests. (For more information, see app. V.) FMCSA attempted to compensate for the infrequency of violations by, among other things, evaluating aggregate data to establish a broader relationship between a group of violations and crash risk.²⁷ However, evaluations completed by outside groups have found weaker relationships between SMS scores and the crash risk of individual carriers than FMCSA's evaluations of

²⁶ While SMS includes approximately 800 of FMCSA's regulations, our analysis looked at the 754 regulations available for the time frame of our analysis in order to limit violations to those that had sufficient violation data to examine over time. To conduct our analysis, a regulation needed to be present both during our analysis observation period, December 2007 to December 2009, and our evaluation period, December 2009 to June 2011.

²⁷ See Volpe, 2008. Volpe National Transportation Systems Center, the American Transportation Research Institute, and FMCSA have conducted studies examining the association between violations and crash risk. These studies evaluated grouped or aggregate data rather than studying the statistical association between violation and individual carrier behavior. Our analysis focused on the relationship between violations and crash risk at the carrier level, which is the level of analysis at which SMS calculates scores and uses them to make high-risk determinations and guide interventions.

aggregate data (for more information, see app. IV). SMS is intended to provide a safety measure for individual carriers, and FMCSA has not demonstrated relationships between groups of violations and the risk that an individual motor carrier will crash. Therefore, this approach of aggregating data does not eliminate the limitations we identified.

Most Carriers Lack Sufficient Information to Reliably Compare Safety Performance across Carriers

Most carriers lack sufficient safety performance information to ensure that FMCSA can reliably compare them with other carriers. As mentioned, SMS is designed to compare violation rates across carriers for the purposes of prioritizing intervention resources. These violation rates are calculated by summing a carrier's weighted violations relative to each carrier's exposure to committing violations, which for the majority of the industry is very low. About two-thirds of carriers we evaluated operate fewer than four vehicles and more than 93 percent operate fewer than 20 vehicles. Moreover, many of these carriers' vehicles are inspected infrequently. (See table 14 in app. VI) Generally, statisticians have shown that estimations of any sort of rate—such as the violation rates that are the basis for SMS scores—become more reliable when they are calculated from more observations. In other words, as observations increase, there is less variation and thus more confidence in the precision of the estimated rate. Given that SMS calculates violation rates for carriers having a very low exposure to violations, such as operating one or two vehicles or subject to a few inspections, many of the SMS scores based on these violation rates are likely to be imprecise.²⁸ Carriers with few inspections or vehicles will potentially have estimated violation rates that are artificially high or low and thus not sufficiently precise for comparison across carriers. Further, because SMS scores are calculated by ranking carriers in relation to one another, imprecise rate estimates for some carriers can cause other carriers' SMS scores to be higher or lower than they would be if they were ranked against only carriers with more reliable violation rates. This creates the likelihood that many SMS scores do not represent an accurate or precise safety assessment for a carrier. As a result, there is less confidence that SMS scores are effectively

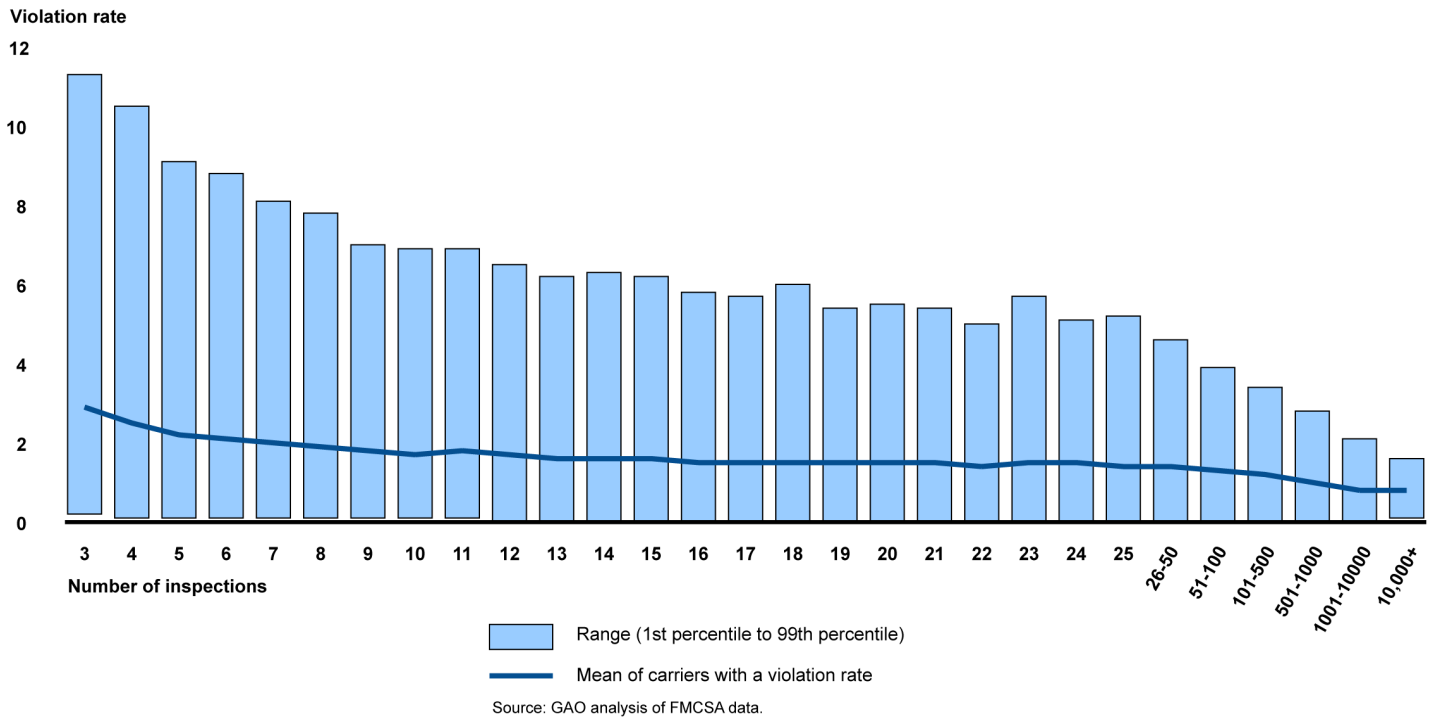
²⁸ Both statistical theory and our analysis show that the precision of estimated rates for carriers with low exposure, measured by vehicles or inspections, is lower than for carriers with more exposure, and that rate estimates can have artificially low or high values for these low-exposure carriers. The amount of data required depends on the degree of imprecision that the user is willing to accept for a given purpose. We describe these principles and provide references in appendix II. Prior evaluations discuss similar issues about SMS scores as measures of safety, see appendix IV.

determining which carriers are riskier than others. (App. II provides a more technical discussion of these issues.)

For the five SMS BASICs for which FMCSA uses relevant inspections as a measure of exposure—Hours-of-Service Compliance, Driver Fitness, Controlled Substances and Alcohol, Vehicle Maintenance, and Hazardous Materials—estimated violation rates can change by a large amount for carriers with few inspections even when the number of their violations changes by a small amount. For example, for a carrier with 5 inspections, a single additional violation could increase that carrier's violation rate 20 times more than it would for a carrier with 100 inspections.²⁹ This sensitivity can result in artificially high or low estimated violation rates that are potentially imprecise for carriers with few inspections. As an example, our analysis of FMCSA's method shows that among carriers for which we calculated a violation rate for the Hours-of-Service Compliance BASIC, violation rate estimates are more variable for carriers with fewer inspections. As shown in figure 1, violation rates tend to vary by a larger amount across carriers with few inspections than across carriers with more inspections. As a consequence, a high estimated violation rate for a carrier with few inspections may reflect greater safety risk, an imprecise estimate, or both. Further, comparisons among carriers are meaningful only to the extent they involve carriers with sufficient inspections and thus more precise estimated violation rates.

²⁹ This example is illustrative; actual changes to a carrier's SMS score would vary based on the number of previous violations, the severity of the violation, and other factors.

Figure 1: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Hours-of-Service Compliance BASIC

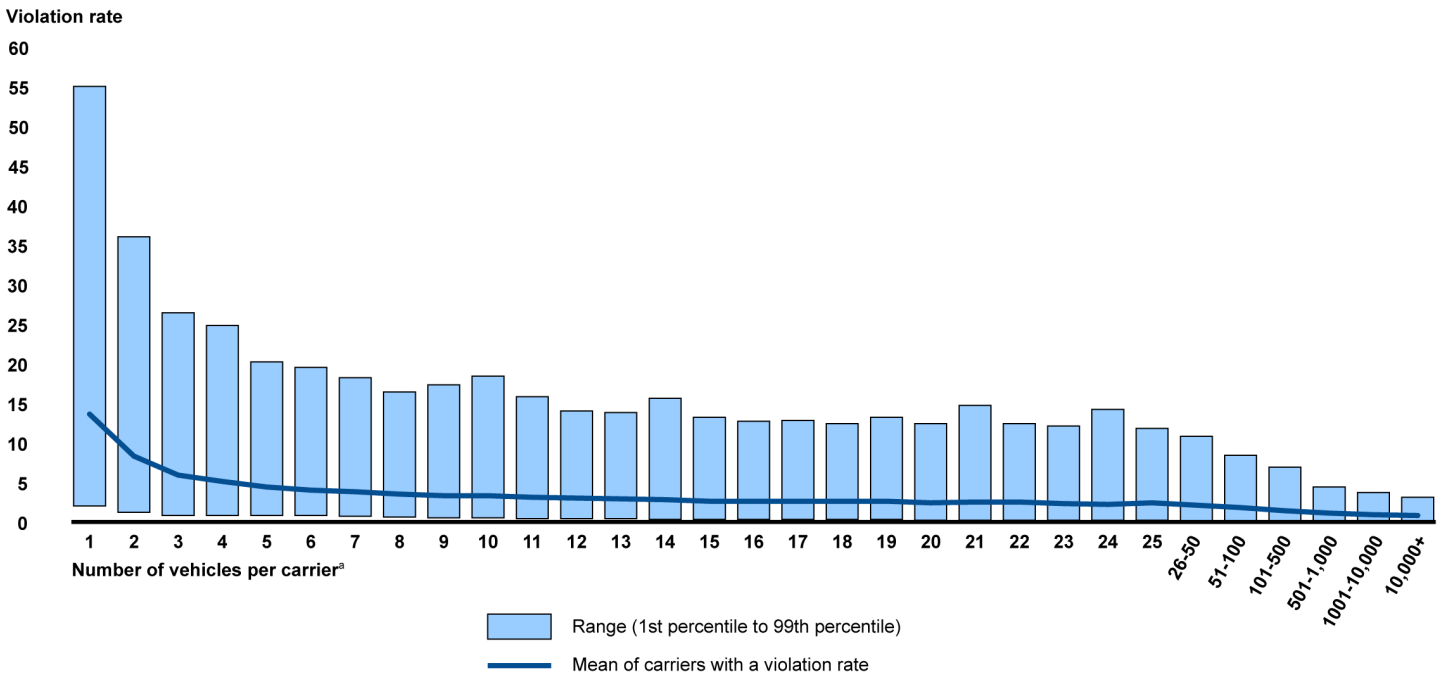


Similar to carriers with few inspections, carriers with few vehicles are also subject to potentially large changes in their estimated violation rates, which can affect a carrier’s SMS scores. For the Unsafe Driving BASIC and the Crash Indicator, FMCSA measures exposure using a hybrid approach that considers a carrier’s number of vehicles and its vehicle miles traveled—when the latter information is available.³⁰ Figure 2 shows that among carriers for which we calculated a violation rate using FMCSA’s method for the Unsafe Driving BASIC, carriers that operate fewer vehicles, for example fewer than 5, experience a greater range in violation rates per vehicle than carriers operating more vehicles, for

³⁰ Unsafe driving violations—such as a speeding infractions—and crashes are not tied to inspections conducted by law enforcement, which justifies the different measure of exposure.

example, greater than 100. (For similar results on other BASICs, see figures 10 to 16 in app. VI.)

Figure 2: Average and Range (between the 1st and 99th Percentiles) of Violation Rates for Carriers in the Unsafe Driving BASIC



Source: GAO analysis of FMCSA data.

^aThis number is an adjusted average number of vehicles that FMCSA uses to calculate an SMS score for carriers in the Unsafe Driving BASIC.

Researchers have raised additional concerns about the quality and accuracy of the data FMCSA uses to calculate SMS scores that could potentially compound the problems with the precision of violation rate estimates.³¹ These issues further limit the precision of carriers' estimated violation rates, and consequently their SMS scores. For example:

- The frequency of an individual carrier's inspections varies depending on where the carrier operates. States vary on inspection and enforcement practices. Some studies have shown that inspectors or

³¹ We did not directly assess the reliability of the data for purposes other than our use in an effectiveness test.

law enforcement officers in some states cite vehicles for certain violations more frequently than in other states.

- Delays in reporting crash data to FMCSA, as well as missing or inaccurate data, can affect a carrier's Crash Indicator SMS scores. These delays can vary by state.
- Data elements used to calculate violation rates for the Unsafe Driving BASIC and Crash Indicator are based on information that is self reported by the carrier. Inaccurate, missing, or misleading reports by a carrier could directly influence their SMS scores. Additionally, among carrier data we evaluated, more than 50 percent did not report their vehicle miles traveled to FMCSA.

FMCSA Has Worked to Address Issues with Precision, but Its Methods Do Not Fully Address Limitations

FMCSA acknowledges that violation rates for carriers with low exposure can be less precise and they attempt to address this limitation in two main ways, but the methods incorporated do not solve the underlying problems. As a result, SMS scores for these carriers are less reliable as relative safety performance indicators, which may limit FMCSA's ability to more effectively prioritize carriers for intervention.

Data Sufficiency Standards

FMCSA established minimum data sufficiency standards to eliminate carriers that lack what it has determined to be a minimum number of inspections, inspections with violations, or crashes to produce a reliable SMS score. For example, in the Hours-of-Service Compliance BASIC, FMCSA does not calculate SMS scores for a carrier unless it has at least three inspections and at least one violation within the preceding two years. In addition, as previously mentioned FMCSA applies another data sufficiency standard requiring a carrier to have a "critical mass" of inspections with violations in order for an SMS score to be a basis for potential intervention, or to be publicly displayed.³²

While this approach helps address the problems for carriers with low exposure, it is not sufficient to ensure that SMS scores effectively prioritize the riskiest carriers for intervention. For most BASICs, we found FMCSA's data sufficiency standards too low to ensure reliable

³² For the Unsafe Driving and Controlled Substances BASICs, and the Crash Indicator, SMS does not limit carriers based on the measure of exposure—relevant inspections or vehicles. SMS requires that carriers have a critical mass of three inspections with Unsafe Driving violations, two crashes, or one Controlled Substance or Alcohol violation. As a result, for these BASICs, comparisons are drawn between carriers with very low levels of exposure, as low as one vehicle or one relevant inspection.

comparisons across carriers. In other words, many carriers' violation rates are based on an insufficient number of observations to be comparable to other carriers in calculating an accurate safety score. Our analysis shows that rate estimates generally become more precise around 10 to 20 observations, higher than the numbers that FMCSA uses for data sufficiency standards. However, the determination of the exact data sufficiency standard needs to be based on a quantitative measure of confidence to fully consider how precise the scores need to be for the purposes for which the scores are used.³³ (For more information, see app. II.)

Safety Event Groups

FMCSA groups the carriers meeting FMCSA's data sufficiency standards for each BASIC into safety event groups in order to, according to FMCSA, "account for the inherent greater variability in violation rates based on limited levels of exposure and the stronger level of confidence in violation rates based on higher exposure."³⁴ FMCSA assigns carriers to groups based on inspections or inspections with violations depending on the BASIC or on crashes for the Crash Indicator. For example, the first safety event group in the Hours-of-Service Compliance BASIC includes carriers that received from 3 to 10 inspections; the second group includes carriers that received from 11 to 20 inspections, and so forth. Within each safety event group, FMCSA rank orders carriers by violation rate and assigns a percentile as an SMS score.

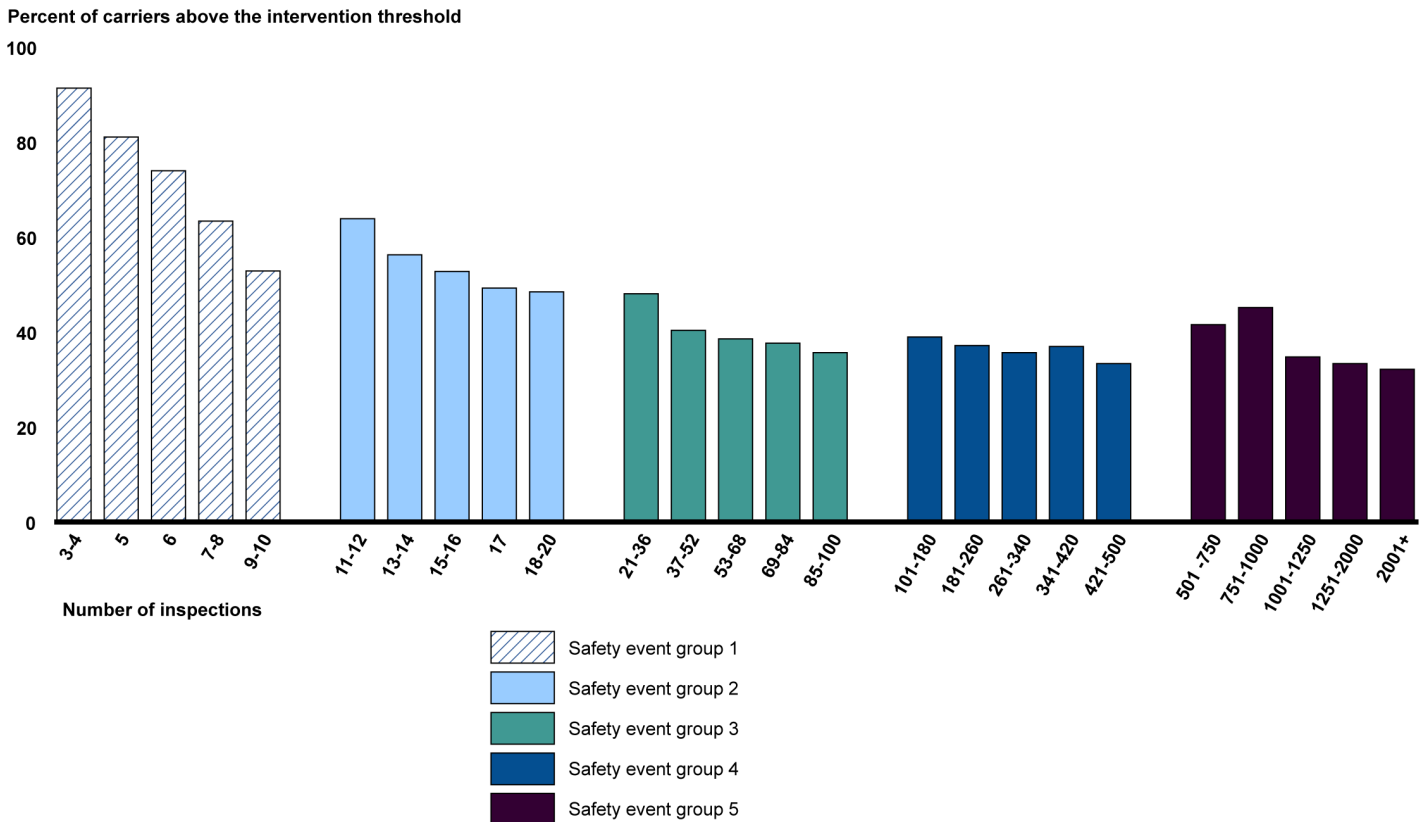
However, assigning carriers to safety event groups does not eliminate the imprecision of the violation rates that are the basis for SMS scores. Instead, for carriers with lower exposure, this approach makes comparisons across carriers within a safety event group with similarly imprecise violation rates. These comparisons are only as precise as the violation rate estimates that go into them. Our analysis shows that carriers with lower exposure within the safety event groups tend to

³³Rate estimates become more precise with each additional observation estimates based on 10 to 20 observations are more precise than those based on 1 to 5 observations, as we show in figure 1, figure 2, and appendix II. However, the amount of data required in practice depends on the degree of imprecision the user is willing to accept for a given purpose. This trade-off, in turn, depends on how the user considers the consequences of inaccuracy. As an example from another policy area, thresholds of 16 are consistent with criteria used by the Centers for Disease Control and Prevention (CDC) to suppress or caveat rate estimates for the purpose of public display.

³⁴ CSA, CSMS Methodology, Version 3.0.1, Revised August 2013.

exceed FMCSA's intervention thresholds at disproportionately higher rates than carriers with more exposure. For example, FMCSA's Hours-of-Service Compliance BASIC has five safety event groups. The group of carriers with the fewest number of inspections in each safety event group tends to have a higher percentage of carriers identified as above the intervention threshold than the group of carriers with a greater number of inspections (see fig. 3). This suggests that FMCSA's methodology is not adequately accounting for differences in exposure, as it is intended to do, but rather is systematically assigning higher scores for carriers with fewer inspections. (See figs. 17 to 25 in app. VI for other BASICs.)

Figure 3: Percentage of FMCSA Scored Carriers in the Hours-of-Service BASIC above the Intervention Threshold by Number of Inspections

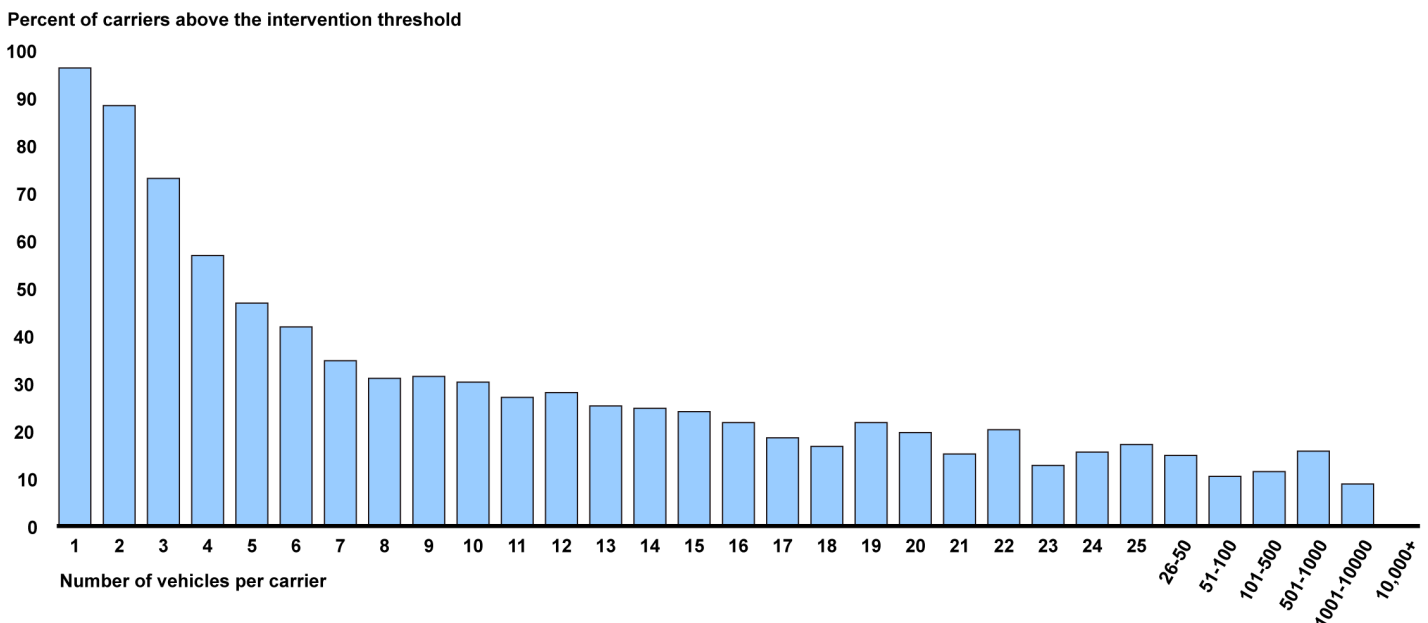


Source: GAO analysis of FMCSA data.

FMCSA's method of categorizing the carriers into safety event groups for the remaining BASICs also demonstrates how imprecision

disproportionately affects small carriers. For the Unsafe Driving and Controlled Substances BASICs, FMCSA forms safety event groups based on the number of inspections with violations. Similarly, for the Crash Indicator, safety event groups are based on a carriers' number of crashes. By using infractions or crashes to categorize carriers, FMCSA is not addressing its stated intent of having safety event groups account for differences in variability due to exposure. As a result, FMCSA derives SMS scores for the Unsafe Driving BASIC and the Crash Indicator by directly comparing small carriers with greater variability in their violation rates—including many carriers with a violation rate based on one vehicle—to larger carriers for which violations rates can be calculated with greater confidence. We found that among carriers that received an SMS score in Unsafe Driving, carriers with fewer than 20 vehicles are more than 3 times as likely to be identified as above the intervention threshold than carriers with 20 or more vehicles (see fig. 4). Of the carriers operating one vehicle, nearly all were identified as above the intervention threshold. (See figs. 26 to 32 in app. VI for other BASICs.)

Figure 4: Distribution of FMCSA Scored Carriers above the Unsafe Driving BASIC Threshold by Carrier Size



Source: GAO analysis of FMCSA data.

FMCSA contends that these results are expected because only small carriers that exceed critical mass standards receive an SMS score, and

small carriers that exceed this threshold have demonstrated several occurrences of risky behavior despite their limited exposure. However, this illustrates the volatility of rates and the disproportionate effect a single violation can have given how FMCSA has structured SMS. For example, using FMCSA's data sufficiency standards, a carrier with one vehicle (forty percent of the carriers in our analysis population have one vehicle) and two inspections with unsafe driving violations does not have sufficient information to be displayed or considered for intervention. However, a single additional violation, regardless of the severity of the violation, would likely mean that the carrier would be scored above threshold and prioritized for intervention. A relatively small difference in the number of violations could change a carrier's status from "insufficient information", to "prioritized for intervention" with potentially no interim steps. Conversely, a carrier such as this will have a very difficult time improving its SMS score to be below threshold.

Strengthened Data Sufficiency Standards Can Improve FMCSA's Ability to Identify High Risk Carriers

Our analysis shows that FMCSA could improve its ability to identify carriers at higher risk of crashing by applying a more stringent data sufficiency standard. As previously discussed, FMCSA uses SMS scores to identify carriers with safety performance problems—those above the threshold in any BASIC—for prioritization for intervention, and considers carriers with SMS scores above the intervention threshold in multiple BASICs as high risk. Overall, SMS is successful at identifying a group of high risk carriers that have a higher group crash rate than the average crash rate of all carriers that we evaluated. However, further analysis shows that a majority of these high risk carriers did not crash at all, meaning that a minority of carriers in this group were responsible for all the crashes. As a result, FMCSA may devote significant intervention resources to carriers that do not pose as great a safety risk as other carriers, to which FMCSA could direct these resources. Given the issues with precision discussed above, we developed and tested an alternative to FMCSA's method that sets a single data sufficiency standard, based on the relevant measure of exposure—either at least 20 inspections or at least 20 vehicles (depending on the BASIC), and eliminates the use of safety event groups. This approach is designed to illustrate how a stronger data sufficiency standard can affect the identification of higher risk carriers and is not meant to be a prescriptive design to replace

current SMS methods.³⁵ The result of this analysis demonstrates the effect that including carriers with low levels of exposure and highly variable violation rates can have on FMCSA's prioritization of carriers for intervention. Using this illustrative alternative, we found that FMCSA would have more reliably identified a higher percentage of carriers that actually had crashed than when compared to its existing methods. (Apps. I and VI provide more detail on this approach.) Specifically:

- This illustrative alternative identified about 6,000 carriers as high risk. During the evaluation period of our analysis, these carriers' group crash rate was approximately the same as the rate for FMCSA's high risk group (about 8.3 crashes per 100 vehicles). However, a much greater percentage of carriers (67%) identified as high risk using alternative higher data sufficiency standards crashed, and these carriers were associated with nearly twice as many crashes (see table 4).

Table 4: FMCSA's Existing Method of Identifying High Risk Carriers Compared with an Illustrative Alternative

	FMCSA's existing method	Illustrative alternative method
Number of carriers identified as high risk	7,201	6,007
(as a percentage of 314,757 carriers analyzed)	2.3%	1.9%
Percentage of carriers identified as high risk that crashed during the post period	39.0%	67.1%

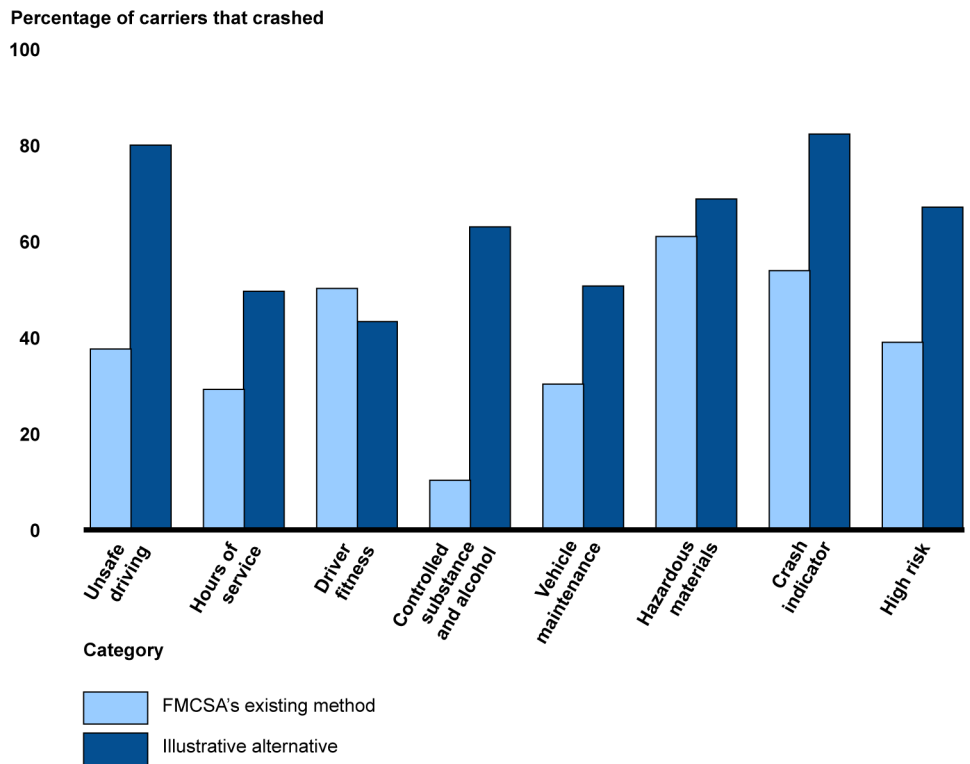
³⁵Our intent is to show the potential effect these changes have on the number of carriers assessed by SMS and the identification of high risk carriers in terms of crash rates, the number of carriers that crash, and the total number of crashes accounted for by the high risk group of carriers. We are including only carriers that recorded at least 20 driver inspections for Hours-of-Service Compliance, Driver Fitness, and Controlled Substances; 20 vehicle related inspections for Vehicle Maintenance; 20 hazardous materials related inspections for Hazardous Materials; or 20 average vehicles for Unsafe Driving and the Crash Indicator. Any carrier that meets these data sufficiency standards is assigned an SMS score for the observation period of analysis, even if that carrier does not have any violations, was free of violations for 12 months, or had a clean last inspection. Because we are imposing a stricter data sufficiency standard, we mitigate the need to segregate carriers with low exposure from those with higher exposure; consequently, we did not divide carriers into safety event groups for the purposes of this illustrative analysis. In addition, because many carriers lack information on vehicle miles traveled, we also simplified the calculation for the Unsafe Driving BASIC and the Crash Indicator by eliminating vehicle miles traveled from consideration.

	FMCSA's existing method	Illustrative alternative method
Number of crashes accounted for by carriers identified as high risk	12,624	22,961
(as a percentage of 120,334 crashes that occurred during the post period)	10.5%	19.1%
Group crash rate (per 100 vehicles) for the carriers identified as high risk	8.38	8.25

Source: GAO analysis of FMCSA data and methodology

- For five out of six BASICs, the Crash Indicator, and the high-risk designation, the illustrative alternative identified a higher percentage of individual carriers above the intervention threshold that actually crashed compared with FMCSA's existing method. (See fig. 5.)

Figure 5: Percentage of Carriers Identified as above FMCSA's Intervention Threshold, or High Risk, That Crashed during the Evaluation Period, Comparing FMCSA's Existing Method and Illustrative Alternative



Source: GAO analysis of FMCSA data.

- Using both FMCSA's method and the illustrative alternative, for most of the BASICs and the Crash Indicator the carriers identified above

the intervention threshold had a higher crash rate (crashes per 100 vehicles) than those below the intervention threshold (see table 5). However, using FMCSA's method, crash rates for the Controlled Substances and Alcohol BASIC have the opposite, negative association (3.2 crashes per 100 vehicles for carriers above threshold versus 5.2 crashes per 100 vehicles for carriers below threshold), whereas the illustrative alternative produces a positive association (4.7 crashes per 100 vehicles for carriers above threshold versus 3.8 crashes per 100 vehicles for carriers below threshold).

Table 5: Crash Rates per 100 Vehicles for Carriers with an SMS Score above and below FMCSA's Intervention Thresholds Using FMCSA's Method and Illustrative Alternative

		Unsafe Driving	Hours-of-Service Compliance	Driver Fitness	Controlled Substances and Alcohol	Vehicle Maintenance	Hazardous Materials	Crash Indicator
FMCSA's existing method	Carriers above threshold	7.1	6.6	2.9	3.2	5.6	5.5	7.2
	Carriers below threshold	3.6	3.6	3.9	5.2	3.6	3.5	3.2
Illustrative alternative	Carriers above threshold	6.1	6.7	2.6	4.7	6.4	5.1	6.8
	Carriers below threshold	1.8	3.4	4.1	3.8	3.7	3.6	2.2

Source: GAO analysis of FMCSA data and methodology

Overall, these results raise concerns about the effectiveness of the existing SMS as a tool to help FMCSA prioritize intervention resources to most effectively reduce crashes. FMCSA's existing SMS method successfully identified as high risk more than 2,800 carriers whose vehicles were involved in 12,624 crashes. However, FMCSA would have potentially prioritized limited resources to investigate more than 4,000 carriers that did not crash at all. Prioritizing resources to these carriers would limit FMCSA's ability to reduce the number of overall crashes, resulting in lost opportunities to intervene with the carriers associated with many crashes.

Implementing a stronger data sufficiency standard as presented involves tradeoffs between the number of carriers FMCSA can score, and the reliability of those scores. Our analysis found that by increasing the data sufficiency standards, fewer carriers would receive at least one SMS

score (approximately 44,000 carriers [14%] in the illustrative alternative versus approximately 89,000 [28%] using FMCSA’s method). The carriers assigned an SMS score under the illustrative alternative accounted for 78.2 percent of all crashes during our evaluation period. FMCSA’s existing method scores carriers responsible for about 85.9 percent of all crashes (see table 6). On the other hand, by setting a higher standard for data sufficiency, the illustrative alternative focuses on carriers that have a higher level of road activity, or exposure, to more reliably calculate a rate that tracks violations and crashes over the 2-year observation period. In addition, exposure itself is a large determinant of overall risk, when defined as a combination of threat and consequence, and could be used as a factor to identify carriers that analysis suggest present a higher future crash risk. This is consistent with the results in table 4 above, which show that a larger proportion of the higher risk carriers in the illustrative alternative crashed and were associated with a larger number and proportion of crashes.

Table 6: Comparison of FMCSA’s Method and Illustrative Alternative to Identify Carriers with an SMS Score in at Least One BASIC

	FMCSA’s existing method	Illustrative alternative method
Number of carriers with at least one SMS score calculated	89,212	44,008
(as a percentage of 314,757 carriers analyzed)	28.3%	14.0%
Number of vehicles associated with these carriers	2,705,485	2,733,240
(as a percentage of 3,565,363 vehicles analyzed)	75.9%	76.7%
Number of crashes associated with these carriers	103,350	94,143
(as a percentage of 120,334 crashes that occurred during the post period)	85.9%	78.2%

Source: GAO analysis of FMCSA data

Regardless of where the data sufficiency standard is set, using only SMS scores limits risk assessment for carriers that do not have sufficient performance information. Our analysis shows that using FMCSA’s existing method, about 28% of carriers have at least one SMS score, leaving approximately 72% of carriers without any SMS scores—largely due to insufficient information. The illustrative alternative scores fewer carriers—14%, leaving 86% of carriers without any SMS scores. However, according to an FMCSA official, there are other enforcement

mechanisms to assess and place unsafe carriers out-of-service, including when a carrier fails to improve from an unsatisfactory safety rating during a comprehensive review, fails to pay a fine, or FMCSA determines a carrier is an imminent hazard. Further, the FMCSA official said carriers that do not receive an SMS score can still be monitored because the officials can initiate investigations and remove carriers based on complaints and other initiatives. For example, FMCSA conducts inspection strike forces targeting unsafe drivers and carriers in a particular safety aspect, such as drug and alcohol safety records. These tools used in conjunction with the performance data, including roadside inspection and crash data, could provide FMCSA with complementary means to assess and target carriers that do not otherwise have sufficient data to reliably calculate SMS scores.

Precision Required in SMS Scores Depends on How They Are Used

The safety scores generated by SMS are used for many purposes, thus the appropriate level of precision required depends on the nature of these applications. According to FMCSA's methodology, SMS is intended to prioritize intervention resources, identify and monitor carrier safety problems, and support the safety fitness determination process.³⁶ In setting a data sufficiency standard, FMCSA needs to consider how precise the scores need to be, and a score's required precision depends on the purposes for which the scores are used.³⁷

FMCSA officials told us the primary purpose of SMS is to serve as a general radar screen for prioritizing interventions. However, as discussed above, due to insufficient data, SMS is not as effective as it could be for this purpose. Further, if the same safety performance data used to inform SMS scores are intended to help determine a carrier's fitness to operate, most of these same limitations will apply. According to FMCSA, the Safety Fitness Determination rulemaking would seek to allow FMCSA to determine if a motor carrier is not fit to operate based on a carrier's performance in five of the BASICs, an investigation, or a combination of roadside and investigative information.³⁸ FMCSA has postponed the

³⁶ CSA, CSMS Methodology, Version 3.0.1 Motor Carrier Preview, Revised August 2013.

³⁷ GAO data reliability standards suggest that the reliability of data depend on the degree of risk and strength of corroborating evidence. GAO, *Assessing the Reliability of Computer-Processed Data*, [GAO-09-680G](#) (Washington, D.C.: July 2009).

³⁸ 79 Fed. Reg. 896, 1038 (Jan. 7, 2014), Department of Transportation, Semiannual Regulatory Agenda.

planned rulemaking until May 2014. However, basing a carrier's safety fitness determination on limited performance data may misrepresent the safety status of carriers, particularly those without sufficient data from which to reliably draw such a conclusion.³⁹

In addition to using SMS for internal purposes, FMCSA has also stated that SMS provides stakeholders with valuable safety information, which can “empower motor carriers and other stakeholders...to make safety-based business decisions.”⁴⁰ FMCSA includes a disclaimer with the publicly released SMS scores stating that the data are intended for agency and law enforcement purposes, and readers should not draw safety conclusions about a carrier's safety condition based on the SMS score, but rather the carrier's official safety rating. Nonetheless, entities outside of FMCSA are also using SMS scores to assess and compare the safety of carriers. For example:

- The Department of Defense has written SMS scores into its minimum safety criteria for selecting carriers of hazardous munitions.
- FMCSA has released a mobile phone application—SaferBus—that is designed to provide safety information, including SMS scores, for consumers to use in selecting a bus company.
- Multiple stakeholders have reported that entities such as insurers, freight shippers and brokers, and others use SMS scores.

Given such uses, it is important that any information about SMS scores⁴¹ make clear to users, including FMCSA, the purpose of the scores, their precision, and the context around how they are calculated. Stakeholders have said that there is a lot of confusion in the industry about what the SMS scores mean and that the public, unlike law enforcement, may not understand the relative nature of the system and its limitations.

³⁹ In noting its upcoming Safety Fitness Determination proposed rulemaking, FMCSA states that “[a] risk of incorrectly identifying a compliant carrier as non-compliant—and consequently subjecting the carrier to unnecessary expenses—has been analyzed and has been found to be negligible under the process being proposed.” 79 Fed. Reg. at 1038.

⁴⁰ CSA, CSMS Methodology, Version 3.0.1 Motor Carrier Preview, Revised August 2013.

⁴¹ As noted above, FMCSA's publication of carriers' SMS scores on its website and encouragement to the public to use the scores to make safety-based business decisions is the subject of ongoing litigation.

Conclusions

With the establishment of its CSA program, FMCSA has implemented a data-driven approach to identify and intervene with the highest risk motor carriers. CSA helps FMCSA to reach more carriers through interventions and provides the agency, state safety authorities, and the industry with valuable information regarding carriers' performance on the road and problems detected during roadside inspections.

GAO continues to believe a data-driven, risk-based approach holds promise and can help FMCSA effectively identify carriers exhibiting compliance or safety issues—such as violations or involvement in crashes. However, assessing risk for a diverse population of motor carriers—many of which are small and inspected infrequently—presents several significant challenges for FMCSA. As a result, the precision and confidence of many SMS scores is limited, a limitation that raises questions about whether SMS is effectively identifying carriers at highest risk for crashing in the future.

As presented in the report, strengthening data sufficiency standards is one of several potential reforms that might improve the precision and confidence of SMS scores. However, strengthening data sufficiency standards involves a trade-off between assigning scores to more carriers and ensuring that those scores are reliable. Our analysis shows how improving the reliability of SMS scores by strengthening data sufficiency standards could better account for limitations in available safety performance information and help FMCSA better focus intervention resources where they can have the greatest impact on reducing crashes. In addition, if these same safety performance data are going to be used to determine whether a carrier is fit to operate, FMCSA needs to consider and address all identified data limitations, or these determinations will also be at risk.

Recommendations for Executive Action

To improve the CSA program, the Secretary of Transportation should direct the FMCSA Administrator to take the following two actions:

Revise the SMS methodology to better account for limitations in drawing comparisons of safety performance information across carriers; in doing so, conduct a formal analysis that specifically identifies:

- limitations in the data used to calculate SMS scores including variability in the carrier population and the quality and quantity of data available for carrier safety performance assessments, and

-
- limitations in the resulting SMS scores including their precision, confidence, and reliability for the purposes for which they are used.

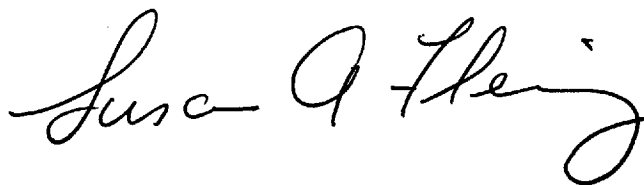
Ensure that any determination of a carrier's fitness to operate properly accounts for limitations we have identified regarding safety performance information.

Agency Comments

We provided a draft of this report to the USDOT for review and comment. USDOT agreed to consider our recommendations, but expressed what it described as significant and substantive disagreements with some aspects of our analysis and conclusions. USDOT's concerns were discussed during a meeting on January 8, 2014, with senior USDOT officials, including the FMCSA Administrator. Following this meeting, we made several clarifications in our report. In particular, FMCSA understood our draft recommendation to be calling for specific changes to its SMS methodology. It was not our intent to be prescriptive, so we revised our first recommendation to state that FMCSA should conduct a formal analysis to inform potential changes to the SMS methodology. In addition, we clarified in the analysis and conclusions our meaning of reliability in context of the purpose for which SMS is used.

We are sending copies of this report to relevant congressional committees and the Secretary of Transportation. In addition, the report is available at no charge on GAO's website at <http://www.gao.gov>.

If you or your staff have any questions about this report, please contact me at (202) 512-2834 or flemings@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made major contributions to this report are listed in appendix VII.



Susan Fleming
Director, Physical Infrastructure Issues

List of Congressional Committees

The Honorable Patty Murray
Chairman
The Honorable Susan Collins
Ranking Member
Subcommittee on Transportation, Housing
and Urban Development and Related Agencies
Committee on Appropriations
United States Senate

The Honorable Tom Latham
Chairman
The Honorable Ed Pastor
Ranking Member
Subcommittee on Transportation, Housing
and Urban Development and Related Agencies
Committee on Appropriations
House of Representatives

Appendix I: Scope and Methodology

This report addresses the effectiveness of the Compliance, Safety, Accountability (CSA) program in assessing safety risk for motor carriers. To assess how effectively CSA assesses the safety risk of motor carriers, we reconstructed the models the Federal Motor Carrier Safety Administration (FMCSA) uses to compute the SMS scores for all six Behavior Analysis and Safety Improvement Categories (BASICS) and the crash indicator. We then assessed the effect of changes to key assumptions made by the models. Using data collected by the U.S. Department of Transportation's Motor Carrier Management Information System (MCMIS) and historical SMS scores, and referencing the SMS algorithm and methodological documentation, we replicated the algorithm for calculating the SMS BASIC scores for the SMS 3.0 methodology.¹ Reconstructing FMCSA's models and replicating the SMS scores FMCSA produced for carriers was a necessary step to ensure that we understood the complexities of the models, the data used in the calculation of the SMS scores, and that the results we present in this report are comparable to FMCSA's outcomes. To corroborate our models with FMCSA's, we compared the SMS violation rates (measure scores) to FMCSA's results for December 2012. We assessed the reliability of data used, for our purposes, by reviewing documentation on FMCSA's data collection efforts and quality assurance processes, talking with FMCSA and Volpe National Transportation Systems Center officials about these data, and checking the data for completeness and reasonableness. We determined that the data were sufficiently reliable for the purpose of our data analysis.

We established a population of about 315,000 carriers for analysis that were under FMCSA's jurisdiction and showed indicators of activity over a 3 and a half year analysis period from December 2007 through June 2011.² The criteria used to identify these carriers were:

- U.S.-based carriers;

¹ CSA, Carrier Safety Measurement System Methodology (CSMS), Version 3.0.1, Revised August 2013. This update was issued after our analysis, based on the CSMS Version 3.0, was completed. However, version 3.0.1 did not include changes that substantively affected our analysis.

² We requested carrier data from FMCSA for December 2007 to June 2011. However, we received carrier data dated December 2008 through June 2012. Instead of submitting another data request, we were able to use the historical carrier files and to capture the relevant data from these snapshots to conduct our analysis for the earlier specified time period.

- interstate or intrastate hazardous materials carriers;
- carriers with at least one inspection or crash during the 2-year analysis observation period (December 18, 2007 to December 17, 2009); and
- carriers with a positive average number of vehicle count at any point during the analysis observation period (December 18, 2007, to December 17, 2009) and at any point during the evaluation period (December 17, 2009, to June 17, 2011).

During the first 2 years of this period, December 2007 through December 2009, we used each carrier's inspection, crash, and violation history to calculate SMS scores. This period is referred to as the observation period. The remaining 18 months, December 2009 through June 2011, were classified as the evaluation period. We used data from this period to identify carriers involved in a crash and estimate crash rates for these carriers. For the approximately 315,000 carriers in our analysis, there were approximately 120,000 crashes during the evaluation period. We chose the lengths of time for observation and evaluation, in part, to match FMCSA's effectiveness testing methods.

We tested the effectiveness of SMS by identifying and making changes to key assumptions of the model. Given FMCSA's use of these scores as quantitative determinations of a carrier's safety performance, we assessed the reliability of SMS scores as defined by the precision, accuracy, and confidence of these scores when calculated for carriers with varying levels of carrier exposure—measured by FMCSA as either inspections or an adjusted number of vehicles.³ We tested changes to the following characteristics of the model: the SMS measures of exposure, the method used to calculate time weights, the organization of the violations to the six BASICs, and the data sufficiency standards. To evaluate the results produced by each model, including FMCSA's, we examined the SMS scores and classifications of carriers into the high risk group. We compared the results from our revised models to the results from a baseline model, SMS 3.0. For each model, we measured whether carriers were involved in a crash, calculated group crash rates, and calculated total crashes in the evaluation period for carriers that were and were not classified as high risk in the observation period. Due to ongoing litigation related to CSA and the publication of SMS scores, we did not

³ GAO, *Assessing the Reliability of Computer-Processed Data*, [GAO-09-680G](#) (Washington, D.C.: February 2009).

assess the potential effects or tradeoffs resulting from any public use of these scores.⁴

To determine the extent to which CSA identifies and intervenes with the highest risk carriers, we examined how our changes to FMCSA's key assumptions affected the safety scores and identification of high risk carriers. Specifically, we identified the carriers with SMS scores above FMCSA's intervention threshold in each BASIC and the carriers considered high risk according to FMCSA's high risk criteria. Using this analysis, we designed an illustrative alternative method that incorporates the following changes:

- including only carriers with at least 20 observations in the following measures of exposure:
 - driver inspections when calculating scores for the Hours-of-Service Compliance, Driver Fitness, and Controlled Substances BASICs;
 - vehicle related inspections for the Vehicle Maintenance BASIC;
 - vehicle related inspections where placardable quantities of hazardous materials are being transported for Hazardous Materials BASIC; and
 - average power units for the Unsafe Driving and Crash Indicator BASICs;⁵
- assigning an SMS score to any carrier meeting these data sufficiency standards (e.g., 20 inspections), even if that carrier does not have any

⁴ See *Alliance for Safe, Efficient and Competitive Truck Transportation v. FMCSA*, No. 12-1305, D.C. Cir. (filed July 16, 2012; oral argument Sept. 10, 2013). The litigation has been brought against FMCSA by a number of motor carrier trade associations and challenges, among other things, the agency's public disclosure of the SMS scores and its encouragement of the use of these public data to help make sound business judgments. The carriers have requested the court to order that the SMS scores not be publicly available until alleged flaws in the methodology are addressed in the context of the planned rulemaking. Under GAO's policy to avoid addressing the merits of matters pending in litigation, we did not assess these matters.

⁵ Our analysis only included carriers with a recorded crash at any time in the MCMIS crash tables.

violations, was free of violations for 12 months, or had a clean last inspection;⁶

- eliminating safety event groups because of the stricter data sufficiency standard; and
- using only the average number of vehicles as the measure of exposure for carrier's assessed in the Unsafe Driving and Crash Indicator BASICS.

Appendix VI provides the complete results of our replication of FMCSA's existing SMS and our illustrative revision to it.

We also examined the extent to which the regulatory violations that largely determine SMS scores can predict future crashes. We developed eight model groups to test the relationship between violations and violation rates, and crashes. We tested only the violations that had non-zero variance and observations for at least 1 percent of the test population. To control for small exposure measures when estimating rates, we estimated models comparing carriers' observed crash status to Bayesian crash rates; used observed violation rates versus Bayesian violation rates; and compared a full model sample to a restricted model sample of carriers with at least 20 vehicles.⁷ We also conducted a sensitivity analysis to validate the predictive power of the models we developed. We ran multiple variations of these models to determine the number and types of violations that were predictive versus unstable. For

⁶ FMCSA doesn't calculate an SMS score for carriers if they haven't had a violation in the last 12 months in a particular BASIC and if that carrier had no violations in the most recent inspection. For 4 BASICS—Hours-of-Service Compliance, Driver Fitness, Vehicle Maintenance, and Hazardous Materials—carriers' SMS scores are eliminated if the carrier has not had a violation recorded in that BASIC in the last 12 months and did not have a violation recorded in the BASIC during the last inspection. Carriers meeting these criteria for these BASICS are removed from the rank order before SMS scores are assigned. For the other two BASICS—Unsafe Driving and Controlled Substances and Alcohol—and the Crash Indicator, carriers' SMS scores are eliminated if their violations in the BASIC, or crashes, are older than 12 months. For these BASICS, SMS scores are assigned to all carriers; carriers meeting the criterion have their SMS scores removed, but the remaining carriers retain their previously assigned SMS score. Our analysis shows that more than 57,000 carriers had SMS scores excluded using FMCSA's method.

⁷ Empirical Bayesian methods prevent estimates from converging to artificially extreme values for carriers whose raw rate estimates are based on small samples (low exposure). The estimator does this by effectively "borrowing information" from other, larger carriers whose rates can be estimated more precisely. Appendix II describes our use of Bayesian methods in more detail.

more information on this specific analysis and model results, please see appendix V.

In addition, we spoke with FMCSA officials in Washington, D.C., and at the Western Service Center and the Colorado Division Office in Lakewood, Colorado, and reviewed existing studies and stakeholder concerns about the SMS model and its outcomes. To understand the impact of CSA on law enforcement, we spoke with law enforcement officials at the Colorado State Patrol. We selected Colorado because it was one of the initial pilot states for CSA, and has been implementing the program since early 2008. We also interviewed representatives from industry and safety interest groups from the Colorado Motor Carriers Association, the Commercial Vehicle Safety Alliance, and the American Trucking Associations. Additionally, we attended meetings of the Motor Carrier Safety Advisory Committee's CSA subcommittee and reviewed the minutes and related documentation from other meetings we did not attend. We also reviewed congressional testimony from industry and safety interest representatives from a September 2012 hearing for the House Transportation and Infrastructure Committee. We reviewed stakeholder comments submitted between March 2012 and July 2012 in response to FMCSA's planned improvements to SMS.

We conducted this performance audit from August 2012 to February 2014 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Appendix II: Estimating Rates of Regulatory Violations in the Safety Measurement System

The FMCSA Safety Measurement System (SMS) methodology involves the calculation of weighted violation rates for regulations within each of six Behavioral Analysis and Safety Improvement Categories (BASICs) and a given time period. (A seventh indicator measures weighted crash rates in previous time periods, or “crash history.”) Carriers are assigned to Safety Event Groups based on measures of their exposure to committing violations, such as the number of driver or vehicle inspections, depending on the BASIC, and the weighted violation rates are transformed into percentiles for carriers within the same group. These percentiles ultimately determine carriers’ alert or high-risk statuses. Because regulatory violation rates strongly influence SMS scores, the precision with which these rates can be calculated becomes important for developing reliable measures of safety, as we discuss in the body of this report.

In this appendix, we summarize statistical methods for estimating rates and assessing their precision, or sampling error. We use these methods to estimate crash rates and their sampling error for a population of motor carriers that were active from December 2007 through December 2009. Carriers may vary widely in their level of activity, known as “exposure.” Both statistical theory and our analysis show that the precision of estimated rates for carriers with low exposure, measured by vehicles or inspections, is lower than for carriers with more exposure, and that rate estimates can become distorted to artificially low or high values for these low-exposure carriers. These results support our findings in the body of this report on the precision of FMCSA’s current approach to calculating safety risk scores and setting data sufficiency standards.

Statistical Methods for Estimating Violation Rates and Their Sampling Variance

Estimating rates of regulatory violations requires data on the number of violations that carriers incur within a given time period. If one makes the assumption that the number of violations is proportional to some measure of exposure (activity) and also assumes that the probability of observing violations within a large number of small independent exposure periods is small, the sampling error of a rate estimate decreases as exposure increases.

Specifically, assume that each carrier in a population of interest has a unique violation rate, λ_i . For a fixed time period and known exposure, t_i , the number of violations, V_i , is distributed as $V_i \sim \text{Poisson}(\lambda_i t_i)$, with $E(V_i) = \text{Var}(V_i) = \lambda_i t_i$. Since λ_i is unknown, it must be estimated from data on regulatory violations and exposure.

The maximum likelihood (ML) estimator for a single carrier's λ_i , given the model above, is $\hat{\lambda}_i = v_i / t_i$, with $\text{Var}(\hat{\lambda}_i) = \hat{\lambda}_i / t_i = v_i / t_i^2$.¹ The variance of the rate estimate increases exponentially as exposure decreases. Accordingly, an estimated rate for a specific carrier and time period can vary substantially from λ_i , particularly when exposure is low.

SMS is primarily concerned with measuring how regulatory violation rates vary over a population of active motor carriers. Even though ordinary methods of estimating these rates are unbiased and consistent, the collection of estimated rates for the population, $\hat{\lambda} = \{\lambda_1, \dots, \lambda_N\}$, may not accurately approximate the distribution of rates in the population, due to the errors associated with each estimate.² Statistics derived from these estimates, such as the percentiles that SMS uses to place carriers into alert and high-risk status, may be similarly prone to error.

Empirical Bayesian methods correct for this problem by estimating $\hat{\lambda}_i$ for each carrier to better estimate the distribution of rates across a population.³ Bayesian methods prevent estimates from converging to artificially extreme values for carriers whose raw rate estimates are based on small samples (low exposure). The estimator does this by effectively "borrowing information" from other, larger carriers whose rates can be estimated more precisely. In the evaluation of the CSA Pilot Test for FMCSA, the University of Michigan Transportation Research Institute used empirical Bayesian rate estimation methods to evaluate the association between SMS scores and crash risk, and cited similar benefits to those we discuss here.⁴

¹ For example, see Roger J. Marshall, "Mapping Disease and Mortality Rates Using Empirical Bayes Estimators," *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 40, no. 2 (1991): 284, or J. N. K. Rao, *Small Area Estimation* (Hoboken, NJ, 2003), 206.

² Rao, 206, and David Clayton and John Kaldor, "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics* 43 (September 1987): 672.

³ Rao, 205-208, and Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin, *Bayesian Data Analysis*, 2d ed (Boca Raton, FL: Chapman and Hall/CRC), 51-60, provide a more detailed discussion of these methods, which we apply in this appendix.

⁴ Paul E. Green and Daniel Blower, "Evaluation of the CSA 2010 Operational Model Test," FMCSA-RRA-11-019, August 2011, 43-48.

Specifically, assume that regulatory violation rates over a population of carriers are distributed as $\hat{\lambda}_1 \sim \text{Gamma}(\alpha, \beta)$, the prior distribution of the parameter of interest. Parameter values for the prior distribution can be assumed, based on historical data on the population of interest, or estimated using a particular sample. Conditional on these rates, the data on regulatory violations are distributed as $V_i | \lambda_i, t_i \sim \text{Poisson}(\lambda_i t_i)$, and the posterior distribution for a specific carrier is given by

$$\lambda_i | v_i, t_i \sim \text{Gamma}(\alpha + v_i, \beta + t_i) \quad (1)$$

Since the mean of a Gamma variate is α / β and the variance is α / β^2 , the posterior mean and variance of the rate for a given carrier are given by

$$E(\lambda_i | v_i, t_i) = (\alpha + v_i) / (\beta + t_i) \quad (2)$$

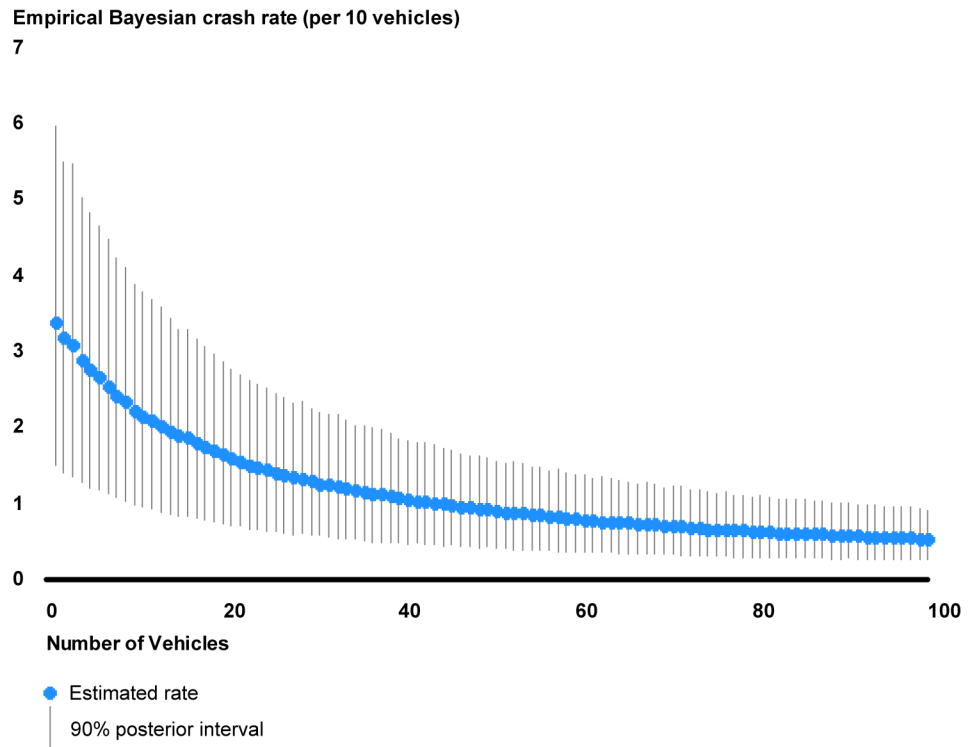
$$\text{Var}(\lambda_i | v_i, t_i) = (\alpha + v_i) / (\beta + t_i)^2 \quad (3)$$

The Bayesian rate estimate—the posterior mean—is a weighted average of the raw estimate for a specific carrier, v_i / t_i , and the mean of the prior distribution, α / β . When enough data are available, as indicated by a large exposure term relative to the violation term, the estimate converges to the ordinary, carrier-specific rate estimate. When exposure is low, however, the method combines data from the specific carrier with the mean rate for all carriers.

The variance of Bayesian rate estimates decreases with increased exposure, similar to the variance of ordinary rate estimates. Figure 6 shows how hypothetical rate estimates and 90% posterior intervals for a carrier that experienced 5 crashes vary with the carrier's exposure, as measured by the number of vehicles. (Although we illustrate rate estimation issues using crash rates, we likely would have obtained similar results if we had estimated regulatory violation rates.) As expected, the precision of the estimates decreases exponentially as the number of vehicles increases. The variance is high in the range of 1 to 5 vehicles and begins to decrease less quickly at approximately 20 vehicles, consistent with our discussion in the body of this report and prior evaluations of SMS.⁵

⁵ Green and Blower, 46-48. James Gimpel, "Statistical Issues in the Safety Measurement and Inspection of Motor Carriers," n.d.

Figure 6: Example of the Relationship between Exposure and the Precision of Rate Estimates



Source: GAO analysis of MCMIS data.

Thresholds in this approximate range are consistent with criteria used by the Centers for Disease Control and Prevention (CDC) to suppress or caveat rate estimates for the purpose of public display.⁶ For example, in its compendium of health statistics in the United States, CDC cautions that “[w]hen the number of events is small and the probability of such an event is small, considerable caution must be observed in interpreting the conditions described by the figures.”⁷

⁶ U.S. Cancer Statistics Working Group, *United States Cancer Statistics: 2004 Incidence and Mortality*. (Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, 2007), 10.

⁷ National Center for Health Statistics, *Health, United States, 2012: With Special Feature on Emergency Care*. (Hyattsville, MD: 2013), 10, 70.

Even though the Bayesian estimates do not converge to extremely low or high values when exposure is low, the uncertainty around the estimates remains high. As figure 6 shows, statistical methods for modeling and estimating rates can quantify this uncertainty explicitly, in order to reflect the varying precision of estimates for motor carriers with more or less observed data. Although the amount of uncertainty that is acceptable in practice depends on the purpose of the estimates, both statistical theory and government agencies estimating rates similar to those involved in the calculation of SMS scores have recognized the need to express the uncertainty of these estimates, particularly when the derived from small samples. This contrasts with FMCSA's approach, which reports SMS scores as safety risk estimates with no quantitative measures of precision.

Applying Rate Estimation Methods to Motor Carrier Data

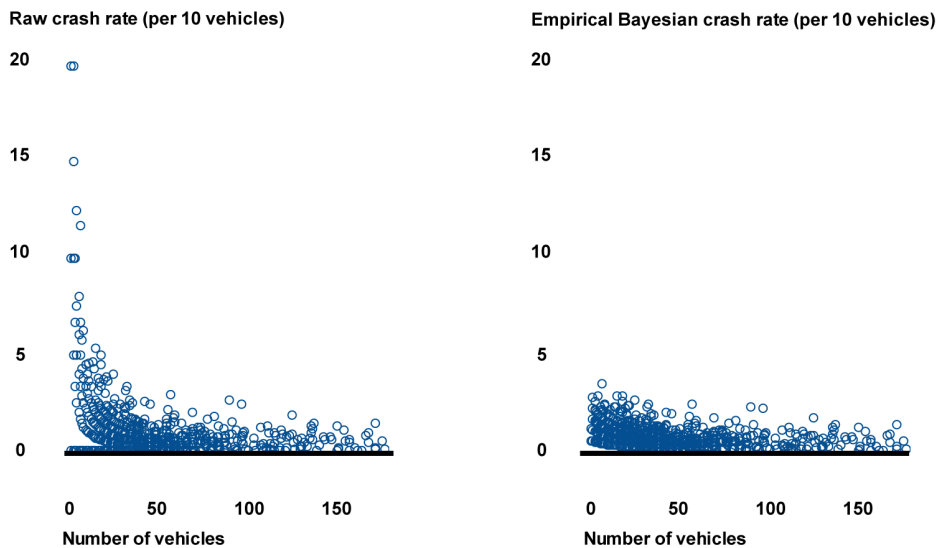
To illustrate the rate estimation issues discussed above in the context of motor carrier safety, we estimated individual crash rates for a population of motor carriers that were actively operating in each of two time periods, December 2007 through December 2009, and December 2009 through June 2011, as measured in FMCSA's Motor Carrier Management Information System (MCMIS). An "active" carrier was one that, in each time period, had at least one inspection or crash and had been recorded as a US-based interstate or intrastate Hazmat carrier. This definition resembled the one we used in replicating SMS, as described in the body of this report and appendix I. We obtained these data from the December 2010 and December 2012 MCMIS "snapshot" data files, as well as a historical file of carrier-specific information that covered all snapshots.

We estimated the raw and empirical Bayesian crash rates for each carrier in the first time period, using data on the number of crashes and vehicles for these carriers and the formulas above. We used the "empirical Bayes" version of the rate estimator, in which the parameters of the prior distribution were estimated from the data. Specifically, we fit the observed rate data for all carriers in the first time period to the negative binomial distribution, parameterized with exposure measured by number of vehicles, and estimated α and β using standard methods of maximum likelihood estimation. The final rate estimates for each carrier were a combination of these parameter estimates and carrier-specific data, according to equation 2 above.

As theory would predict, Bayesian methods prevented crash rates from converging to zero or extremely high values for carriers with low exposure. The left half of figure 7 presents the raw crash rates for our

analysis carriers, while the right half presents the empirical Bayesian estimates. The raw estimates for carriers with about 1 to 10 vehicles can be 10 to 20 times higher than for carriers with more than 10 vehicles. In addition, the raw rates cluster at zero for a large number of carriers, particularly for those with low exposure. An underlying crash rate of zero is implausible for active carriers. In contrast, the Bayesian rate estimates are more stable, with no inflation or deflation to extreme values. Since the body of this report finds that 93 percent of carriers in our replication of SMS had fewer than 20 vehicles, Bayesian methods may provide more stable estimates for many specific carriers and may better approximate the distribution of rates across carriers.

Figure 7: Relationships between Exposure and Rate Estimates for a Population of Motor Carriers Active from December 2007 through June 2011



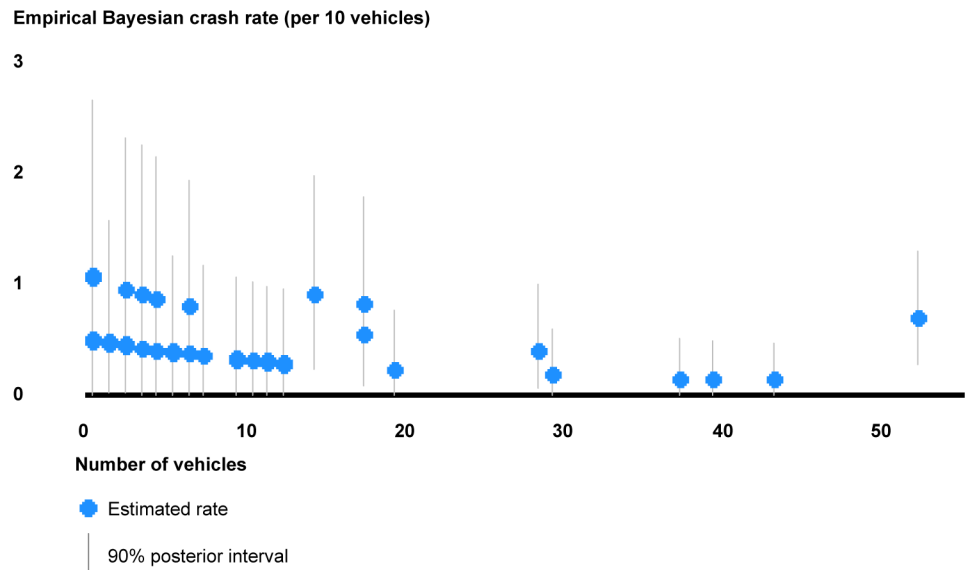
Source: GAO analysis of MCMIS data.

In addition to stabilizing rates for small carriers, Bayesian rate estimation methods provide an explicit measure of precision for each carrier's rate, regardless of size. In figure 8, we show the Bayesian rate estimates for a random sample of 109 carriers in the first period of our analysis population, along with 90 percent Bayesian posterior intervals.⁸ (We

⁸ Due to the small number of discrete counts for small carriers, the estimates for many of these carriers take the same values. As a result, the estimates overlap in the figure and may appear to involve a smaller number of carriers.

present these results for a sample to make the intervals readable.) The posterior interval expresses the range over which the true rate exists with a 90 percent probability. Consistent with theory, the precision of the rate estimates increases with exposure—in this case, the number of vehicles. These results apply to actual carriers in the sample, but the results are consistent with those expected by theory. The width of the posterior intervals does not decrease monotonically, however, because the relative number of crashes also affects the variance and is not held constant in the plot.

Figure 8: Examples of Empirical Bayes Rate Estimates for a Sample of Carriers Active from December 2007 through June 2011



Source: GAO analysis of MCMS data.

Appendix III: Evaluating the Statistical Validity of the Safety Measurement System

In this appendix, we express the Safety Management System (SMS) as a statistical measurement model, in order to make its assumptions explicit, and describe how estimating the model could validate those assumptions. We find that FMCSA's SMS makes a number of strong assumptions about motor carrier safety that empirical data cannot easily validate.

The SMS uses administrative data on inspections of commercial motor carriers, violations of regulations, and crashes to measure carrier safety. Statisticians and other researchers have developed methods to validate measures of such broad concepts as safety, referred to as "latent variables," using empirical data.¹ These methods are known as "measurement models." For example, mental health professionals have created scales to measure the existence of broad disorders, such as depression, by combining responses to multiple items on patient questionnaires. SMS has a similar goal: to create scales to measure motor carrier safety risk on several dimensions, such as "Unsafe Driving" or "Vehicle Maintenance," by combining violation rate data across multiple regulations. Latent variable measurement methods can assess whether these broader measures are valid and reliable, and whether the empirical indicators that go into them actually measure the intended concepts. Estimating the degree to which various indicators measure a broader concept helps confirm and often improve the reliability and validity of the scales constructed.

Structure and Assumptions of SMS

Much of the SMS involves calculating weighted regulatory violation rates for motor carriers in a given time period.² FMCSA assigns weights that, in principle, reflect the violations' associations with one of six dimensions of safety, known as Behavioral Analysis and Safety Improvement Categories (BASICS), such as "Unsafe Driving" and "Vehicle Maintenance."³ The weights represent what FMCSA considers to be the

¹ For example, see Kenneth A. Bollen, *Structural Equations with Latent Variables* (New York: John Wiley and Sons, 1989).

² FMCSA refers to these as "measures."

³ The other BASICS that reflect regulatory violations include "Controlled Substances/Alcohol," "Driver Fitness," "Fatigued Driving (Hours of Service)," and "Hazardous Materials." A seventh BASIC measures crash history. Portions of this appendix do not apply to the crash history BASIC, because it is not a function of regulatory violation rates.

strength of each violation's association with safety, relative to other violations in the same BASIC. All violations that are categorized in a BASIC get a positive weight ranging from 1 to 10, which implies that they have some association with safety. These weighted violation rates strongly influence the final SMS measures of safety on these dimensions. Each BASIC is linked to a set of violations, which are all assumed to measure the same dimension of safety. Each violation maps to exactly one BASIC, though BASICs map to multiple violations in their associated groups.⁴

For a carrier i , the violation rates influencing scores in each of the $p = 1, 2, \dots, 6$ BASICs can be expressed as

$$R_{ip} = \frac{\sum_j^k \lambda_j V_{ij}}{T_i}.$$

V_{ij} measures the number of times that carrier i violated regulation j in a given time period. λ_j is a weight for each violation. It is the product of a "severity" weight, measuring what FMCSA considers the violation's "crash risk relative to the other violations comprising the BASIC measurement," in addition to outcomes thought to be particularly severe (e.g., out-of-service violations), and a time weight, measuring what FMCSA considers the importance of violations from different time periods to estimating a carrier's current level of safety. By defining V_{ij} for fixed time periods, such as 6 or 12 months prior to the measurement time, we collapse the separate weights used in SMS into λ_j , in order to simplify the notation. Lastly, T_i measures exposure to committing violations in the time period, which is either a function of carrier's vehicles and vehicle miles traveled (VMT) or the time-weighted sum of relevant inspections, depending on the BASIC.⁵

SMS transforms the weighted violation rates for each carrier into percentile ranks, after applying a number of "data sufficiency standards" to exclude carriers with few violations, inspections, and/or vehicles. Carriers with percentiles that exceed established thresholds are "alerted"

⁴ For a detailed specification of SMS, see John A. Volpe National Transportation Systems Center, CSA, Carrier Safety Measurement System (CSMS) Methodology: Version 3.0—Motor Carrier Preview, (August 2012).

⁵ FMCSA refers to vehicles as "power units."

on the relevant BASICs and, if enough alerts or other conditions exist, are identified as “high risk.” As a result, the ultimate measures of safety risk are ordered groups, with cut-points defined by BASIC percentiles for carriers that meet FMCSA’s standards for data sufficiency.

SMS as a Latent Variable Measurement Model

The SMS can be viewed as an attempt to measure latent concepts of “safety,” such as “Unsafe Driving” or “Vehicle Maintenance,” using observed data on regulatory violations and the opportunity to commit them (exposure). Consider the latent variable measurement model below, using notation from a prominent textbook⁶:

$$r = \Lambda \xi + \delta$$

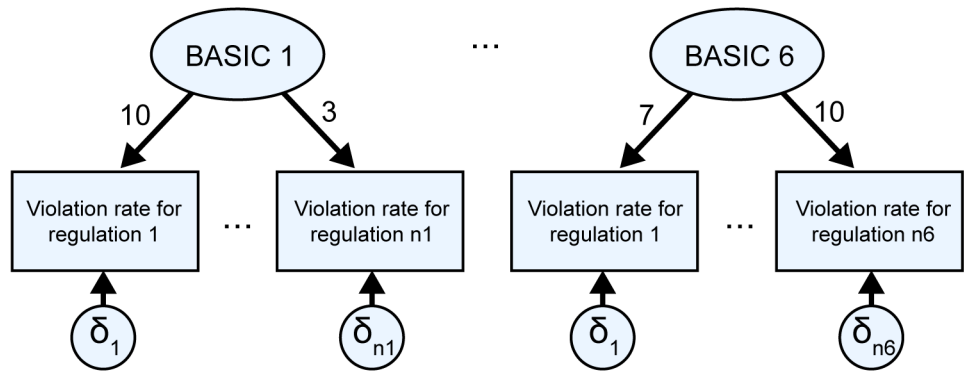
$$r = \begin{bmatrix} R_{11} \\ \vdots \\ R_{n_11} \\ R_{12} \\ \vdots \\ R_{n_22} \\ \vdots \\ R_{1p} \\ \vdots \\ R_{n_pp} \end{bmatrix} \quad \Lambda = \begin{bmatrix} \lambda_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ \lambda_{n_11} & 0 & \dots & 0 \\ 0 & \lambda_{12} & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \lambda_{n_22} & \dots & \lambda_{1p} \\ 0 & 0 & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_{n_pp} \end{bmatrix} \quad \xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_p \end{bmatrix} \quad \delta = \begin{bmatrix} \delta_{11} \\ \vdots \\ \delta_{n_11} \\ \delta_{12} \\ \vdots \\ \delta_{n_22} \\ \vdots \\ \delta_{1p} \\ \vdots \\ \delta_{n_pp} \end{bmatrix}$$

The model assumes that a vector of $k = \sum_g^p n_g$ observed variables, r , are determined by p latent variables, ξ , and random measurement error, δ . The weights describing the relationship between the latent and observed variables make up the block diagonal matrix Λ , with p blocks of weights applied to the corresponding blocks of observed variables. This structure implies that each group of observed variables is related to exactly one latent variable. In many applications, the model assumes that $\text{Cov}(\xi, \delta) = 0$ and $E(\delta) = 0$ but allows other variances and covariances to be estimated from the data as parameters or fixed to known values.

⁶ Kenneth A. Bollen, *Structural Equations with Latent Variables* (New York: John Wiley and Sons, 1989), 16-18, 179-225.

The SMS is a particular form of the model above. Specifically, SMS defines r as violation rates for $k = 826$ regulations, where r may include variables measured at different times. It sets $p = 6$ and relates the violation rates to the BASICs, or latent variables ξ measuring safety, through the weighting matrix Λ . FMCSA created fixed time and severity weights for each regulation through a combination of statistical analysis and the opinions of stakeholders.⁷ Since SMS is not a stochastic model, it assumes that $\delta = 0$. A graphical version of SMS as a measurement model appears in figure 9 below.⁸

Figure 9: SMS as a Measurement Model



Source: GAO analysis of SMS methodology.

When expressed as a measurement model, the strong assumptions of SMS—and their potential detrimental effect on its usefulness—become clear. FMCSA’s assumption of zero measurement error is unusual for statistical approaches to measurement, given that any particular violation is likely to represent variation in latent variables (in this case, safety) as well as unmeasured variables summarized by the error term. SMS makes specific assumptions about the number of safety dimensions—the latent variables assumed by the model above—as well as their relationships to

⁷ Volpe National Transportation Systems Center, A-1 – A-2.

⁸ SMS expresses the latent measurement of safety as a weighted function of violation rates, rather than the reverse, as in the model above. The difference is immaterial, because we can express the SMS model as $\Lambda^{-1}r = \xi$. If the weights in Λ were measured on the same scale as in the SMS, one could use the weights in Λ^{-1} to express the latent safety variables as linear combinations of violation rates, as in the SMS.

violation rates. Exactly six dimensions of safety exist (involving regulations), and each violation rate measures only one of them. In other efforts to measure broad concepts using numerous indicators, inference about the existence and relationships among observed and latent variables are endogenous parameters (determined by the model) to be estimated, rather than exogenous parameters (determined outside the model) that are fixed *ex ante*, ahead of time, as they are here. Finally, SMS takes the unusual step of fixing the values of the weights relating the latent variables measuring safety to violation rates at values other than 0. This assumes a high degree of prior knowledge about the relationships between latent and observed variables. Although FMCSA has conducted several studies of how regulatory violation rates are associated with crash risk, these studies do not directly estimate the degree to which each type of violation reflects one of several dimensions of safety.

One approach to validating the assumptions of SMS is to estimate the parameters of the measurement model above using empirical data on regulatory violation rates. This approach is known as Confirmatory Factor Analysis, which is a special type of measurement model. Because SMS makes specific assumptions about the number of BASICs and the violations that go into them, we can express the system as a measurement model, as discussed above, and estimate the degree to which its assumptions are consistent with reality. For example, SMS assumes that six dimensions of safety exist—labeled BASICs in SMS—and that each violation reflects only one dimension. However, a model that assumes three BASICs and allows violations to reflect multiple dimensions of safety might be a plausible alternative. High violation rates for brake maintenance regulations may indicate worse performance on both the Vehicle Maintenance and Unsafe Driving dimensions of safety. Measurement modeling can identify which of these approaches better fits empirical patterns of regulatory violations. More generally, analyzing SMS as a measurement model can validate its assumptions, such as the values of the severity and time weights, and suggest improvements to better measure safety.

We can extend the SMS measurement model to predict empirical data on crash risk, in order to further validate its ability to identify high-risk carriers. This structural equation modeling (SEM) approach combines the measurement model above with a model that describes how the latent dimensions of safety predict crash risk, generically known as “endogenous observed variables.”

To incorporate outcomes, we extend the measurement model above to assume that the six BASICs are directly related to an empirical measure of crash risk:

$$C_i = \gamma\xi + \varepsilon_i$$

C_i measures crash risk; γ are parameters describing how the latent safety dimensions are related to crash risk; ξ are the safety dimensions; and ε_i is a random error term. Estimating this larger model would yield the original parameters of the measurement model, in addition to the parameters describing how the SMS scores relate to crash risk, γ . Strong correlations between SMS scores and crash risk would further support their ability to identify higher-risk carriers. This is known as “criterion validity” in statistics and social research.

A key strength of this validation approach is that it accounts for the error in measuring broad dimensions of safety when predicting crash risk. Because empirical data on violation rates and SMS scores are indicators of latent concepts of safety, measurement error can distort the underlying relationships between these broader concepts and crash risk. For example, poor vehicle maintenance may be positively associated with higher crash risk, but empirical data on violations of vehicle maintenance regulations may measure both the concept of interest and the enforcement efforts of state and local governments. As a result, the violation rates may be uncorrelated with crash risk simply due to error in measuring the concept of interest. SEM models estimate the relationships among latent variables more precisely by accounting for this measurement error. This contrasts with simpler regression models of crash risk as a function of observed violation rates, which assume that violation rates measure the dimensions of safety without error.

Appendix IV: Prior Evaluations of SMS Scores as Measures of Safety for Specific Carriers and Risk Groups

Previous evaluations of SMS have focused on estimating the correlations between crash risk and regulatory violation rates and Safety Measurement System (SMS) scores. These evaluations have found mixed evidence that SMS scores predict crash risk with a high degree of precision for specific carriers or groups of carriers. This appendix synthesizes the results of these prior evaluations.

Several prior evaluations of SMS have analyzed grouped data, rather than directly analyzing how a carrier's individual regulatory violation rates and SMS scores predict its own future crash risk. For example, in a pilot evaluation conducted for FMCSA, the University of Michigan Transportation Research Institute (UMTRI) estimated group crash rates within percentiles of SMS scores for each Behavioral Analysis and Safety Improvement Category (BASIC), pooling several hundred carriers in each percentile, to trace out the aggregate relationship between SMS scores and crash risk.¹ Similarly, FMCSA's Violation Severity Assessment Study analyzed grouped violation data from roadside inspections conducted from 2003 through 2006, in order to compare rates cited in post-crash reports to rates in the general population of carriers.²

Aggregation addresses a key statistical obstacle to validating SMS: a large proportion of regulations are violated too infrequently to have enough meaningful variation across carriers for analysis. Even after aggregating 4 years of carrier-level data, the Violation Severity Assessment Study had insufficient data—which the study defined as less than 10 inspections—to estimate the association between crash risk and 69 to 73 percent of the violations available to the authors, depending whether the analysis considered crash severity.³ The study noted that many regulations were “not being cited” or not “being cited at a sufficient rate to meet the study's data sufficiency requirements.”⁴ Evaluations conducted by FMCSA, known as “SMS Effectiveness Testing,” have taken similar approaches, calculating aggregate crash rates for carriers

¹ Paul E. Green and Daniel Blower, *Evaluation of the CSA 2010 Operational Model Test*, FMCSA-RRA-11-019 (August 2011, 40-43).

² John A. Volpe National Transportation Systems Center, *Violations Severity Assessment Study: Final Report* (October 2008, 3-3 – 3-4).

³ *Ibid.*, 4-2, 4-6.

⁴ *Ibid.*, 5-1.

that did and did not exceed the SMS thresholds to be placed in “alert” or “high risk” statuses.

Aggregate approaches, such as those used in several prior evaluations, do not directly assess the ability of SMS and regulatory violations to predict future crash risk for specific carriers. Well-known findings in statistics on “ecological fallacies” show that associations at higher levels of analysis are not guaranteed to exist at lower levels of analysis.⁵ In this application, carriers that crash may have higher violation rates or SMS scores as a group than carriers that do not crash, but this pattern does not necessarily apply to specific carriers within the groups. Because less variation exists at the carrier level, aggregation can overstate the strength and precision of these correlations for individual carriers.

Even when similar correlations exist at the carrier level, comparing average crash rates for SMS percentiles or risk groups does not assess the prediction error for any particular carrier. The average crash rate may be higher for groups of carriers with increasingly high SMS percentiles, but crash rates may vary significantly around these means. This residual variation, not differences in means or other aggregate statistics, is more directly relevant for assessing the quality of predicted crash rates for a particular carrier. In statistical terms, the prediction error summarized by the residual variance of a linear regression model or the classification matrix of a categorical model is what matters for assessing predictive power for individual carriers, not the models’ coefficients, which estimate mean crash rates conditional on these percentiles.

Thus, it is not surprising that previous evaluations of carrier-level data have found weaker relationships between crash risk and SMS scores and regulatory violations than have the evaluations of aggregated data.

UMTRI estimated the relationship between exceeding thresholds in the six non-crash BASICs and mean crash rates, using an empirical Bayesian negative binomial model estimated on carrier-level data. The results showed that carriers exceeding the thresholds for the Unsafe Driving and Vehicle Maintenance BASICs had average crash rates that were 1.1 to

⁵ W.S. Robinson, “Ecological Correlations and the Behavior of Individuals,” *American Sociological Review*, vol. 15, no. 3 (June 1950): 351-357. David A. Freedman, “Ecological Inference and the Ecological Fallacy,” *International Encyclopedia of the Social and Behavioral Sciences*, Technical Report 549 (October 1999).

1.8 times higher than carriers not exceeding the thresholds⁶—usually lower than the rate ratios of 1.0 to 5.4 reported by UMTRI’s aggregate analysis and FMCSA’s December 2012 Effectiveness Testing.⁷ However, this relationship was negative for the Driver Fitness and Loading/Cargo (currently Hazardous Materials) BASICS, with mean crash rates for alerted carriers that were 0.85 and 0.91 times the rates of non-alerted carriers, respectively. The ratios were not significantly greater than 1 for the Fatigued Driving and Substance Abuse/Alcohol BASICS.⁸ Similarly, the American Transportation Research Institute (ATRI) found that alerted carriers in the Unsafe Driving, Vehicle Maintenance, Hours-of-Service, and Controlled Substances/Alcohol BASICS had mean crash rates that were 1.3 to 1.7 times larger than scored carriers not in alert status, but carriers exceeding the Driver Fitness thresholds had mean crash rates that were 0.87 times those of non-alert scored carriers.⁹

Although UMTRI and ATRI analyzed carrier-level data, they validated SMS measures using regression coefficients and similar statistics that describe aggregate correlations. As we discuss above, this approach does not directly quantify predictive power for specific carriers.

Two studies that have directly estimated prediction error for specific carriers, conducted by Wells Fargo Securities and James Gimpel of the University of Maryland, found weaker evidence of the model’s predictive effectiveness. Gimpel found that mean crash rates increased by small amounts as SMS scores increased on the Unsafe Driving, Hours-of-Service, and Vehicle Maintenance BASICS increased.¹⁰ Wells Fargo found a similarly positive association for the Unsafe Driving BASIC, but a

⁶ Green and Blower, 47. We calculated these results using the negative binomial regression coefficient estimates for carriers with SMS scores that did and did not exceed the BASIC thresholds, as reported by UMTRI.

⁷ *Ibid.*, 31, 34.

⁸ *Ibid.*, 47.

⁹ Micah D. Lueck (American Transportation Research Institute), “Compliance, Safety, Accountability: Analyzing the Relationship of Scores to Crash Risk,” (October 2012) 19-24.

¹⁰ James Gimpel, “Statistical Issues in the Safety Measurement and Inspection of Motor Carriers,” 3-9. This study was commissioned by The Alliance for Safe, Efficient, and Competitive Truck Transportation, which is currently in litigation with FMCSA over the public use of SMS data.

negative association for the Hours-of-Service BASIC, in its analysis of 4,600 carriers with at least 25 vehicles and 50 inspections.¹¹ More critically, the authors showed that scores on these BASICs predict crash rates with a large amount of error, with most R-squared fit statistics ranging from nearly zero to 0.07 for reasonably large analysis samples.¹² Although these studies do not report critical estimates of the residual variance, the R-squared statistics likely imply confidence intervals around predicted crash rates for individual carriers with widths that are several times larger than the predictions themselves. This implies that SMS scores predict future crash risk for specific carriers with substantial error, even though mean crash rates can be higher among carriers with higher SMS scores.

FMCSA used aggregate data to dispute the findings of the Wells Fargo evaluation. Specifically, the agency cited the UMTRI findings that aggregate crash rates were 3.0 to 3.6 times higher for carriers exceeding thresholds for the Unsafe Driving and Hours-of-Service BASICs than for carriers that did not exceed thresholds for any BASIC.¹³ In addition, FMCSA highlighted analyses by UMTRI and the Volpe Center of aggregate crash rates across percentiles of SMS scores in the Unsafe and Fatigued Driving BASICs, respectively, which they claimed to show a stronger correlation to crash risk.¹⁴ FMCSA's approach to evaluating the predictive power of SMS scores resembles its Effectiveness Testing, which compares aggregate crash rates for carriers above and below thresholds for various BASICs.

However, as we discuss above and Wells Fargo discussed in its response to FMCSA, the fact that SMS scores predict aggregate crash rates more strongly at the alert-group or percentile level does not necessarily imply that the scores will predict the crash risk of individual carriers. Recognizing this, the UMTRI evaluation analyzes the data at both the aggregate and carrier levels, and finds that mean crash rate

¹¹ Anthony P. Gallo and Michael Busche (Wells Fargo Securities), "CSA: Another Look with Similar Conclusions," (July 2, 2012) 10-13, 17.

¹² Gimpel finds slightly larger R-squared statistics for one subsample of carriers.

¹³ FMCSA, "Review of Wells Fargo Equity Research Report on Compliance, Safety, Accountability" (March 16, 2012) 2, 9.

¹⁴ *Ibid.*, 6-9.

ratios are far smaller at the carrier level than at the alert-group or percentile levels. It should be intuitive that aggregate evidence of effectiveness, stressed in some FMCSA evaluations, shows stronger predictive power than the carrier-level analyses of ATRI, Gimpel, UMTRI, and Wells Fargo. Aggregating violation and crash rates within larger groups effectively increases the sample size used to calculate rates, which reduces their sampling error when compared to the equivalent carrier-level measures. The reduction of sampling error can strengthen the correlations between violation rates and SMS scores and crash risk.¹⁵

Evaluations of SMS that focus on carrier-level prediction error provide the most appropriate evidence of effectiveness for assessing the safety of individual carriers. FMCSA has stated that one purpose for SMS scores is to predict the future crash risk of individual motor carriers, in order to prioritize resources for intervention and enforcement. In addition, FMCSA reports SMS scores as measures of safety on a public website and the SaferBus Mobile app. To assess the validity of SMS scores for this purpose, evaluations should focus on the system's ability to predict the crash risk at the carrier level, not its ability to identify groups of carriers with larger crash rates on average or collectively. Measures of predictive accuracy—such as the residual error made when predicting crash rates or the classification error made when assigning carriers to risk groups—are the critical metrics of success, not aggregated crash rate ratios and regression coefficients. When evaluated on these criteria, prior studies show that SMS predicts future crash risk for individual carriers with substantial imprecision.

None of the prior studies has explicitly incorporated measurement error into evaluations of SMS. Since SMS is ultimately a method of creating measures of latent variables, as we discuss in appendix III, the regulations used to calculate scores and the scores themselves have some degree of measurement error. Because existing studies have used statistical methods that assume zero measurement error, more comprehensive attempts to model the measurement structure of SMS and validate its assumptions and predictive power, such as those we discuss in appendix III, may produce different results. The correlations among SMS scores, violation rates, and crash risk may reflect

¹⁵ Kenneth A. Bollen, *Structural Equations with Latent Variables* (New York: John Wiley and Sons, 1989) 154-156.

**Appendix IV: Prior Evaluations of SMS Scores
as Measures of Safety for Specific Carriers and
Risk Groups**

measurement error as much as the underlying relationships among the variables of interest. This more complex analysis is critical for future evaluations of SMS and its ability to measure safety risk.

Appendix V: Analysis of Regulatory Violations and Crash Risk

As a more basic approach to validating SMS, which focuses on the ability of data on regulatory violations in one time period to predict crash risk in a subsequent period, we analyzed the relationship between violation rates and crash risk using a series of statistical models. These models predicted the probability of a crash and crash rates as a function of regulatory violation rates for a population of motor carriers that were actively operating over a recent 3.5-year time period (described below).

We find that a substantial portion of regulatory violations in SMS cannot be empirically linked to crash risk for individual carriers. Consistent with prior research,¹ about 160 of the 754 regulations with data available in this time period had sufficient variation across carriers for analysis. Of the approximately 160 regulations with sufficient violation data, less than 14 were consistently associated with crash risk, across statistical models. These results suggest that the specific weights that SMS assigns to many regulations when calculating safety risk cannot be directly validated with empirical data, and many of the remaining regulations do not have meaningful associations with crash risk at the carrier level.

Data and Methods

We assembled data for a population of motor carriers using the MCMIS snapshot files dated December 2010 and 2012. Specifically, we identified carriers that were actively operating in each of two time periods: from December 2007 through December 2009 (the “pre-period”) and from December 2009 through June 2011 (the “post-period”). We defined an active carrier as one that is as outlined in Appendix I, consistent with FMCSA’s definition of active carriers for its Effectiveness Testing and other analyses. For each of the approximately 315,000 carriers that met these criteria, we extracted data on the number of regulatory violations and crashes incurred in each time period, along with the number of inspections, vehicles, and use of straight versus combo trucks, among other variables, from the crash and inspection tables in MCMIS.

The goal of our analysis was to predict crash risk in the post-period, using data on regulatory violations, crash data, and carrier characteristics

¹ For example, even after aggregating four years of data, the Violation Severity Assessment Study sites insufficient data to estimate the association between crash risk and about 70 percent of the violations available to the authors. John A. Volpe National Transportation Systems Center, *Violations Severity Assessment Study: Final Report* (October 2008) 4-2.

measured in the pre-period. We developed a series of linear and generalized linear regression models to predict two measures of crash risk for individual carriers: a binary indicator for having crashed in the post-period and the ratio of crashes to vehicles. Estimating and evaluating all potential models and model types was not the goal of these analyses. Rather, we sought to estimate the associations between regulatory violation rates and crash risk at the carrier level, in order to validate the violations' severity weights in SMS.

We reduced the list of 754 regulations whose violations are tracked in MCMIS to those that had enough variation across carriers for analysis. After excluding 593 violations that had zero variance or zero counts for more than 99 percent of the analysis carriers, we retained data on the violation of approximately 160 regulations for use in predicting crash risk.

As we discuss in appendix II and the body of this report, crash and violation rates based on small exposure measures, generally resulting from carriers with few vehicles, may be estimated with less precision than rates based on larger exposure measures. To better understand and attempt to overcome these rate estimation issues and assess the sensitivity of our results, we used both ordinary and empirical Bayesian estimators of crash and violation rates.² In addition, we estimated separate models limited to carriers that had more than 20 vehicles.

These methodological choices produced 8 groups of models, as described in table 7. The groups were defined by the combined categories of crash measure (binary crash status versus Bayesian crash rate), methods of violation rate estimation (ordinary versus Bayesian), and carrier size (full data or restricted to more than 20 vehicles). These parallel analyses allowed us to assess the sensitivity of our results to different assumptions.

² The potentially rare occurrence of crashes and violations may also contribute to higher variability of crash and violation rates since for small carriers, the effect of small exposure is being compounded by a rare event.

Table 7: Model Groups Based on Crash Status Measure, Violation Rate Measure, and Carrier Size Restrictions

Model group	Crash status	Violation rate	Model building data
1	Crash status (yes/no)	Observed	Restricted to carriers with >20 vehicles
2	Crash status (yes/no)	Bayesian	Restricted to carriers with >20 vehicles
3	Crash status (yes/no)	Observed	Full carrier sample
4	Crash status (yes/no)	Bayesian	Full carrier sample
5	Bayesian crash rate	Observed	Restricted to carriers with >20 vehicles
6	Bayesian crash rate	Bayesian	Restricted to carriers with >20 vehicles
7	Bayesian crash rate	Observed	Full carrier sample
8	Bayesian crash rate	Bayesian	Full carrier sample

Source: GAO.

For each of the eight model groups, we include three sets of covariates to predict crash risk in the post-period:

- “Simple model:” indicator (binary) for crashing in the pre-period, carrier size, and carrier type (percent straight versus combo).
- “Full model:” predictors in the simple model, plus all violation rates with viable data in the pre-period.
- “Stepwise full model:” We applied a stepwise selection algorithm applied to all predictors in the “full model,” in order to select the most predictive covariates. The algorithm’s constraints required a p-value of 0.30 for a covariate to enter the model and 0.35 to remain in the model.

To avoid over-fitting our models to any particular sample of data, we divided our data using a random method to form a model-building sample and a validation sample. We used the model-building sample to estimate the models described above and the validation sample to assess the accuracy of the model’s predictions of crash probability against new data. When seeking to develop statistical methods for predictive purposes, this type of out-of-sample validation is extremely useful to ensure that any method identified can consistently predict well on all samples of data, not just the sample that was used to develop the method. This is an important limitation of prior evaluations of SMS, which, to our knowledge, have not used replication samples to avoid over-fitting when identifying predictive violation types or methods of identifying higher-risk carriers.

Model selection required addressing statistical estimation issues, such as instability of the parameter estimates caused by co-linearity of predictors or lack of variability in the predictors, and other model fitting concerns.

For the linear crash rate models, the dependent variable required a log transformation to remove non-constant error variance, which would invalidate results if left untreated. These statistical issues resulted in sub-models within the major model groups that were explored until a stable model resulted. Therefore, the results within each model group focus on three sub models, when applicable: simple, stepwise and full, where stepwise is the model that eliminated independent variables until a stabilized model with estimable coefficients resulted. See table 8 for the final list of 30 models and subsamples.

Table 8: A list of Sub-Model Descriptions according to Data Restrictions (Restricted to Data for Carriers with Greater Than 20 Vehicles versus Full Data with All Carriers), Violation Rates (Observed versus Bayesian), and Sample (Model Building versus Validation)

Sample	Carrier vehicle restriction			
	Restricted (carriers with more than 20 vehicles)		All carriers	
	Violation rates		Violation rates	
	Observed	Bayesian	Observed	Bayesian
Model building sample for crash (yes/no)	1. Simple	n/a	6. Simple	n/a
	2. Stepwise	4. Stepwise	7. Stepwise	9. Stepwise
	3. Full	5. Full	8. Full	10. Full
Validation sample for crash (yes/no)	11. Simple	n/a	16. Simple	n/a
	12. Stepwise	14. Stepwise	17. Stepwise	19. Stepwise
	13. Full	15. Full	18. Full	20. Full
Model building sample for Bayesian crash rate	21. Simple	n/a	26. Simple	n/a
	22. Stepwise	24. Stepwise	27. Stepwise	29. Stepwise
	23. Full	25. Full	28. Full	30. Full

Source: GAO.

Notes: The simple models do not include violation rates inputs and thus the observed and Bayesian models produce the same results.

Model groups 1 through 4 in table 7 are represented by sub-models 1 through 20; Model groups 5 through 8 in table 7 are represented by sub-model groups 21 through 30.

Evaluation of Models

Models that use the SMS violation information do not fit well according to various measures discussed below. In addition, the violation rates, as measured in SMS, do not have a strong predictive relationship with crashes, regardless of whether the observed or the Bayesian violation rates are used as inputs.

Models for crash status (yes/no) were examined for stability of parameter estimates, fit statistics,³ number and types of violations that were predictive and that were stable,⁴ and future predictive performance according to these measures. Models for Bayesian crash rates were examined for stability of parameter estimates, fit statistics, number and types of violations that were predictive, predictive power and future predictive power. Some of the diagnostics cannot be compared in absolute terms, but rather should be compared across models fit to the same data. For example, the AIC must be compared across competing models fit on the same data.

The crash status (yes/no) model was evaluated in the out-of-sample validation data, where each model was re-fit on the validation sample, and the diagnostics were examined and compared to those from the model-building sample. As an additional sensitivity analysis, the same set of inputs for each of the model groups one through four were also fit using a Bayesian crash rate outcome, via a linear regression fit to the model-building sample. Results were compared.

Model Results

Since diagnostics will differ according to the outcome measure, crash status (yes/no) versus crash rate, information for these outcome types is displayed separately. For results of models for the crash status (yes/no), see tables 9 and 10. For results for the Bayesian crash rates, see table 11. Given that a high value of the H-L p-value (close to 1) indicates good model fit, according to this measure, most of the models fail to fit acceptably, and none of the models fit well.

Within the same data, a lower value of the AIC indicates better fit; therefore, the stepwise models perform best, and do nearly as well regarding the ROC and generalized R-squared when compared to the

³ For logistic regression models with post-crash status (yes/no), fit measures included: Hosmer-Lemeshow (H-L) p-value, AIC, percent concordant/discordant, area under the ROC (receiver operating characteristic) curve and classification rates— true positive (sensitivity), true negative (specificity), false positive, false negative. For a linear regression model with post-crash rates, fit measures included R-squared, AIC (Akaike information criterion), Mallows' Cp, and regression diagnostic plots.

⁴ For models, a flag was created after the models were finalized to define the number of unstable effects based on the coefficient of variation (cv). Define the cv as the standard error of an estimate divided by the estimate.

more complicated full model. But even for the stepwise models, the ROC and R-squared do not indicate a strong predictive relationship. This finding is echoed by the number of effects in the model, relative to the number of potential violations (about 160) and the number of stable effects.

Table 9: Logistic Regression Results for Sub-Models Simple, Stepwise, and Full-of-Outcome Crash Status (Yes/No); Note That the Simple Model Is Redundant for Model Groups 2 and 4 Since No Violation Rates Are Included in the Simple Model

Model group	Model group description: Crash Vio rate Data	Sub-Model description	AIC	H-L Pvalue	Percentage concordant	Percentage discordant	R-squared	ROC	Number of covariate effects in model	Number of stable covariate effects as defined in footnote 4.
1	<ul style="list-style-type: none"> Crash status (yes/no) Observed vio rate Restricted 	1. Simple	4,105	0.004	77.3	21.2	0.205	0.781	7	7
		2. Stepwise	3,782	0.022	81.8	18.0	0.264	0.819	72	50
		3. Full	3,945	0.190	82.2	17.6	0.269	0.823	169	46
2	<ul style="list-style-type: none"> Crash status (yes/no) Bayesian vio rate Restricted 	4. Stepwise	3,723	0.008	81.7	18.2	0.261	0.817	37	24
		5. Full	3,883	0.005	82.9	17.0	0.281	0.829	168	25
3	<ul style="list-style-type: none"> Crash status (yes/no) Observed vio rate Full 	6. Simple	41,059	<0.001	70.0	19.6	0.158	0.752	9	8
		7. Stepwise	36,628	<0.001	76.9	22.6	0.177	0.771	81	63
		8. Full	36,784	<0.001	77.0	22.6	0.177	0.772	171	61
4	<ul style="list-style-type: none"> Crash status (yes/no) Bayesian vio rate Full 	9. Stepwise	36,155	0.420	76.9	22.6	0.184	0.771	47	37
		10. Full	36,287	0.154	77.0	22.5	0.187	0.772	170	40

Source: GAO analysis of FMCSA data.

One aspect of predictive power is the ability for a model to discriminate the observed outcomes based on model predictions. Classification tables describe a model's classification accuracy with correct and incorrect classifications, as measured by sensitivity (correctly predict an event) and specificity (correctly predict a non-event), and false positive (incorrectly predict a non-event) and negative rates (incorrectly predict an event).

Classification tables for the simple, full, and stepwise model within a model group are presented in table 10. The observed proportion of crashes, approximately 0.2 for the unrestricted data and 0.66 for the data restricted to carriers with more than 20 vehicles, is used as the cut-point to classify predicted probabilities for a carrier into a predicted event (crash) versus non-event (no crash). The predicted crash status for a particular model is compared to the actual post-crash status, resulting in a series of table rows, one for each model, that examine the false positives, false negatives, and other quantities that help evaluate the predictive quality of a model.

For unrestricted data, the false negative rate (or the rate that results from incorrectly classifying a carrier to a non-alert status), is relatively low (around 11 percent) compared to the false positive rate (ranges from about 56 to 58 percent). This is a desired result if it is considered more appropriate to be conservative and put a carrier in alert status, even if that alert status is incorrect (false positive), compared to misclassifying a carrier into non-alert when an alert would be called for (false negative). The restricted data have a higher false negative rate (from 42 to 44 percent) than false positive rate (around 14 to 19 percent), and this false negative rate is also higher than the full data false negative rate. For the restricted data with higher false negative rates, this means a higher percentage of carriers are being classified in non-alert when they have crashed than the percent classified as alert, but that did not crash, and such a scenario is not desirable under a conservative preference toward low false negative rates. In addition, the sensitivity and specificity are both moderate at best within data (restricted versus full), further evidence of the inability for models to discriminate.

Table 10: Classification of Predicted Values from Models for the Crash-Status (Yes/No) Using the Average Observed Predicted Rate as the Cut-Point, Based on the Model-Building Sample

Model Group	Crash Vio Rate Data	Sub-Model description	Correct events	Correct nonevents	Incorrect events	Incorrect nonevents	Percent correct	Sensitivity	Specificity	False positive	False negative	
1	• Crash status (yes/no)	1. Simple	1,888	897	434	647	72.0	74.5	67.4	18.7	41.9	
		2. Stepwise	1,836	898	363	645	73.1	74.0	71.2	16.5	41.8	
		3. Full	1,824	859	402	657	71.7	73.5	68.1	18.1	43.3	
2	• Observed vio rate	• Restricted	4. Stepwise	1,763	979	282	718	73.3	71.1	77.6	13.8	42.3
			5. Full	1,724	955	306	757	71.6	69.5	75.7	15.1	44.2
			6. Simple	5,905	31,455	8,100	3,994	75.5	59.7	79.5	57.8	11.3
3	• Bayesian vio rate	• Restricted	7. Stepwise	6,008	25,599	8,308	3,245	73.2	64.9	75.5	58.0	11.3
			8. Full	5,996	25,548	8,359	3,257	73.1	64.8	75.3	58.2	11.3
			9. Stepwise	5,846	26,382	7,525	3,407	74.7	63.2	77.8	56.3	11.4
4	• Full	• Full	10. Full	5,823	26,429	7,478	3,430	74.7	62.9	77.9	56.2	11.5

Source: GAO analysis of FMCSA data.

To address whether crash status (yes/no) has a different relationship with violations than the crash rate, we compare conclusions of crash status (yes/no) versus crash rate models. Examining sensitivity to the prediction of crash status (yes/no) versus crash rate, the stepwise selected model will be compared to logistic regression results for the model-building and

the validation sample (see Table 11).⁵ Generally, the linear regression model indicates that the numbers of effects that are related to crash rate are small, and that the better fitting models tend to have only a few predictors included. Specifically, Mallows' Cp statistic indicates a model is preferable when Cp is around or smaller than the number of effects (p), and the model is more parsimonious than competing models. The model fit to the restricted data, where carriers have greater than 20 vehicles, (stepwise model number 22), includes only 34 stable effects, and 72 effects altogether, but the model fit is more stable (i.e., relatively fewer unstable effects) and has the best (lowest) Cp, while also having similar explained variance and low AIC. However, it is interesting to note that the simple model, model 21, performs similarly according to some measures, such as Root MSE and R-squared, though this model does not contain violation rate information.

Table 11: Linear Regression Model Results for a Bayesian Crash-Rate Model, Using the Model Developed for the Crash Status (Yes/No) Outcome, Estimated with the Model-Building Sample

Model Group	Description: Crash Vio rate Data	Sub-Model description	AIC	Mallow's Cp	R-squared	Root MSE	Number of covariate effects in model	Number of stable covariate effects as defined in footnote 4
5	<ul style="list-style-type: none"> • Bayesian crash • Observed vio rate • Restricted 	21. Simple	-713	89	0.44	0.55	7	5
		22. Stepwise	-765	13	0.45	0.54	72	34
		23. Full	-610	169	0.46	0.55	169	31
6	<ul style="list-style-type: none"> • Bayesian crash • Bayesian vio rate • Restricted 	24. Stepwise	-980	81	0.47	0.53	37	25
		25. Full	-897	168	0.50	0.53	168	51
7	<ul style="list-style-type: none"> • Bayesian crash • Observed vio rate • Full 	26. Simple	-68,413	-2910	0.23	0.30	9	9
		27. Stepwise	-57,065	22	0.23	0.31	81	42
		28. Full	-56,917	171	0.23	0.31	171	46

⁵ When examining the relationship between violations and crash rates, as a sensitivity test related to using a linear model to predict crash rates, we also examined a negative binomial regression model for the number of crashes with an exposure measure of vehicles. We examined the consistency between the full models 23, 25, 28 and 30, when a linear versus a negative binomial regression is used by comparing the proportion of 161 violations within each of the four models that had the same significance and sign. Between 70 and 83 percent of the violations considered resulted in the same sign and significance status, regardless of whether a linear or negative binomial regression was used.

Model Group	Description: Crash Vio rate Data	Sub-Model description	AIC	Mallow's Cp	R-squared	Root MSE	Number of covariate effects in model	Number of stable covariate effects as defined in footnote 4
		30. Full	-60,338	170	0.29	0.30	170	94

Source: GAO analysis of FMCSA data.

Model Predictive Power

Comparing how well the models perform when applied to the validation sample that consists of new observations—which are not included in the model-building sample—informs the precision of SMS with respect to predicting crashes. We examine the number of violations and the violation types that are included across the model groups (logistic and linear) and sub-models (stepwise and full). We compare this to the number of models within which each violation was found to be a significant and a stable predictor of crash outcomes. Importantly, of the reduced set of approximately 160 violations considered, only 13 violations were significant in at least half of the 24 models that incorporate violations (i.e., stepwise and full models).

There were 10 different possible models for the logistic model-building sample, and these were also evaluated on the validation sample and on the model-building sample, but with a linear regression setting, resulting in 30 possible models. However, we regarded only 24 of these 30 models as informative since we exclude the 6 simple models that ignore the pre-violation information. Of the violations considered, only speeding (violation 3922S) and failure to use a seatbelt while operating CMV (39216) were significant and stable in all 24 models. A similar picture arises for some other violations, though many of the models did not result in a significant relationship between the violation in question and the crash outcome, as indicated in table 12. Only 41 violations were significant in 5 or more models out of 24. However, even for the top 13 violations with respect to frequency of significance and stability across the 24 models, predictive power is still affected by poor model diagnostics. This is echoed in the results from the predictive relationship when compared to the linear regression model for Bayesian crash rates (results in table 11), where the model that excluded all violations performed similarly to models that included some significant violations. Whether modeling crash status (yes/no) or a crash rate, the predictive power of SMS violations is weak.

**Appendix V: Analysis of Regulatory Violations
and Crash Risk**

Table 12: Numbers of Models for which Violations Were Significant and Stable Predictors, for Violations That Were Significant in 5 or More Models

	Input (violation)	Violation description	Violation group	BASIC	Number of models that included the input	Number of models where the input was significant (pvalue<=0.10)	Number of models where input was stable (See footnote 3)
1	39216	Failing to use seat belt while operating CMV	Seat Belt	Unsafe Driving	24	24	24
2	3922S	Speeding	Speeding Related	Unsafe Driving	24	24	24
3	393100A	Failure to prevent cargo shifting	General Securement	Vehicle Mainte	24	20	24
4	39617C	Operating a CMV without periodic inspection	Inspection Reports	Vehicle Mainte	24	20	20
5	3922C	Failure to obey traffic control device	Dangerous Driving	Unsafe Driving	24	17	20
6	39353B	Automatic brake adjuster CMV manufactured on or after 10/20/1994 - air bra	Brakes, All Others	Vehicle Mainte	24	16	19
7	3939H	Inoperative head lamps	Lighting	Vehicle Mainte	24	16	19
8	39141A	Driver not in possession of medical certificate	Medical Certificate	Driver Fitness	24	16	18
9	39260A	Unauthorized passenger on board CMV	Other Driver Violations	Unsafe Driving	21	16	12
10	39328	Improper or no wiring protection as required	Other Vehicle Defect	Vehicle Mainte	24	13	16
11	3958A	No driver's record of duty status	Incomplete/Wrong Log	HOS	21	13	14
12	39347	Inadequate/contaminated brake linings	Brakes, All Others	Vehicle Mainte	24	13	12
13	39343	No/improper breakaway or emergency braking	Brakes, All Others	Vehicle Mainte	21	12	12
14	3958F1	Driver's record of duty status not current	Incomplete/Wrong Log	HOS	24	11	15
15	39271A	Using or equipping a CMV with radar detector	Speeding Related	Unsafe Driving	21	11	7
16	39375F1	Weight carried exceeds tire load limit	Tire vs. Load	Vehicle Mainte	15	10	9
17	39395A	No/discharged/unsecured fire extinguisher	Emergency Equipment	Vehicle Mainte	18	9	14
18	3958E	False report of driver's record of duty status	False Log	HOS	18	9	11

**Appendix V: Analysis of Regulatory Violations
and Crash Risk**

	Input (violation)	Violation description	Violation group	BASIC	Number of models that included the input	Number of models where the input was significant (pvalue<=0.10)	Number of models where input was stable (See footnote 3)
19	39311TL	No retro reflective sheeting or reflex reflectors on mud flaps - Truck Tra	Reflective Sheeting	Vehicle Mainte	18	9	10
20	393207F	Air suspension pressure loss	Suspension	Vehicle Mainte	18	9	10
21	39145B	Expired medical examiner's certificate	Medical Certificate	Driver Fitness	21	8	13
22	3922FC	Following too close	Dangerous Driving	Unsafe Driving	18	8	11
23	39325F	Stop lamp violations	Lighting	Vehicle Mainte	18	8	11
24	393203	Cab/body parts requirements violations	Cab, Body, Frame	Vehicle Mainte	21	7	13
25	3965	Excessive oil leaks	Other Vehicle Defect	Vehicle Mainte	18	7	9
26	39324A	Noncompliance with headlamp requirements	Lighting	Vehicle Mainte	18	7	8
27	3922LC	Improper lane change	Dangerous Driving	Unsafe Driving	15	7	8
28	393130	No/improper heavy vehicle/machinery securement	General Securement	Vehicle Mainte	18	7	6
29	3953A2	Requiring or permitting driver to drive after 14 hours on duty	Hours	HOS	18	6	10
30	39395F	No / insufficient warning devices	Emergency Equipment	Vehicle Mainte	18	6	7
31	39343A	No/improper tractor protection valve	Brakes, All Others	Vehicle Mainte	18	6	6
32	39111	Unqualified driver	License-related: High	Driver Fitness	15	6	5
33	39222B	Failing/improper placement of warning devices	Cab, Body, Frame	Vehicle Mainte	15	6	5
34	39343D	No or defective automatic trailer brake	Brakes, All Others	Vehicle Mainte	18	5	9
35	3929A2	Failing to secure vehicle equipment	General Securement	Vehicle Mainte	18	5	7
36	39375A1	Tire — ply or belt material exposed	Tires	Vehicle Mainte	15	5	7

Appendix V: Analysis of Regulatory Violations and Crash Risk

Input (violation)	Violation description	Violation group	BASIC	Number of models that included the input	Number of models where the input was significant (pvalue<=0.10)	Number of models where input was stable (See footnote 3)
37 39313C3	No upper rear retroreflective sheeting or reflex reflective material as re	Reflective Sheeting	Vehicle Mainte	18	5	6
38 38323A2	Operating a CMV without a CDL	License-related: High	Driver Fitness	21	5	5
39 3953B	60/70 - hour rule violation	Hours	HOS	18	5	5
40 39378	Windshield wipers inoperative/defective	Windshield/Glass/Markings	Vehicle Mainte	15	5	5
41 3929A1	Failing to secure cargo	General Securement	Vehicle Mainte	18	5	2

Source: GAO analysis of FMCSA data.

Note: The number of models that included the input does not always equal 24, because inputs were dropped from step-wise models when they were insignificant or indicated estimation issues. All inputs were, however, tested in 24 models.

When comparing the predictive power of the models that result from the model-building sample, once applied to the validation sample, there is a consistent picture regarding the model fit (see table 13). In particular, the model fit is generally poor according to the H-L value; the stepwise model tends to perform better according to the AIC, but the ROC, adjusted R2, and percent discordant do not indicate the models have a strong ability to discriminate and predict future crashes. Classification tables that result from evaluating the model-building sample models, but estimated from the validation sample, generally resulted in similar results to those presented in table 10.

Table 13: Fit Statistics Based on the Validation Sample, for Crash Status (Yes/No)

Model group	Model group description: crash vio rate data	Sub-Model description	AIC	H-L Pvalue	Percentage concordant	Percentage discordant	R2	ROC	Number of covariate effects in model	Number of stable covariate effects as defined in footnote 3
1	• Crash status (yes/no)	11. Simple	6,864	<0.001	79.2	19.6	0.23	0.80	7	5
		12. Stepwise	6,503	0.001	81.7	18.3	0.26	0.82	72	22
	• Observed vio rate	13. Full	6,572	0.045	82.5	17.5	0.27	0.82	169	45
	• Restricted									

Appendix V: Analysis of Regulatory Violations and Crash Risk

Model group	Model group description: crash vio rate data	Sub-Model description	AIC	H-L Pvalue	Percentage concordant	Percentage discordant	R2	ROC	Number of covariate effects in model	Number of stable covariate effects as defined in footnote 3
2	• Crash status (yes/no)	14. Stepwise	6,375	0.295	82.0	18.0	0.26	0.82	37	21
		15. Full	6,485	0.135	83.0	17.0	0.28	0.83	168	29
	• Bayesian vio rate • Restricted									
3	• Crash status (yes/no)	16. Simple	69,205	<0.001	69.9	19.9	0.16	0.75	9	9
		17. Stepwise	62,155	<0.001	76.5	23.4	0.17	0.77	81	33
	• Observed vio rate • Full	18. Full	62,212	<0.001	76.7	23.3	0.18	0.77	171	52
4	• Crash status (yes/no)	19. Stepwise	61,446	<0.001	76.7	23.3	0.18	0.77	49	31
		20. Full	61,525	<0.001	76.8	23.2	0.18	0.77	170	45
	• Bayesian vio rate • Full									

Source: GAO analysis of FMCSA data.

Conclusions

The predictive power observed in these modeling and sensitivity analyses indicates that SMS may be less precise than what is reported and that the available information on violations is limited for the purpose of scoring carriers or predicting their crash risk. Regardless of which type of model we fit, we see that the predictive power of our models is low, and the use of the SMS violations in predicting future crashes is not very precise. The number of stable and significant effects across the various model-fitting scenarios that include violations is small. For the about 800 violations in SMS, only around 160 met the basic criteria of non-zero variance and non-zero counts for at least 1 percent of the sample. Of these, only two violations (speeding and failure to wear a seatbelt while operating a CMV) consistently appeared as a stable predictor of crashes, regardless of data and model. While some other violations appeared in models, only 13 were significant and stable in at least half of the models, most were significant in no more than half the models examined, and most often in fewer than 5 of the models. The results did not vary substantially according to whether observed versus Bayesian violation rates, crash versus Bayesian crash rates, or restricted data (carriers with more than 20 vehicles) versus full data were used to estimate crashes. Therefore

the modeling attempts did not overcome the issues that result from small exposures. The results were generally confirmed when evaluated on a validation sample, indicating the future prediction is stable, yet not strong. Ultimately, much of the variance in crash predictions remains unexplained, regardless of the model and model-building data, so that the SMS might be less precise when the objective is to predict crashes.

Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

This appendix provides additional information and illustrations of the distribution of motor carrier population included in our analysis such as carrier size, number of crashes, inspections, and high risk status (see table 14). It also provides results of our analysis on the number and percentage of carriers above or below intervention thresholds, as well as the frequency and rate of crashes for each of those groups of carriers within each BASIC using FMCSA's methodology and the illustrative alternative methodology (i.e., using a stronger data sufficiency standard) demonstrated earlier in the report. In addition, this appendix provides summary statistics of the various motor carrier populations used in FMCSA and GAO analysis. These statistics include, among other things, the numbers of carriers with an SMS score (i.e., "measure") and the number of carriers above an intervention threshold in at least one BASIC. Finally, this appendix provides the complete graphical results of our analysis of FMCSA's violation rates, safety event groups, and distribution of SMS scores for carriers above FMCSA's intervention threshold using FMCSA's methodology.

**Appendix VI: Descriptive Statistics on Motor
Carrier Population and Results of GAO's
Analysis**

Table 14: Distribution of Crashes, Power Units, Inspections, and High Risk Status by Carrier Size (GAO Analysis Population)

Carrier Size (Power Units)	Carriers		Power units		Crashes		Fatal crashes		Average number of inspections	Percentage classified as high risk
	No.	(%)	No.	(%)	No.	(%)	No.	(%)	No.	(%)
1	125,902	40.0	125,902	3.5	6,534	5.4	202	5.6	4.0	1.4
2	51,465	16.4	102,930	2.8	4,001	3.3	135	3.7	5.1	1.9
3	29,278	9.3	87,834	2.4	3,118	2.6	93	2.6	6.4	2.2
4	19,846	6.3	79,384	2.2	2,768	2.3	89	2.5	7.9	2.6
5	14,258	4.5	71,290	2.0	2,611	2.2	83	2.3	9.9	3.2
6	10,125	3.2	60,750	1.7	2,201	1.8	76	2.1	10.8	3.2
7	7,382	2.3	51,674	1.4	1,873	1.6	77	2.1	12.4	3.2
8	6,092	1.9	48,736	1.3	1,809	1.5	65	1.8	14.1	3.2
9	4,734	1.5	42,606	1.2	1,651	1.4	66	1.8	15.6	3.0
10	4,624	1.5	46,240	1.3	1,713	1.4	48	1.3	17.4	4.2
11	3,311	1.1	36,421	1.0	1,299	1.1	55	1.5	18.0	3.4
12	3,051	1.0	36,612	1.0	1,322	1.1	47	1.3	20.6	3.5
13	2,496	0.8	32,448	0.9	1,191	1.0	32	0.9	21.9	3.0
14	2,176	0.7	30,464	0.8	1,113	0.9	36	1.0	22.9	4.0
15	2,081	0.7	31,215	0.9	1,134	0.9	39	1.1	24.8	4.2
16	1,772	0.6	28,352	0.8	1,053	0.9	35	1.0	27.2	4.0
17	1,505	0.5	25,585	0.7	886	0.7	31	0.9	28.0	2.9
18	1,437	0.5	25,866	0.7	946	0.8	23	0.6	29.3	3.1
19	1,148	0.4	21,812	0.6	844	0.7	27	0.7	30.6	3.9
20	1,398	0.4	27,960	0.8	1,106	0.9	37	1.0	36.3	5.0
21	918	0.3	19,278	0.5	763	0.6	22	0.6	36.8	4.4
22	898	0.3	19,756	0.5	723	0.6	22	0.6	34.1	3.9
23	742	0.2	17,066	0.5	598	0.5	14	0.4	34.7	3.8
24	719	0.2	17,256	0.5	996	0.8	35	1.0	37.8	3.9
25	798	0.3	19,950	0.6	744	0.6	14	0.4	44.5	4.3
26-50	8,653	2.7	305,778	8.4	11,369	9.4	371	10.3	54.5	4.9
51-100	4,253	1.4	296,923	8.2	11,130	9.2	345	9.6	105.3	4.8
101-500	3,070	1.0	611,360	16.9	20,886	17.4	595	16.5	247.9	6.1
501-1,000	354	0.1	242,553	6.7	7,861	6.5	199	5.5	771.6	7.3
1,001-10,000	256	0.1	587,439	16.2	18,189	15.1	507	14.1	2,181.5	7.4
10,000+	15	0.0	467,889	12.9	7,902	6.6	182	5.1	9,972.5	0.0
Total	314,757	100.0	3,619,329	100.0	120,334	100.0	3,602	100.0	15.9	2.3

Source: GAO analysis of FMCSA data.

Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

Table 15 contains the results of our analysis using FMCSA's SMS 3.0 methodology. This analysis calculated the number and percentage of carriers above and below intervention thresholds for each BASIC using carrier data from December 2007 through December 2009, and determined which carriers subsequently crashed during the 18-month evaluation period, December 2009 through June 2011. The analysis also presents aggregate crash rates for comparison purposes.

Table 15: Comparison of Crash Involvement for Carriers above and below Intervention Threshold Using FMCSA's Methodology (Compare to Illustrative Alternative Analysis in Following Table)

	No. of Carriers Involved in Crash	No. of Carriers Not Involved in Crash	Total (%)	Crashes per 100 vehicles ^a
Unsafe Driving:				
Above Threshold	4,575	7,597	12,172	7.13
[col%] ^b (row %) ^c	[27.2] (37.6)	[47.5] (62.4)	[37.1] (100)	
No. of Crashes ^d	17,268	n.a.		
Below Threshold	12,247	8,402	20,649	3.55
[col%] ^b (row %) ^c	[72.8] (59.3)	[52.5] (40.7)	[62.9] (100)	
No. of Crashes ^d	67,552	n.a.		
Total	16,822	15,999	32,821	3.96
[col%] ^b (row %) ^c	[100] (51.3)	[100] (48.7)	[100] (100)	
Hours-of-Service Compliance:				
Above Threshold	7,702	18,693	26,395	6.63
[col%] ^b (row %) ^c	[42.9] (29.2)	[57.7] (70.8)	[52.4] (100)	
No. of Crashes ^d	26,248	n.a.		
Below Threshold	10,267	13,712	23,979	3.62
[col%] ^b (row %) ^c	[57.1] (42.8)	[42.3] (57.2)	[47.6] (100)	
No. of Crashes ^d	55,838	n.a.		
Total	17,969	32,405	50,374	4.24
[col%] ^b (row %) ^c	[100] (35.7)	[100] (64.3)	[100] (100)	
Driver Fitness:				
Above Threshold	1,892	1,880	3,772	2.87
[col%] ^b (row %) ^c	[37.8] (50.2)	[57.9] (49.8)	[45.7] (100)	
No. of Crashes ^d	11,677	n.a.		
Below Threshold	3,114	1,365	4,479	3.94
[col%] ^b (row %) ^c	[62.2] (69.5)	[42.1] (30.5)	[54.3] (100)	
No. of Crashes ^d	44,957	n.a.		
Total	5,006	3,245	8,251	3.66
[col%] ^b (row %) ^c	[100] (60.7)	[100] (39.3)	[100] (100)	

**Appendix VI: Descriptive Statistics on Motor
Carrier Population and Results of GAO's
Analysis**

	No. of Carriers Involved in Crash	No. of Carriers Not Involved in Crash	Total (%)	Crashes per 100 vehicles^a
Controlled Substance and Alcohol:				
Above Threshold	59	512	571	3.24
[col%] ^b (row %) ^c	[5.2] (10.3)	[36.9] (89.7)	[22.7] (100)	
No. of Crashes ^d	133	n.a.		
Below Threshold	1,069	874	1,943	5.21
[col%] ^b (row %) ^c	[94.8] (55.0)	[63.1] (45.0)	[77.3] (100)	
No. of Crashes ^d	21,317	n.a.		
Total	1,128	1,386	2,514	5.19
[col%]^b (row %)^c	[100] (44.9)	[100] (55.1)	[100] (100)	
Vehicle Maintenance:				
Above Threshold	5,283	12,154	17,437	5.56
[col%] ^b (row %) ^c	[21.5] (30.3)	[29.0] (69.7)	[26.2] (100)	
No. of Crashes ^d	15,216	n.a.		
Below Threshold	19,283	29,766	49,049	3.64
[col%] ^b (row %) ^c	[78.5] (39.3)	[71.0] (60.7)	[73.8] (100)	
No. of Crashes ^d	81,908	n.a.		
Total	24,566	41,920	66,486	3.84
[col%]^b (row %)^c	[100] (36.9)	[100] (63.1)	[100] (100)	
Hazardous Materials:				
Above Threshold	412	263	675	5.47
[col%] ^b (row %) ^c	[31.8] (61.0)	[49.4] (39.0)	[37.0] (100)	
No. of Crashes ^d	14,095	n.a.		
Below Threshold	882	269	1,151	3.46
[col%] ^b (row %) ^c	[68.2] (76.6)	[50.6] (23.4)	[63.0] (100)	
No. of Crashes ^d	12,815	n.a.		
Total	1,294	532	1,826	4.28
[col%]^b (row %)^c	[100] (70.9)	[100] (29.1)	[100] (100)	
Crash Indicator:				
Above Threshold	3,256	2,788	6,044	7.19
[col%] ^b (row %) ^c	[31.1] (53.9)	[56.5] (46.1)	[39.2] (100)	
No. of Crashes ^d	22,219	n.a.		
Below Threshold	7,219	2,143	9,362	3.21
[col%] ^b (row %) ^c	[68.9] (77.1)	[43.5] (22.9)	[60.8] (100)	
No. of Crashes ^d	53,657	n.a.		
Total	10,475	4,931	15,406	3.83
[col%]^b (row %)^c	[100] (68.0)	[100] (32.0)	[100] (100)	

Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

	No. of Carriers Involved in Crash	No. of Carriers Not Involved in Crash	Total (%)	Crashes per 100 vehicles^a
High Risk:				
Above Threshold	2,808	4,393	7,201	8.38
[col%] ^b (row %) ^c	[9.8] (39.0)	[7.3] (61.0)	[8.1] (100)	
No. of Crashes ^d	12,624	n.a.		
Below Threshold	25,876	56,135	82,011	3.55
[col%] ^b (row %) ^c	[90.2] (31.6)	[92.7] (68.4)	[91.9] (100)	
No. of Crashes ^d	90,726	n.a.		
Total	28,684	60,528	89,212	3.82
[col%] ^b (row %) ^c	[100] (32.2)	[100] (67.8)	[100] (100)	

Source: GAO analysis of FMCSA data.

^aThese figures represent the aggregate crash rates for carriers in the respective rows for each BASIC (i.e., above threshold, below threshold, and total). The aggregate crash rate is calculated by dividing the number of crashes for all carriers in the row (e.g., above threshold) by the number of vehicles (i.e., power units) for those carriers, and is expressed per 100 vehicles.

^bColumn percentages are in brackets. For example, note that there were 32,821 carriers (that had an SMS score) for the Unsafe Driving BASIC. Some 16,822 of these carriers experienced a crash in the evaluation period (total at the bottom of the "No. of Carriers that Crashed" column). Just above the 16,822 total one can observe that 4,575 (27.2 percent) of those carriers that crashed were above the intervention threshold and the remaining 12,247 (72.8 percent) carriers were below the intervention threshold.

^cRow percentages are in parentheses. For example, of those carriers (with an SMS score) in the Unsafe Driving BASIC, 12,172 had a score above the intervention threshold. 4,575 (37.6 percent) of those carriers experienced a crash in the evaluation period and the remaining 7,597 (62.4 percent) did not crash.

^dThese figures are the total number of crashes for the carriers represented in the corresponding cell.

Table 16 contains the results of our analysis using an illustrative alternative incorporating a stronger data sufficiency standard, among other things, as described elsewhere in this report (e.g. carriers with 20 or more inspections or 20 or more vehicles, depending upon the BASIC). As in the previous table, this analysis calculated the number of carriers above and below intervention thresholds for each BASIC using carrier data from December 2007 through December 2009, and determined which carriers subsequently crashed during the subsequent 18-month period, December 2009 through June 2011. The analysis also presents aggregate crash rates for comparison purposes.

Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

Table 16: Comparison of Crash Involvement for Carriers above and below Intervention Threshold using Illustrative Alternative (Compare to FMCSA's Methodology in Previous Table)

	No. of Carriers Involved in Crash	No. of Carriers Not Involved in Crash	Total (%)	Crashes per 100 vehicles ^a
Unsafe Driving:				
Above Threshold	6,404	1,606	8,010	6.13
[col%] ^b (row %) ^c	[51.7] (80.0)	[19.5] (20.0)	[38.9] (100)	
No. of Crashes ^d	50,407	n.a.		
Below Threshold	5,979	6,612	12,591	1.76
[col%] ^b (row %) ^c	[48.3] (47.5)	[80.5] (52.5)	[61.1] (100)	
No. of Crashes ^d	29,247	n.a.		
Total	12,383	8,218	20,601	3.20
[col%] ^b (row %) ^c	[100] (60.1)	[100] (39.9)	[100] (100)	
Hours-of-Service Compliance:				
Above Threshold	6,299	6,389	12,688	6.72
[col%] ^b (row %) ^c	[33.7] (49.6)	[36.5] (50.4)	[35.1] (100)	
No. of Crashes ^d	23,631	n.a.		
Below Threshold	12,413	11,093	23,506	3.39
[col%] ^b (row %) ^c	[66.3] (52.8)	[63.5] (47.2)	[64.9] (100)	
No. of Crashes ^d	65,792	n.a.		
Total	18,712	17,482	36,194	3.90
[col%] ^b (row %) ^c	[100] (51.7)	[100] (48.3)	[100] (100)	
Driver Fitness:				
Above Threshold	3,149	4,125	7,274	2.63
[col%] ^b (row %) ^c	[16.8] (43.3)	[23.6] (56.7)	[20.1] (100)	
No. of Crashes ^d	7,967	n.a.		
Below Threshold	15,563	13,357	28,920	4.09
[col%] ^b (row %) ^c	[83.2] (53.8)	[76.4] (46.2)	[79.9] (100)	
No. of Crashes ^d	81,456	n.a.		
Total	18,712	17,482	36,194	3.90
[col%] ^b (row %) ^c	[100] (51.7)	[100] (48.3)	[100] (100)	
Controlled Substance and Alcohol:				
Above Threshold	1,522	893	2,415	4.71
[col%] ^b (row %) ^c	[8.1] (63.0)	[5.1] (37.0)	[6.7] (100)	
No. of Crashes ^d	8,678	n.a.		
Below Threshold	17,190	16,589	33,779	3.83
[col%] ^b (row %) ^c	[91.9] (50.9)	[94.9] (49.1)	[93.3] (100)	
No. of Crashes ^d	80,745	n.a.		
Total	18,712	17,482	36,194	3.90
[col%] ^b (row %) ^c	[100] (51.7)	[100] (48.3)	[100] (100)	

Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

	No. of Carriers Involved in Crash	No. of Carriers Not Involved in Crash	Total (%)	Crashes per 100 vehicles^a
Vehicle Maintenance:				
Above Threshold	2,532	2,458	4,990	6.35
[col%] ^b (row %) ^c	[17.1] (50.7)	[24.6] (49.3)	[20.2] (100)	
No. of Crashes ^d	7,172	n.a.		
Below Threshold	12,245	7,527	19,772	3.71
[col%] ^b (row %) ^c	[82.9] (61.9)	[75.4] (38.1)	[79.8] (100)	
No. of Crashes ^d	76,045	n.a.		
Total	14,777	9,985	24,762	3.84
[col%]^b (row %)^c	[100] (59.7)	[100] (40.3)	[100] (100)	
Hazardous Materials:				
Above Threshold	286	130	416	5.07
[col%] ^b (row %) ^c	[19.1] (68.8)	[21.6] (31.3)	[19.8] (100)	
No. of Crashes ^d	7,019	n.a.		
Below Threshold	1,209	471	1,680	3.57
[col%] ^b (row %) ^c	[80.9] (72.0)	[78.4] (28.0)	[80.2] (100)	
No. of Crashes ^d	21,308	n.a.		
Total	1,495	601	2,096	3.85
[col%]^b (row %)^c	[100] (71.3)	[100] (28.7)	[100] (100)	
Crash Indicator:				
Above Threshold	4,438	956	5,394	6.83
[col%] ^b (row %) ^c	[42.0] (82.3)	[24.3] (17.7)	[37.2] (100)	
No. of Crashes ^d	40,587	n.a.		
Below Threshold	6,130	2,984	9,114	2.19
[col%] ^b (row %) ^c	[58.0] (67.3)	[75.7] (32.7)	[62.8] (100)	
No. of Crashes ^d	36,227	n.a.		
Total	10,568	3,940	14,508	3.42
[col%]^b (row %)^c	[100] (72.8)	[100] (27.2)	[100] (100)	
High Risk:				
Above Threshold	4,032	1,975	6,007	8.25
[col%] ^b (row %) ^c	[19.0] (67.1)	[8.7] (32.9)	[13.6] (100)	
No. of Crashes ^d	22,961	n.a.		
Below Threshold	17,186	20,815	38,001	2.90
[col%] ^b (row %) ^c	[81.0] (45.2)	[91.3] (54.8)	[86.4] (100)	
No. of Crashes ^d	71,182	n.a.		
Total	21,218	22,790	44,008	3.44
[col%]^b (row %)^c	[100] (48.2)	[100] (51.8)	[100] (100)	

Source: GAO analysis of FMCSA data.

Note: See appendix I: Objectives, Scope, and Methodology for information on the carrier population used in this analysis. The illustrative alternative presented here only includes carriers with at least 20 relevant inspections or vehicles (depending upon the BASIC).

Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

^aThese figures represent the aggregate crash rates for carriers in the respective rows for each BASIC (i.e. above threshold, below threshold, and total). The aggregate crash rate is calculated by dividing the number of crashes by the number of vehicles (i.e. power units) and is expressed per 100 vehicles.

^bColumn percentages are in brackets. See note to previous table for interpretation of the numbers and percentages in this table.

^cRow percentages are in parentheses. See note to previous table for interpretation of the numbers and percentages in this table.

^dThese figures are the total number of crashes for the carriers represented in the corresponding cell.

Table 17 contains selected SMS outcomes based on results reported by FMCSA's and from GAO's analysis.

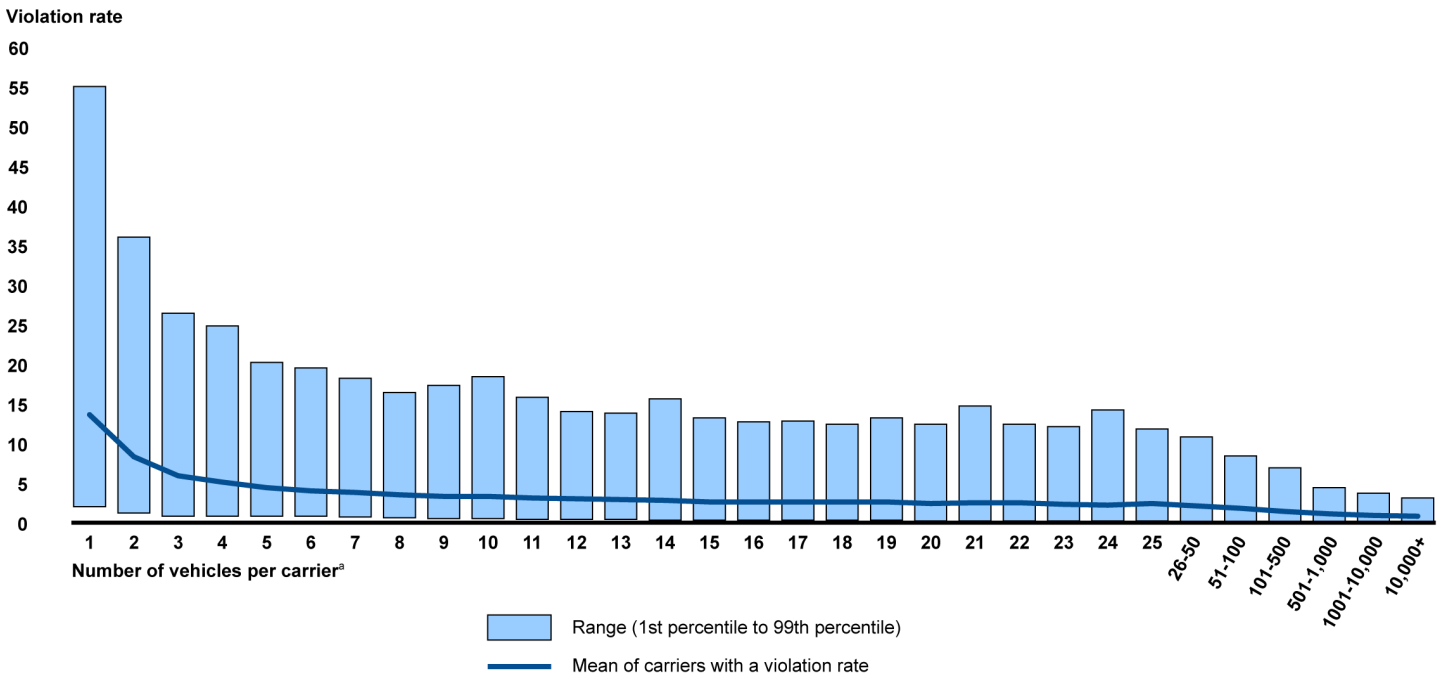
Table 17: SMS Outcomes as Reported by FMCSA Compared to Outcomes from GAO Analysis

	FMCSA	FMCSA effectiveness test	GAO's replication of FMCSA	Illustrative alternative
Population (Number of carriers)	525,000	276,855	314,757	314,757
Carriers with a measure score	200,000	161,555	283,041	283,041
Carriers with a percentile in at least one BASIC (% of total)	92,000 (17.5%)	76,215 (27.5%)	89,212 (28.3%)	44,008 (14.0%)
Carriers above the intervention threshold in 1 or more BASICs	50,000 (9.5%)	41,789 (15.1%)	49,927 (15.9%)	24,696 (7.8%)
Number of crashes for above threshold carriers (crash rate)		58,064 (5.05)	62,825 (5.08)	69,228 (5.15)
High risk carriers		6,731	7,201	6,007
Number of crashes for high risk carriers (evaluation period)(crash rate)		15,391 (8.15)	12,624 (8.38)	22,961 (8.25)
Number of vehicles (Power units)		188,922	150,614	278,280

Source: GAO analysis of FMCSA data.

The following figures are graphical results of our analysis of the average and range of violation rates for carriers, percentage of carriers above FMCSA's intervention thresholds for various safety event group categories, and distribution of SMS scores for carriers above FMCSA's intervention thresholds using FMCSA's methodology as discussed in the body of this report above. Figures 10 through 16 contain the average and range of violation rates for all carriers (where a violation rate could be calculated) by carrier size, for all the BASICS. Figures 17 through 25 contain the percentage of carriers above intervention thresholds within safety event groups for each BASIC. Finally, figures 26 through 32 show the distribution of carriers above intervention thresholds for each BASIC by carrier size.

Figure 10: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Unsafe Driving BASIC



Source: GAO analysis of FMCSA data.

³This number is an adjusted average number of vehicles that FMCSA uses to calculate an SMS score for carriers in the Unsafe Driving BASIC.

Figure 11: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Hours-of-Service Compliance BASIC

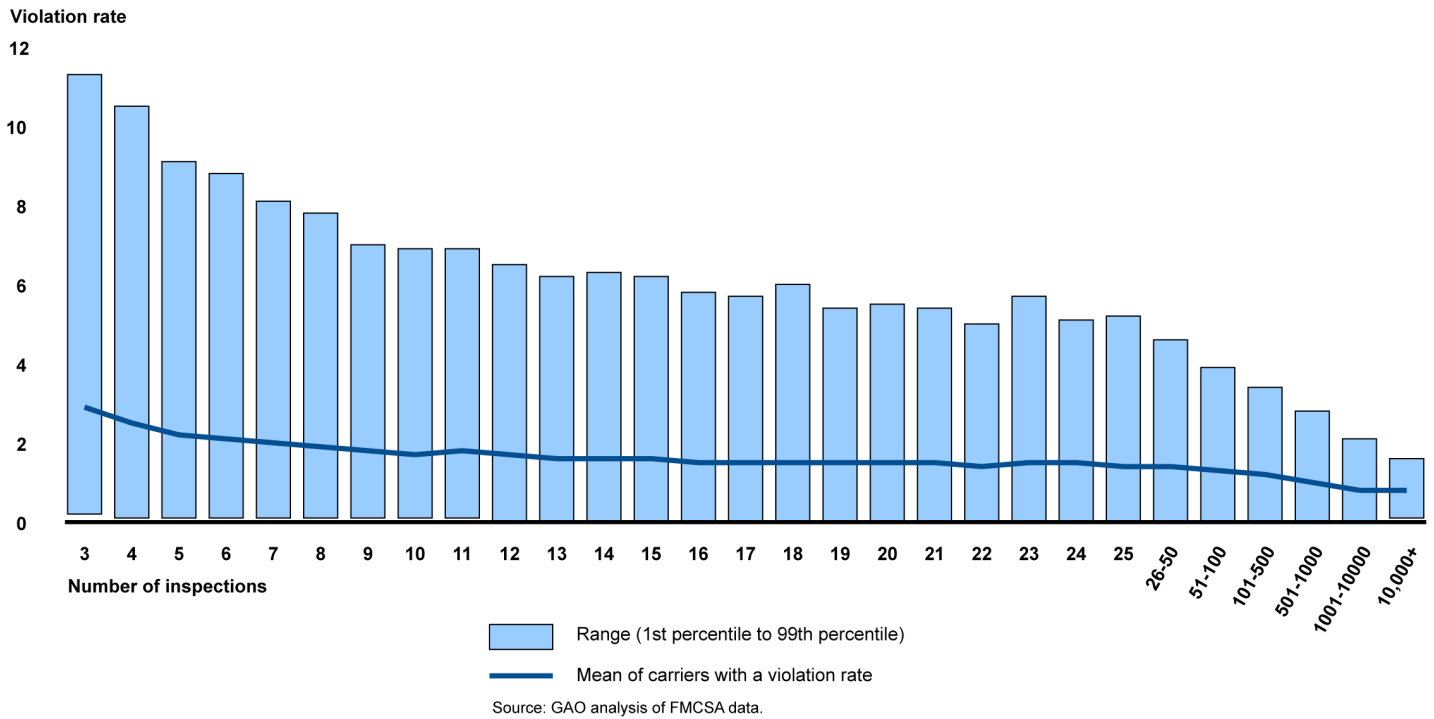
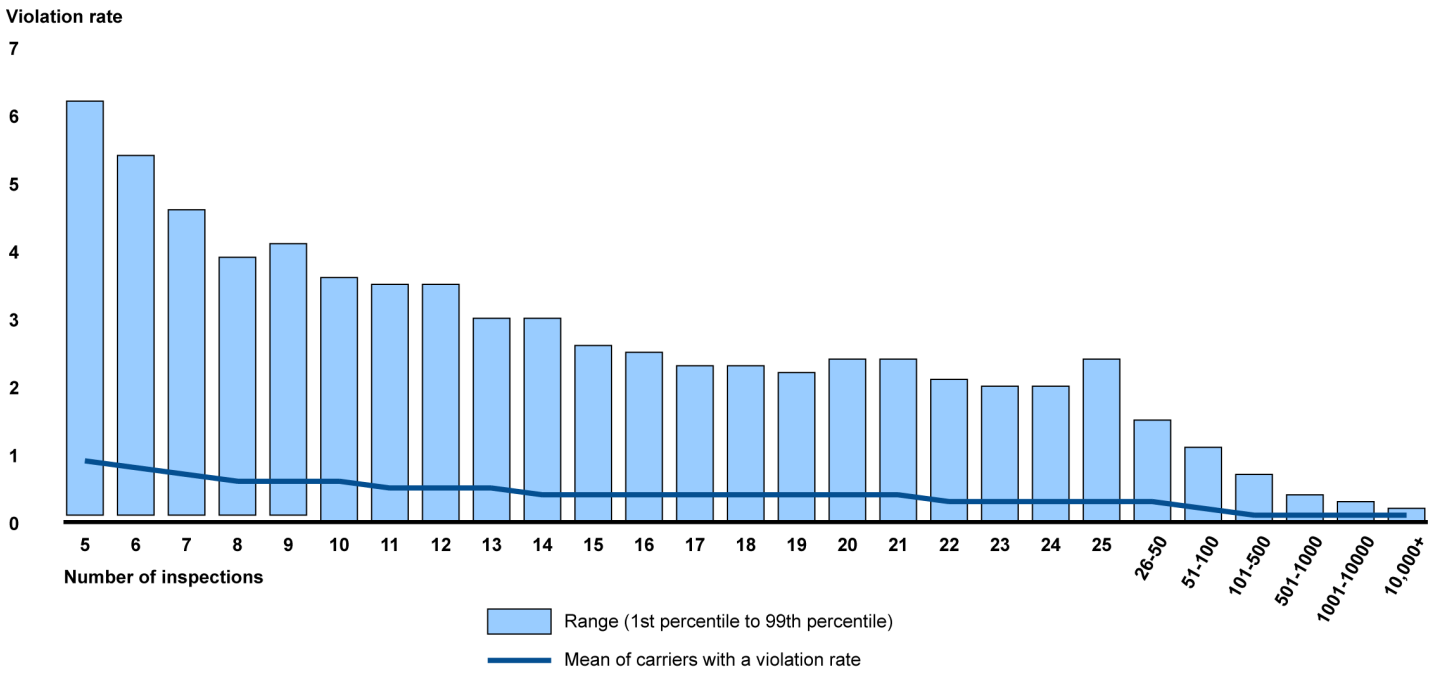
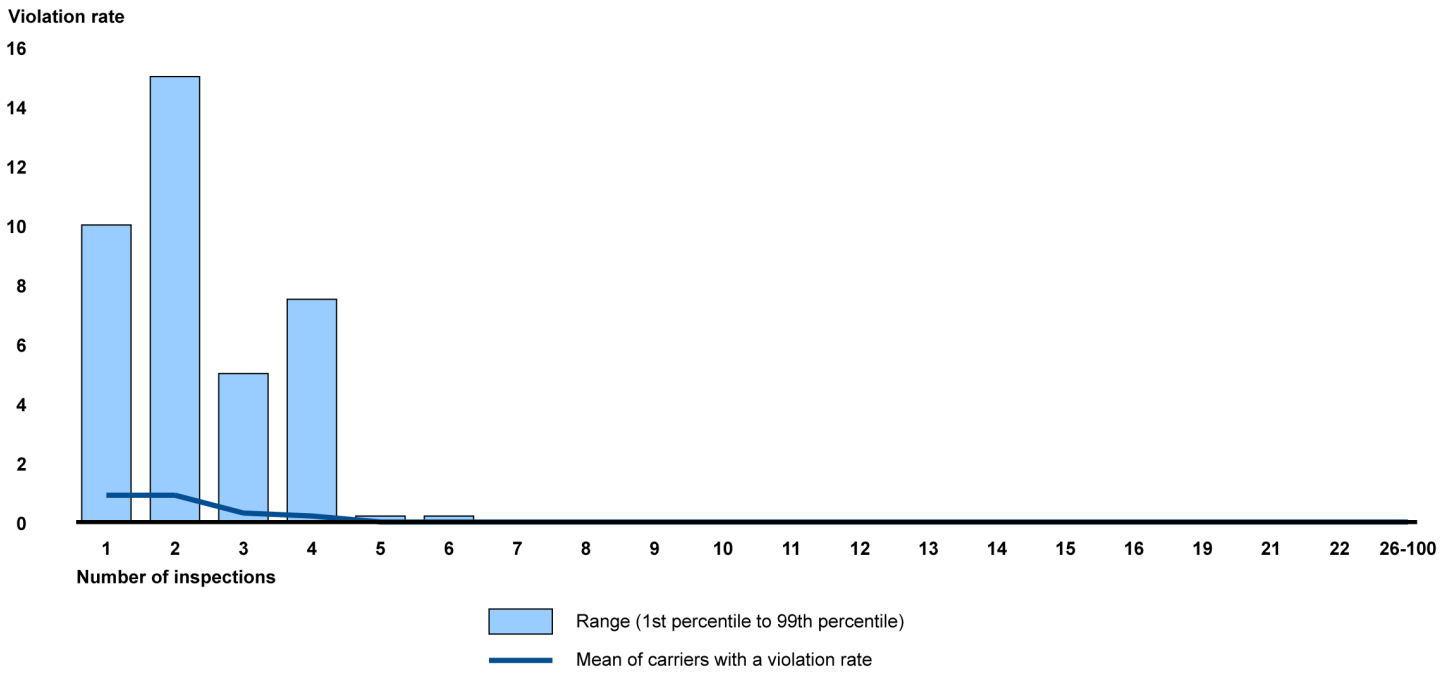


Figure 12: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Driver Fitness BASIC



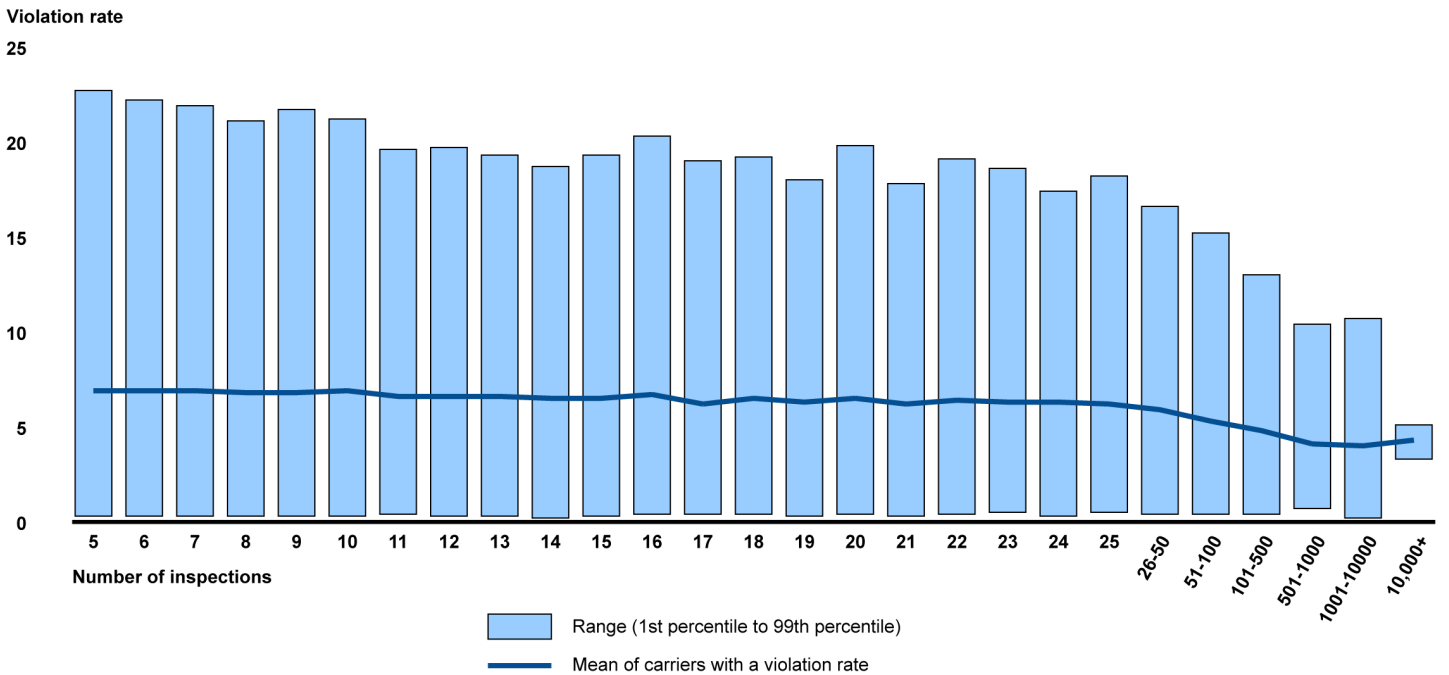
Source: GAO analysis of FMCSA data.

Figure 13: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Controlled Substances and Alcohol BASIC



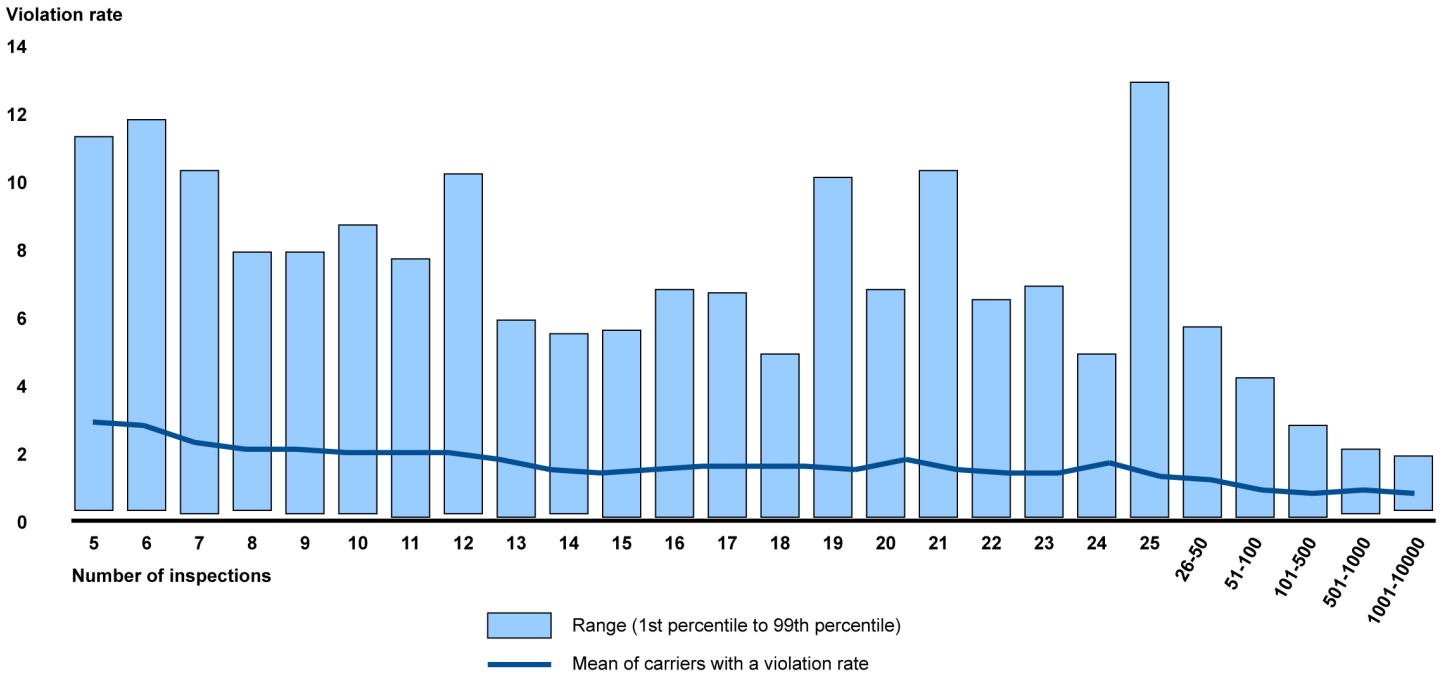
Source: GAO analysis of FMCSA data.

Figure 14: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Vehicle Maintenance BASIC



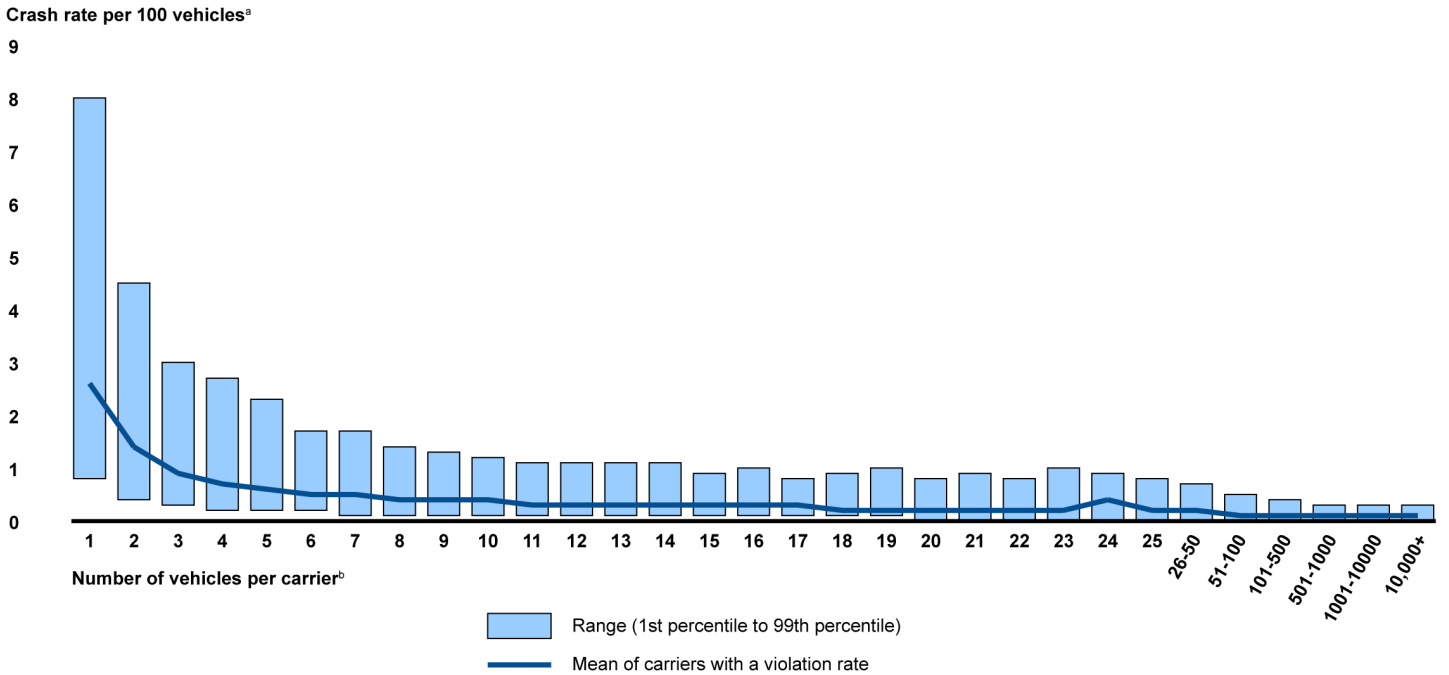
Source: GAO analysis of FMCSA data.

Figure 15: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Hazardous Materials BASIC



Source: GAO analysis of FMCSA data.

Figure 16: Average and Range of Violation Rates (between the 1st and 99th Percentiles) for Carriers in the Crash Indicator BASIC



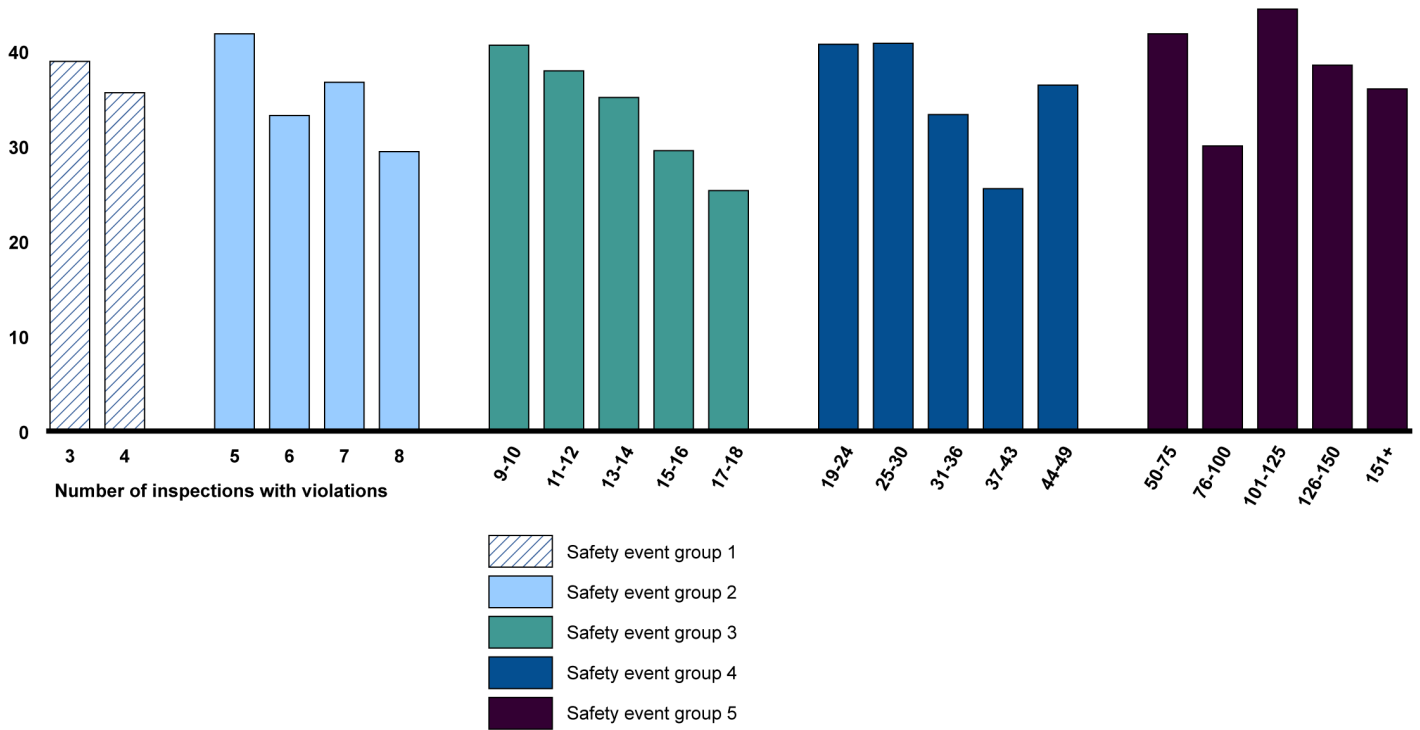
Source: GAO analysis of FMCSA data.

^aThis number is a weighted crash rate based on a weighted average number of vehicles that FMCSA uses to calculate a score.

^bThis number is an adjusted average number of vehicles that FMCSA uses to calculate an SMS score for carriers on the Crash Indicator.

Figure 17: Percentage of FMCSA Scored Carriers in the Unsafe Driving (Straight Segment) BASIC above the Intervention Threshold by Number of Inspections

Percent of carriers above the intervention threshold
50



Source: GAO analysis of FMCSA data.

Figure 18: Percentage of FMCSA Scored Carriers in the Unsafe Driving (Combo Segment) BASIC above the Intervention Threshold by Number of Inspections

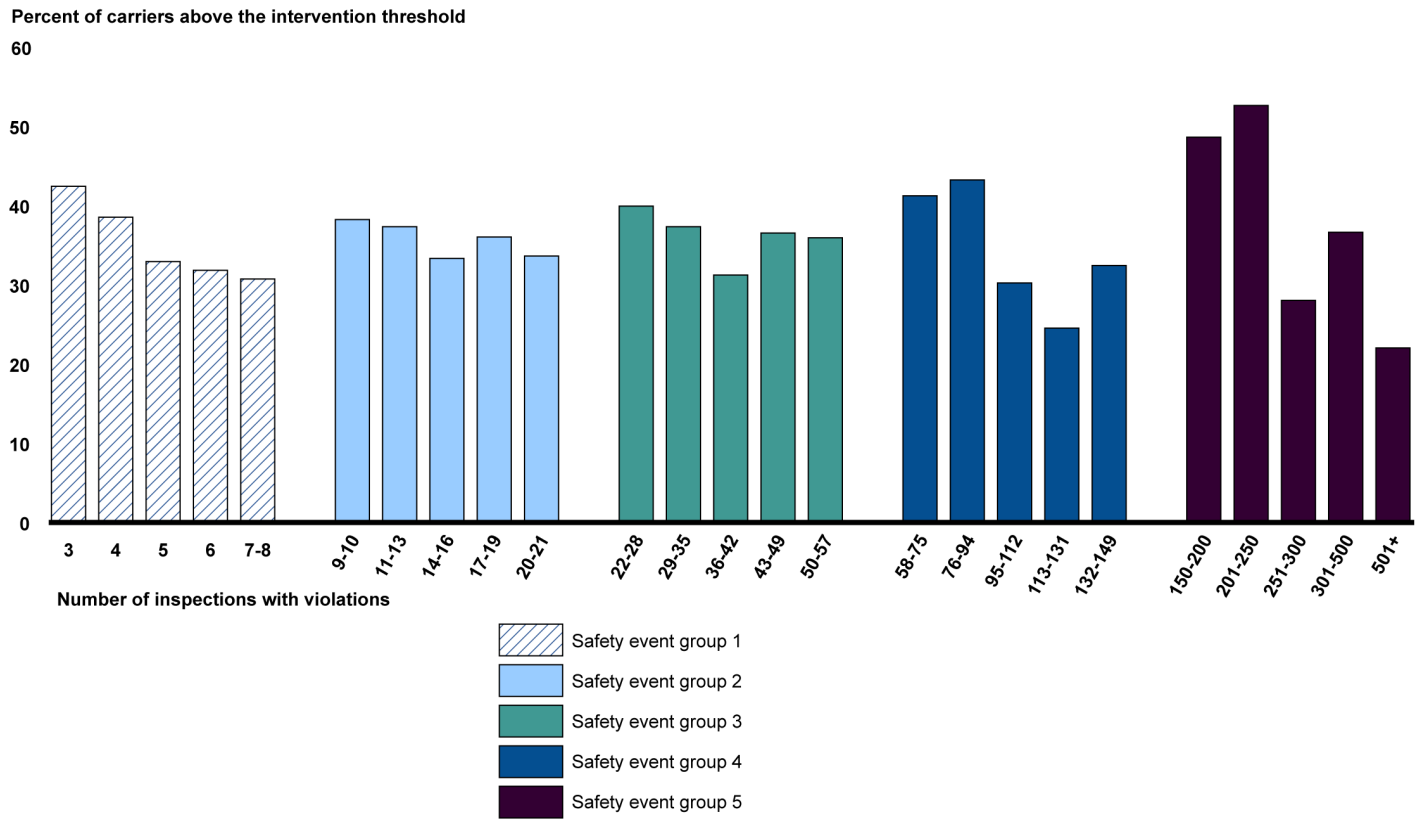
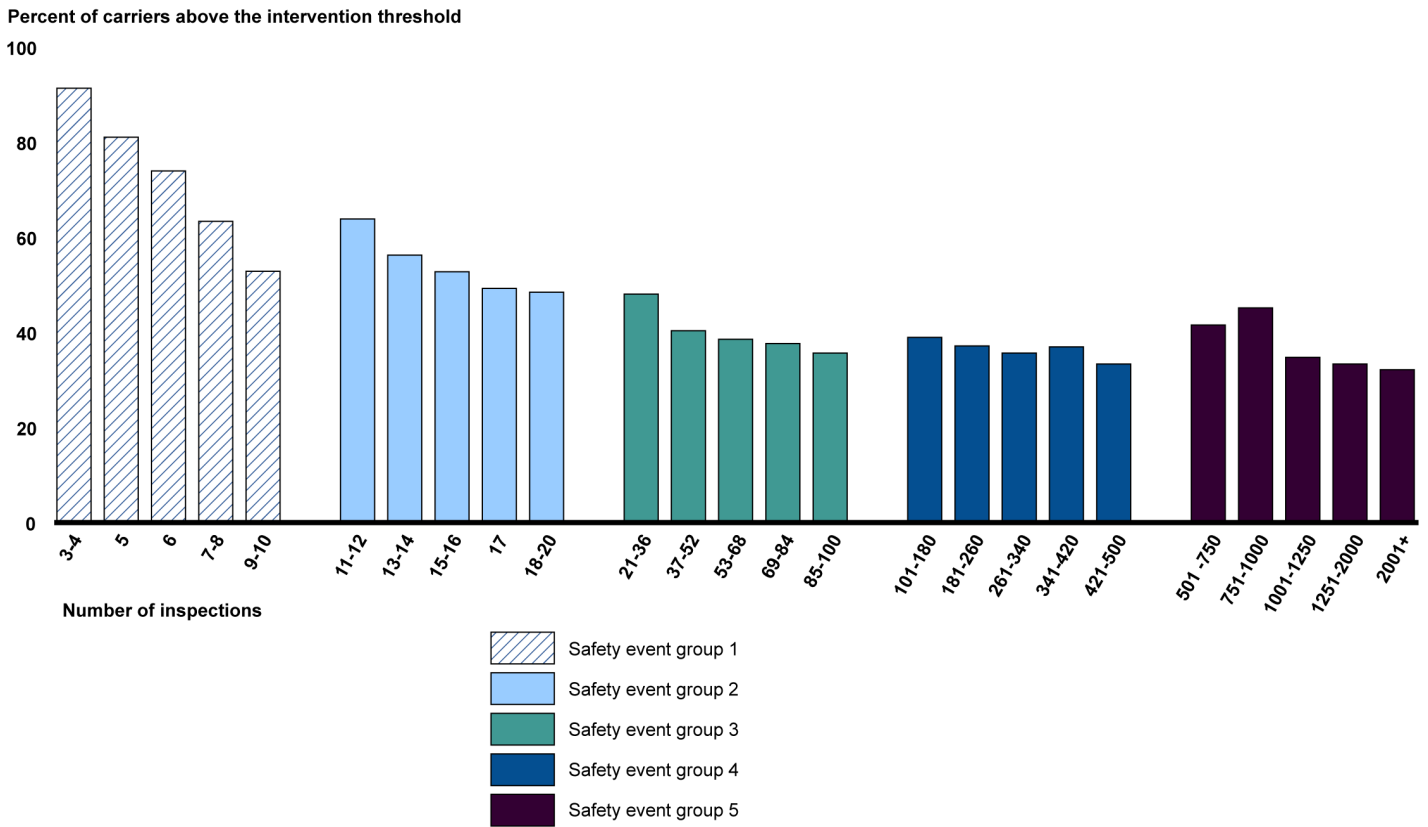
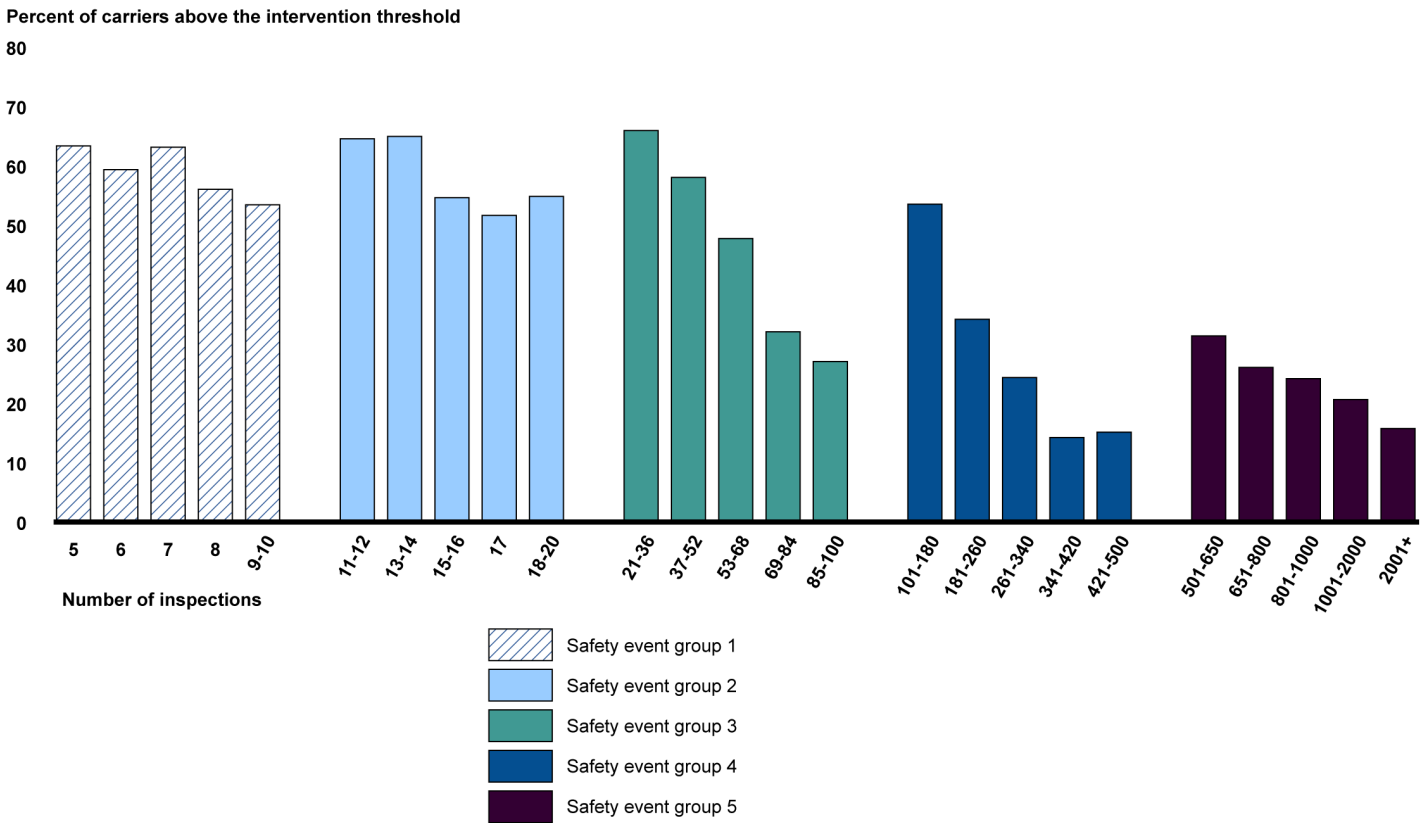


Figure 19: Percentage of FMCSA Scored Carriers in the Hours-of-Service Compliance BASIC above the Intervention Threshold by Number of Inspections



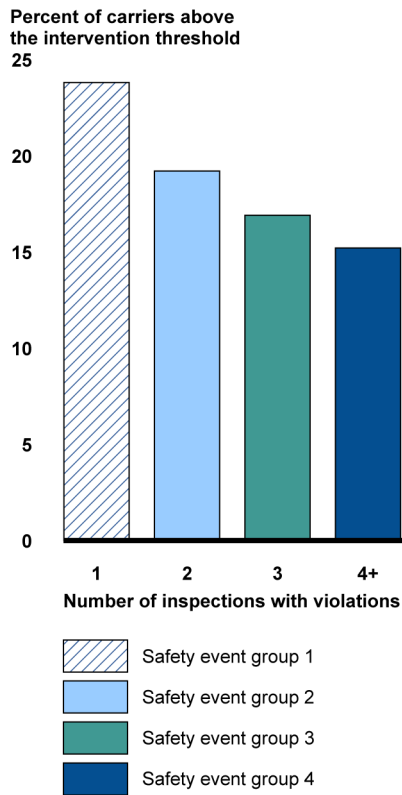
Source: GAO analysis of FMCSA data.

Figure 20: Percentage of FMCSA Scored Carriers in the Driver Fitness BASIC above the Intervention Threshold by Number of Inspections



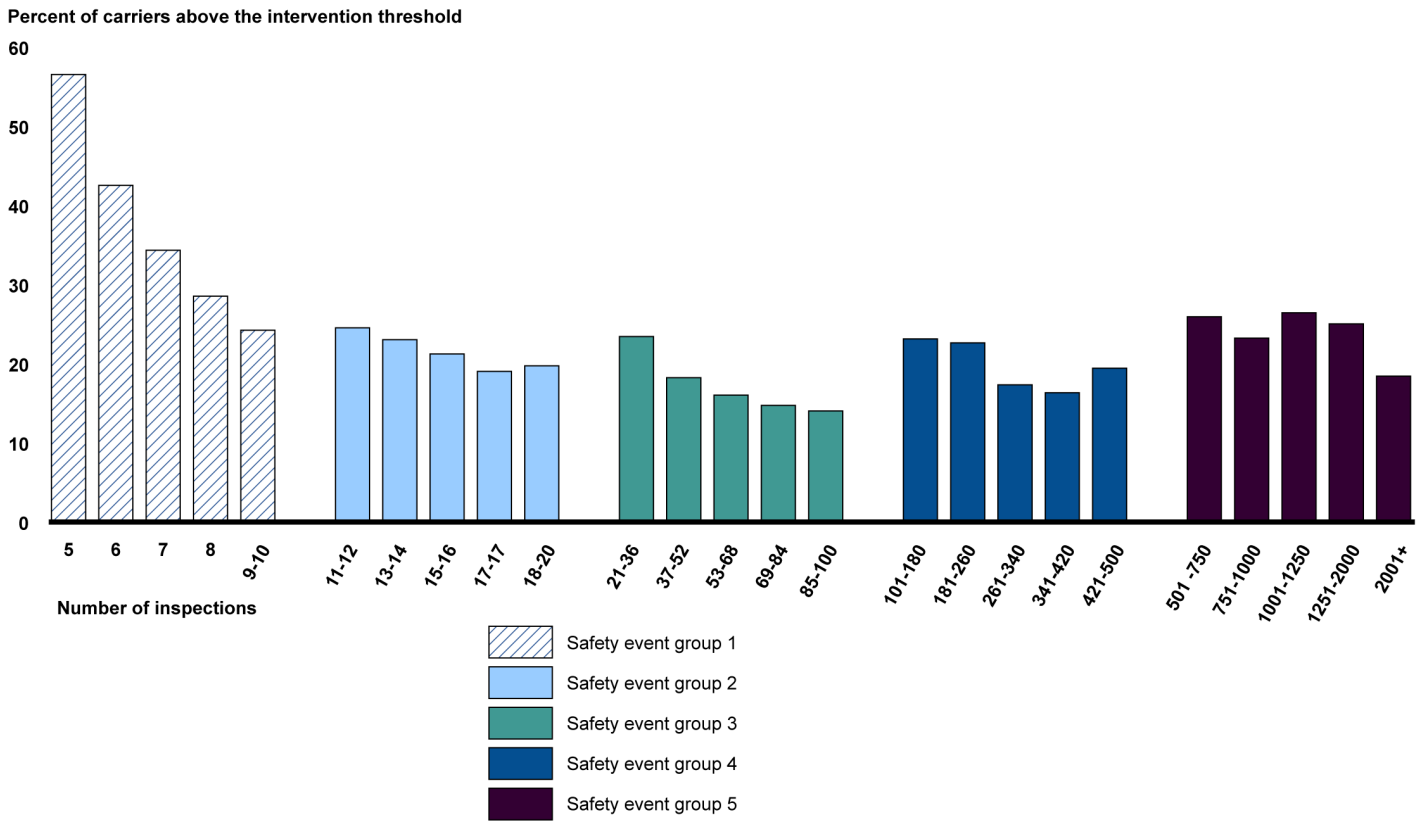
Source: GAO analysis of FMCSA data.

Figure 21: Percentage of FMCSA-Scored Carriers in the Controlled Substances and Alcohol BASIC above the Intervention Threshold by Number of Inspections



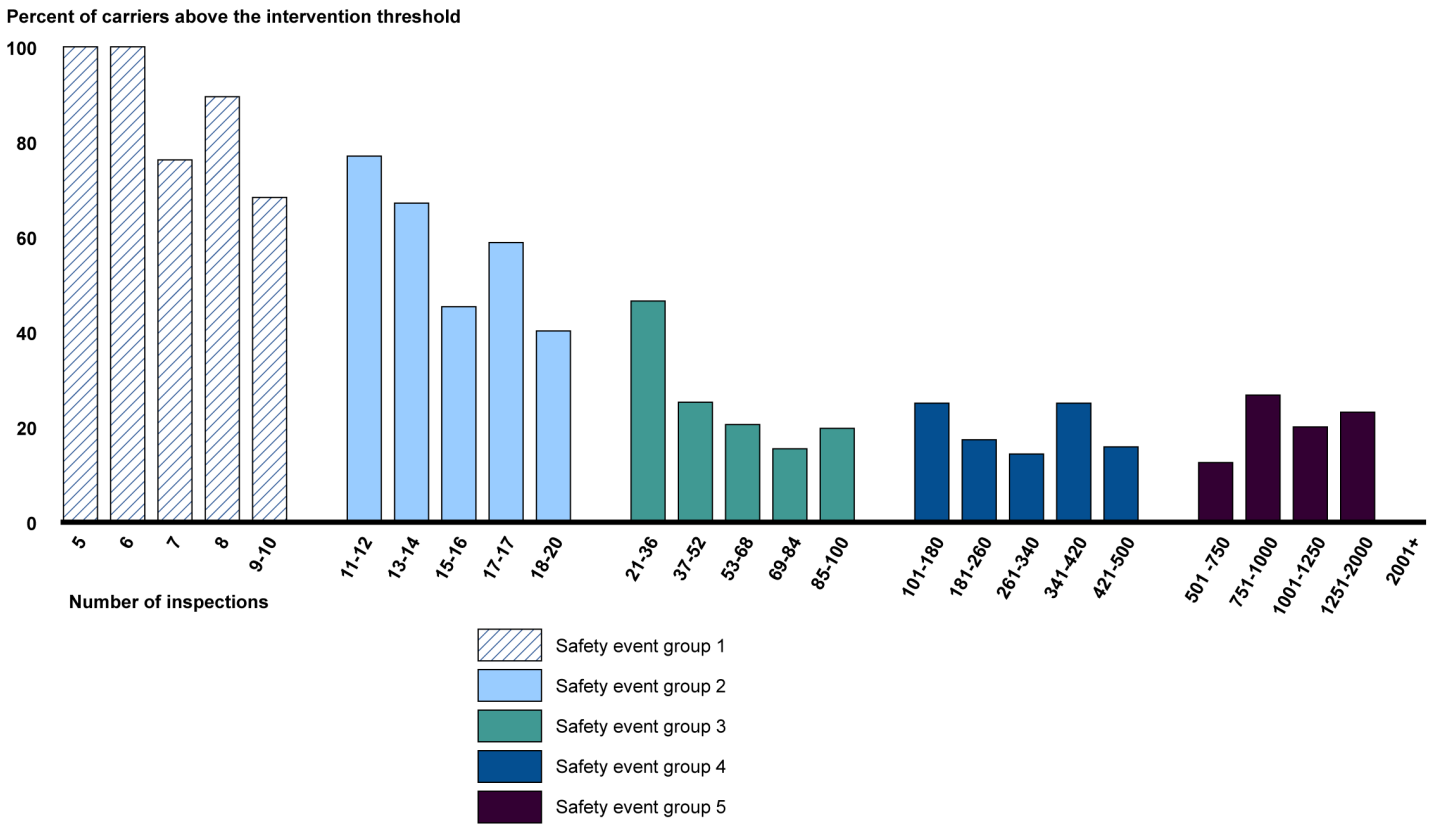
Source: GAO analysis of FMCSA data.

Figure 22: Percentage of FMCSA-Scored Carriers in the Vehicle Maintenance BASIC above the Intervention Threshold by Number of Inspections



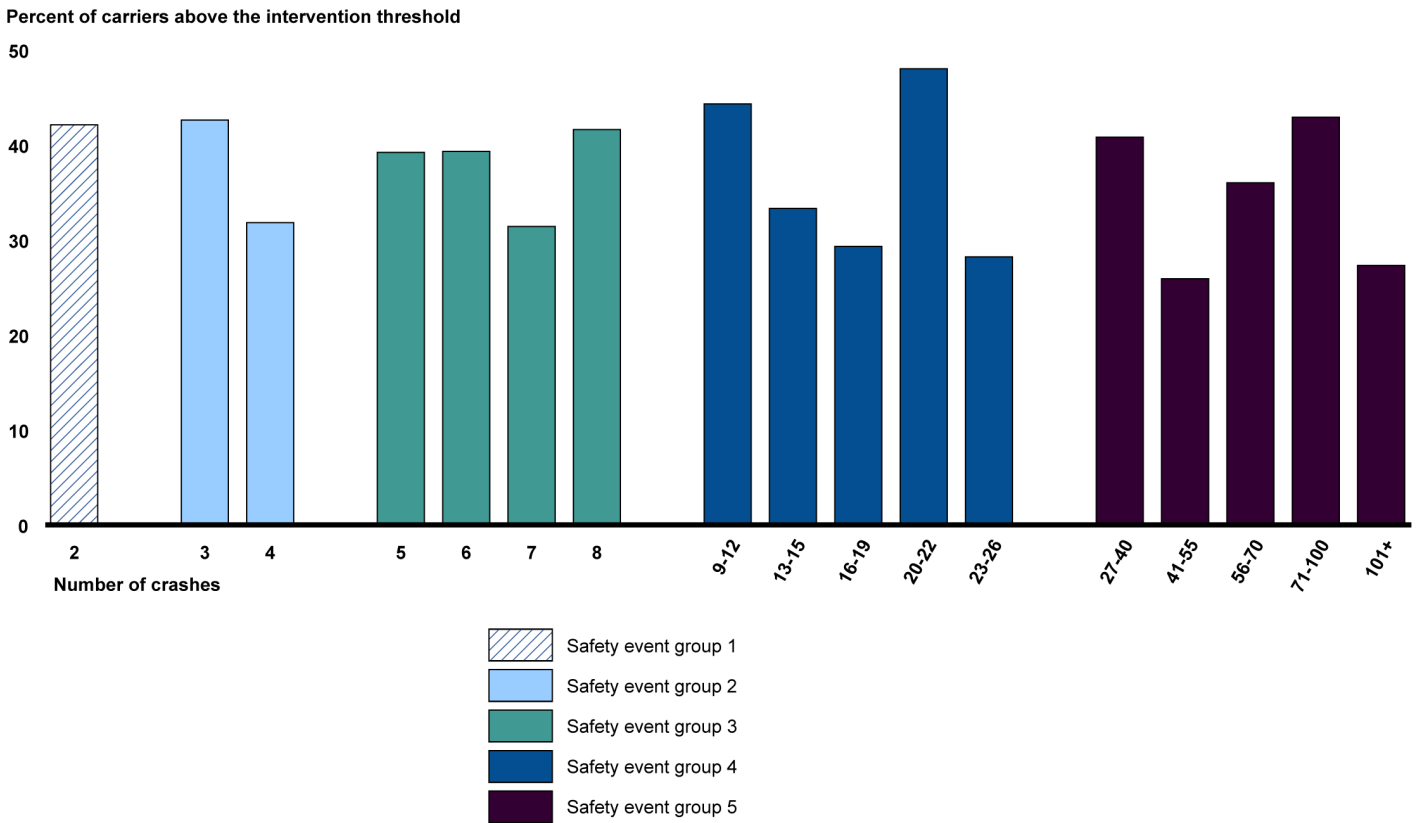
Source: GAO analysis of FMCSA data.

Figure 23: Percentage of FMCSA-Scored Carriers in the Hazardous Materials BASIC above the Intervention Threshold by Number of Inspections



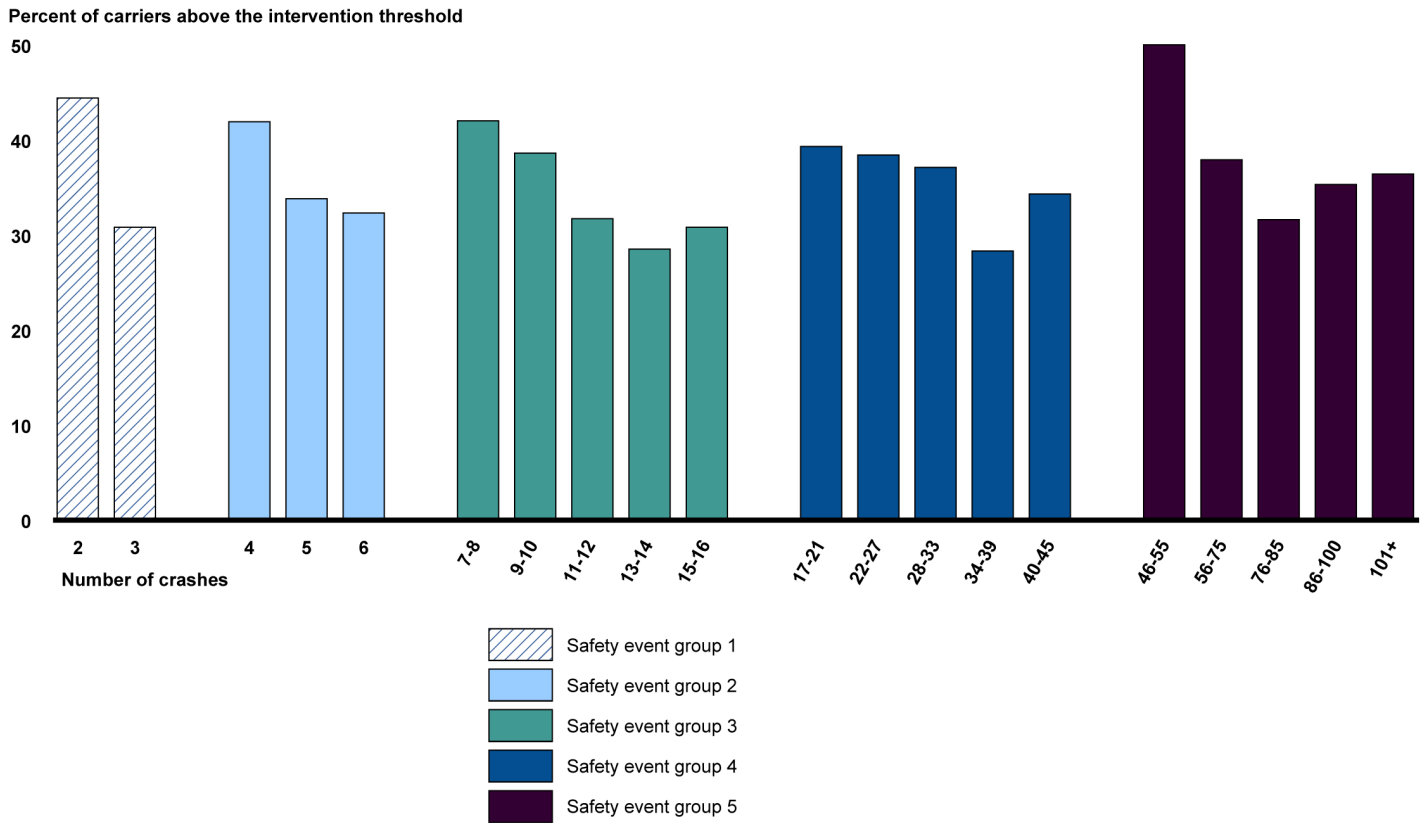
Source: GAO analysis of FMCSA data.

Figure 24: Percentage of FMCSA-Scored Carriers on the Crash Indicator (Straight Segment) above the Intervention Threshold by Number of Inspections



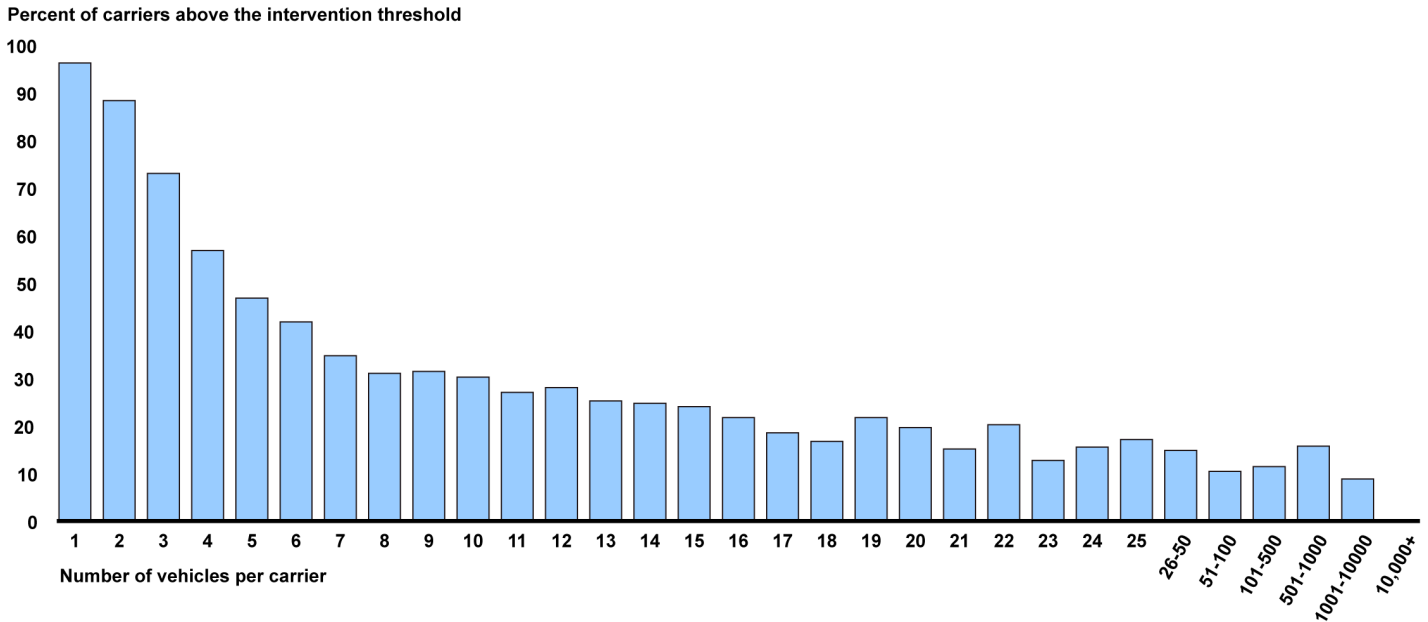
Source: GAO analysis of FMCSA data.

Figure 25: Percentage of FMCSA-Scored Carriers on the Crash Indicator (Combo Segment) above the Intervention Threshold by Number of Inspections



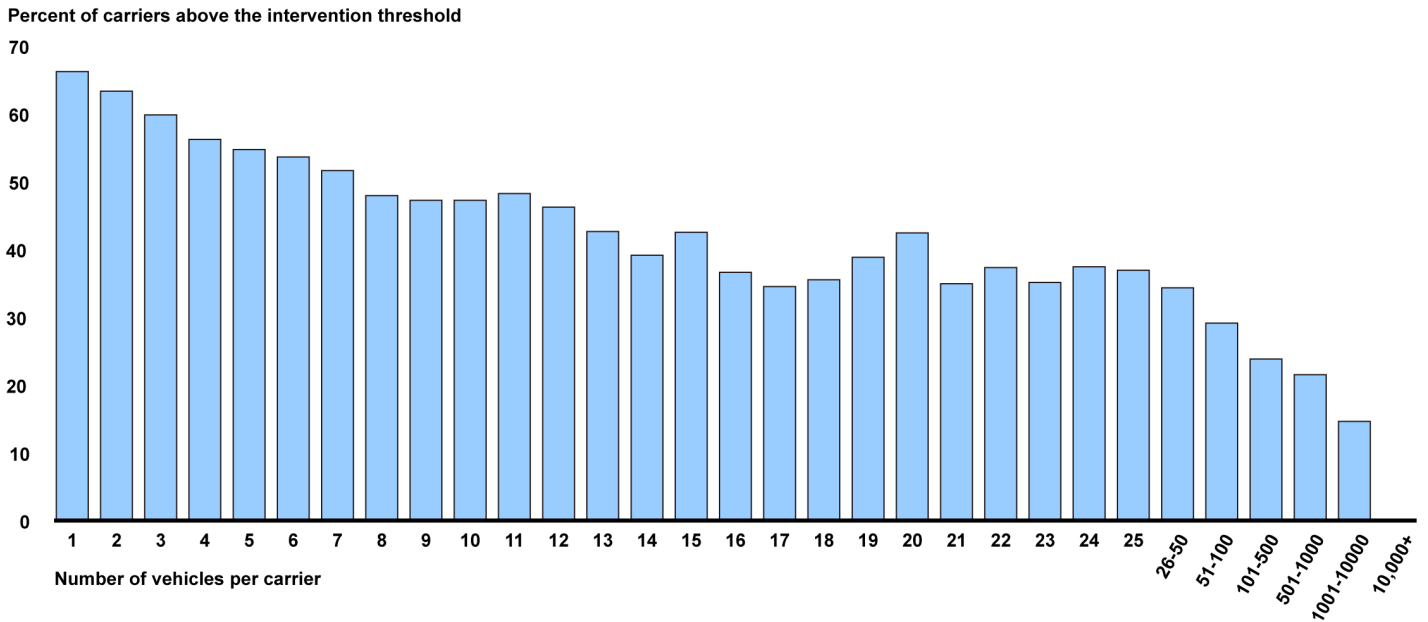
Source: GAO analysis of FMCSA data.

Figure 26: Distribution of FMCSA-Scored Carriers above the Unsafe Driving BASIC Threshold by Carrier Size



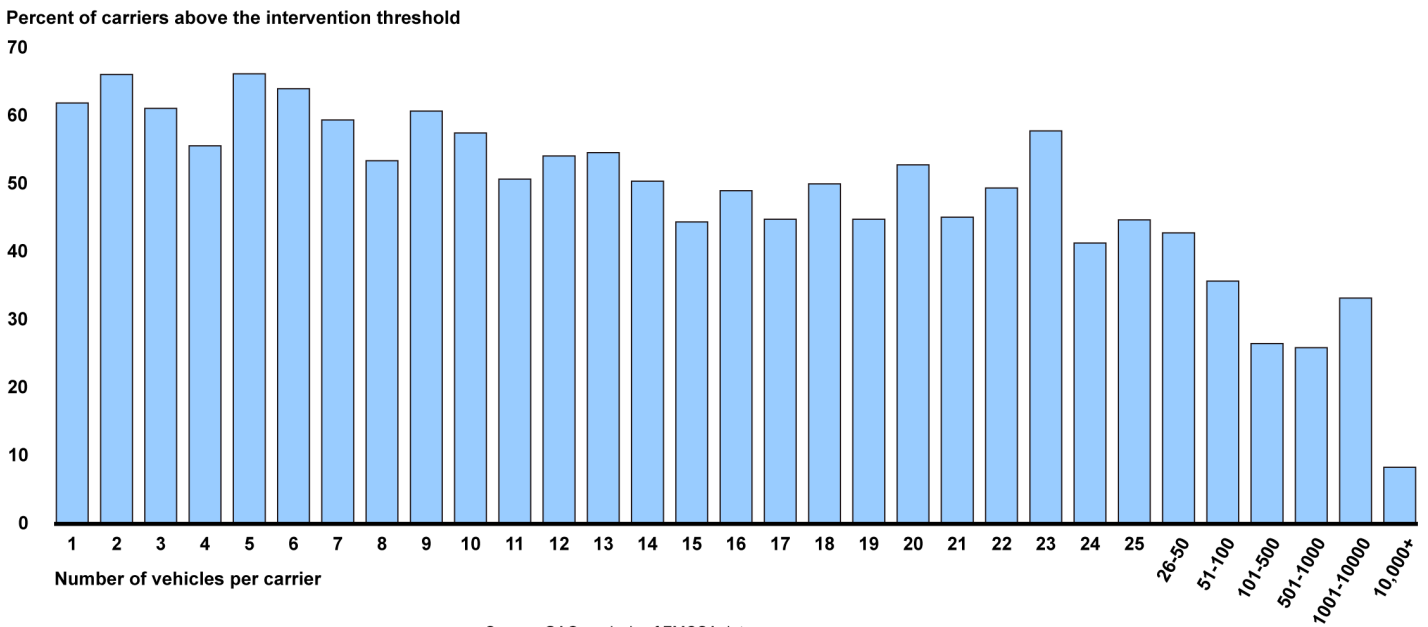
Source: GAO analysis of FMCSA data.

Figure 27: Distribution of FMCSA-Scored Carriers above the Hours-of-Service Compliance BASIC Threshold by Carrier Size



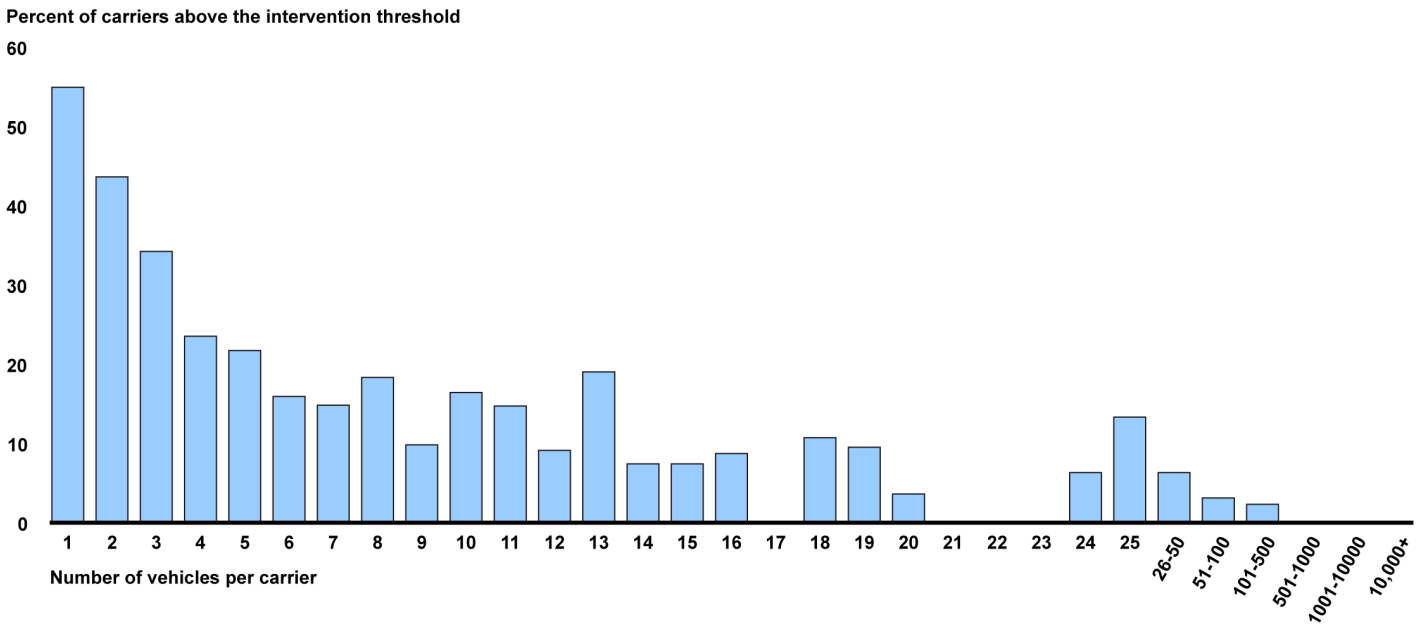
Source: GAO analysis of FMCSA data.

Figure 28: Distribution of FMCSA-Scored Carriers above the Driver Fitness BASIC Threshold by Carrier Size



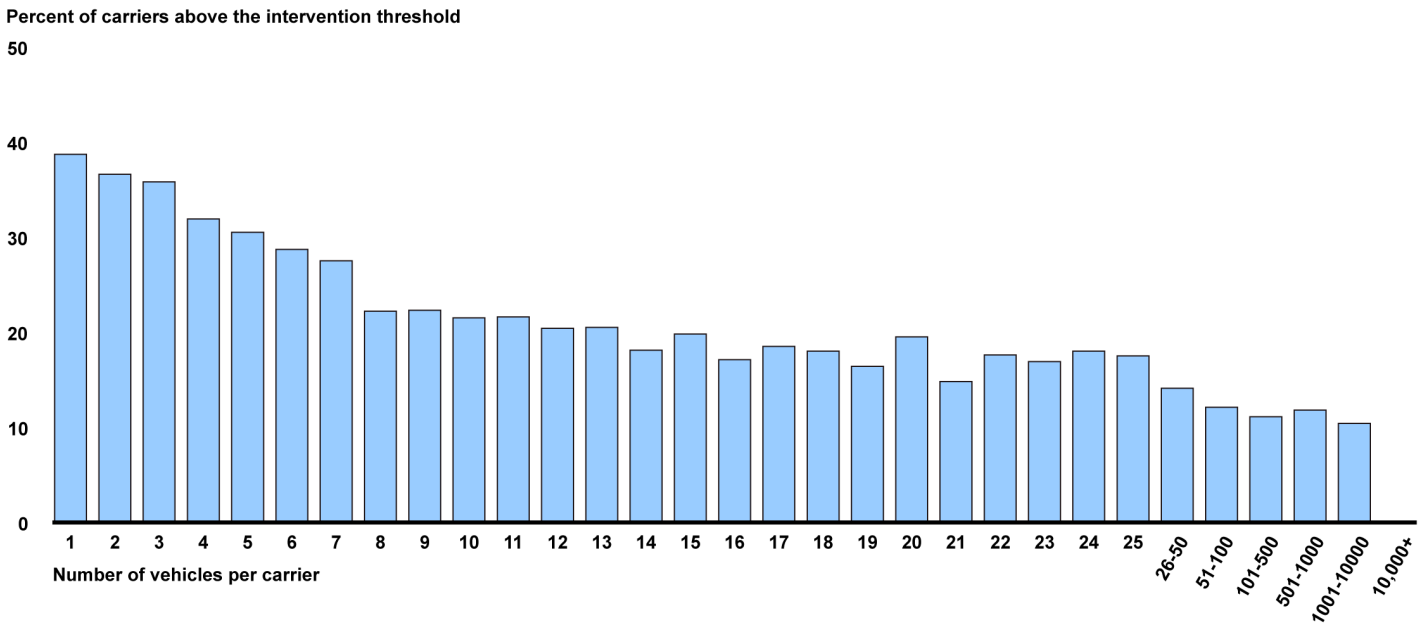
Appendix VI: Descriptive Statistics on Motor Carrier Population and Results of GAO's Analysis

Figure 29: Distribution of FMCSA-Scored Carriers above the Controlled Substance and Alcohol BASIC Threshold by Carrier Size



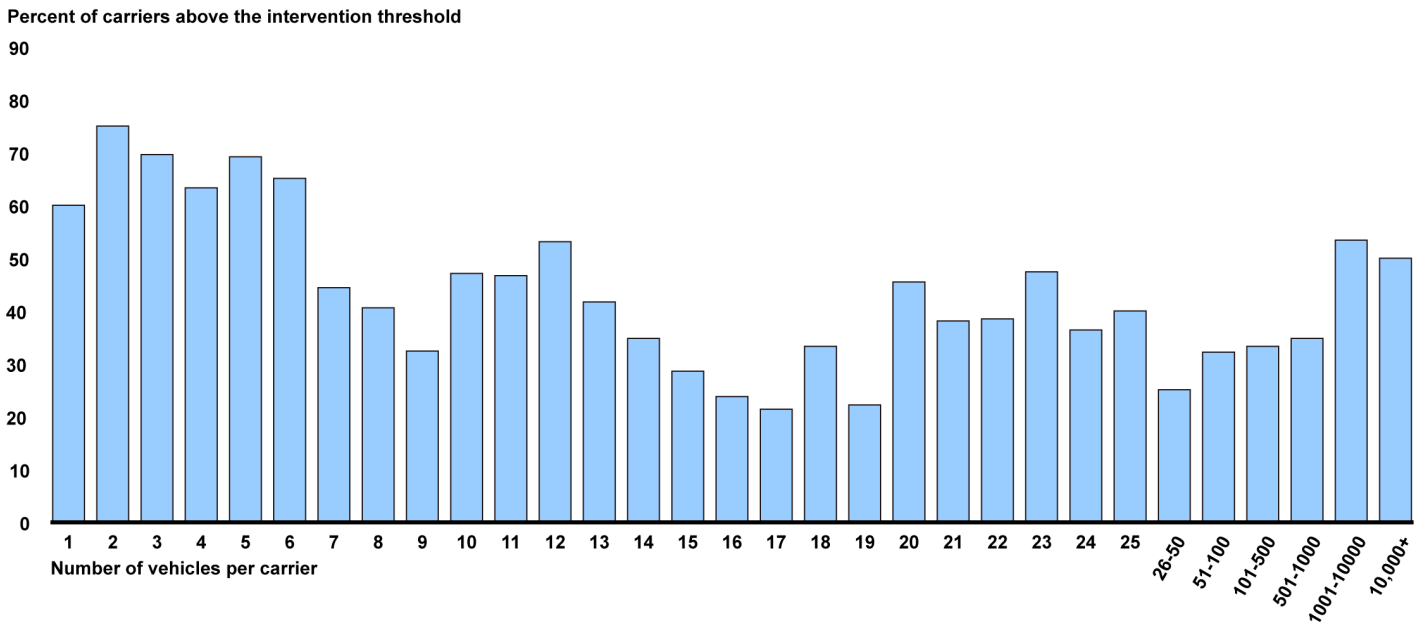
Source: GAO analysis of FMCSA data.

Figure 30: Distribution of FMCSA-Scored Carriers above the Vehicle Maintenance BASIC Threshold by Carrier Size



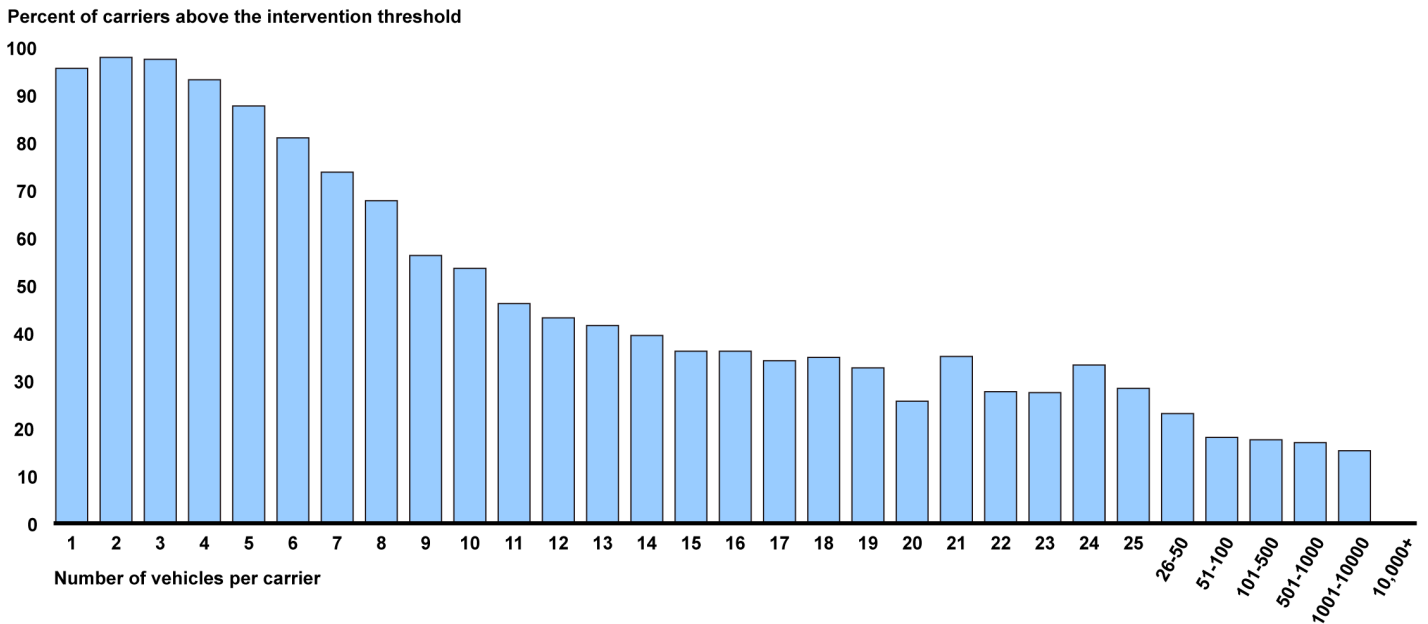
Source: GAO analysis of FMCSA data.

Figure 31: Distribution of FMCSA-Scored Carriers above the Hazardous Materials BASIC Threshold by Carrier Size



Source: GAO analysis of FMCSA data.

Figure 32: Distribution of FMCSA-Scored Carriers above the Crash Indicator Threshold by Carrier Size



Source: GAO analysis of FMCSA data.

Appendix VII: GAO Contact and Staff Acknowledgments

GAO Contact

Susan A. Fleming, (202) 512-2834 or flemings@gao.gov

Staff Acknowledgments

In addition to the individual named above, H. Brandon Haller, Assistant Director, Russell Burnett, Melinda Cordero, Jennifer DuBord, Colin Fallon, David Hooper, Matthew LaTour, Grant Mallie, Jeff Tessin, Sonya Vartivarian, and Joshua Ormond made key contributions to this report.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (<http://www.gao.gov>). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to <http://www.gao.gov> and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, <http://www.gao.gov/ordering.htm>.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on [Facebook](#), [Flickr](#), [Twitter](#), and [YouTube](#). Subscribe to our [RSS Feeds](#) or [E-mail Updates](#). Listen to our [Podcasts](#). Visit GAO on the web at www.gao.gov.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Website: <http://www.gao.gov/fraudnet/fraudnet.htm>

E-mail: fraudnet@gao.gov

Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Katherine Siggerud, Managing Director, siggerudk@gao.gov, (202) 512-4400, U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800 U.S. Government Accountability Office, 441 G Street NW, Room 7149 Washington, DC 20548

