Improving Motor Carrier Safety Measurement

DETAILS

132 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-46201-3 | DOI: 10.17226/24818

CONTRIBUTORS

Panel on the Review of the Compliance, Safety, and Accountability (CSA) Program of the Federal Motor Carrier Safety Administration; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

IMPROVING MOTOR CARRIER SAFETY MEASUREMENT

Panel on the Review of the Compliance, Safety, and Accountability (CSA) Program of the Federal Motor Carrier Safety Administration

Committee on National Statistics
Division of Behavioral and Social Sciences and Education

Transportation Research Board

A Report of
The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by Award No. DTMC7516C0001 from the United States Department of Transportation's Federal Motor Carrier Safety Administration. Support of the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation, a cooperative agreement from the National Agricultural Statistics Service, and several individual contracts. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-XX:
International Standard Book Number-XX:
Digital Object Identifier: https://doi.org/10.17226/24818

Additional copies of this publication are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: The National Academies of Sciences, Engineering, and Medicine. 2017. *Improving Motor Carrier Safety Measurement*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24818.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the National Academies of Sciences, Engineering, and Medicine to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Reports document the evidence-based consensus of an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and committee deliberations. Reports are peer reviewed and are approved by the National Academies of Sciences, Engineering, and Medicine.

Proceedings chronicle the presentations and discussions at a workshop, symposium, or other convening event. The statements and opinions contained in proceedings are those of the participants and have not been endorsed by other participants, the planning committee, or the National Academies of Sciences, Engineering, and Medicine.

For information about other products and activities of the National Academies, please visit nationalacademies.org/whatwedo.

PANEL ON THE REVIEW OF THE COMPLIANCE, SAFETY, AND ACCOUNTABILITY (CSA) PROGRAM OF THE FEDERAL MOTOR CARRIER SAFETY ADMINISTRATION

JOEL GREENHOUSE (*Cochair*), Department of Statistics, Carnegie Mellon University
SHARON-LISE NORMAND (*Cochair*), Department of Biostatistics and Health Care Policy, Harvard Medical School
MICHAEL BELZER, Department of Economics, Wayne State University
DAN BLOWER, University of Michigan Transportation Research Institute
LINDA BOYLE, Department of Civil and Environmental Engineering, University of Washington
MICHAEL DANIELS, Department of Statistics and Data Sciences, University of Texas at Austin
DON HEDEKER, Division of Epidemiology and Biostatistics, The University of Chicago
BRENDA LANTZ, Upper Great Plains Transportation Institute, North Dakota State University
DAN MCCAFFREY, Educational Testing Service
BRISA SANCHEZ, Department of Biostatistics, University of Michigan
ROBERT SCOPATZ, VHB, Inc.
JUNED SIDDIQUE, Feinberg School of Medicine, Northwestern University

MICHAEL L. COHEN, *Costudy Director*
ESHA SINHA, *Costudy Director*
RICK PAIN, *Consultant*
JACOB SPERTUS, *Consultant*
ANDREW YARGER, *Consultant*
AGNES GASKIN, *Administrative Assistant*
MICHAEL J. SIRI, *Program Coordinator*

COMMITTEE ON NATIONAL STATISTICS

T. F. SCOTT DARLING III, Federal Motor Carrier Safety Administration, U.S. Department of Transportation
MARIE THERESE DOMINGUEZ, Pipeline and Hazardous Materials Safety Administration, U.S. Department of Transportation
SARAH FEINBERG, Federal Railroad Administration, U.S. Department of Transportation
LEROY GISHI, Division of Transportation, Bureau of Indian Affairs, U.S. Department of the Interior, Washington, DC
JOHN T. GRAY II, Policy and Economics, Association of American Railroads, Washington, DC
MICHAEL P. HUERTA, Federal Aviation Administration, U.S. Department of Transportation
PAUL N. JAENICHEN, SR., Maritime Administration, U.S. Department of Transportation
THERESE W. MCMILLAN, Federal Transit Administration, U.S. Department of Transportation
MICHAEL P. MELANIPHY, American Public Transportation Association, Washington, D.C.
GREGORY G. NADEAU, Federal Highway Administration, U.S. Department of Transportation
MARK R. ROSEKIND, National Highway Traffic Safety Administration, U.S. Department of Transportation
CRAIG A. RUTLAND, U.S. Air Force Civil Engineer Center, Tyndall Air Force Base, Florida
REUBEN SARKAR, U.S. Department of Energy
BARRY R. WALLERSTEIN, South Coast Air Quality Management District, Diamond Bar, California
GREGORY D. WINFREE, Office of the Secretary, U.S. Department of Transportation
FREDERICK G. (BUD) WRIGHT, American Association of State Highway and Transportation Officials, Washington, DC
PAUL F. ZUKUNFT, U.S. Coast Guard, U.S. Department of Homeland Security

Acknowledgments

To begin, the panel is grateful to the Federal Motor Carrier Safety Administration (FMCSA) for providing the funds that made this study possible. We are particularly indebted to Joseph DeLorenzo, director, Office of Enforcement and Compliance, FMCSA, who provided three enormously useful presentations to the panel on various aspects of the Compliance, Safety Accountability Safety Management System (CSA/SMS), set up a meeting at FMCSA to learn more about the Motor Carrier Management Information System (MCMIS) database, and helped identify stakeholders for various aspects of the workings of SMS. It is difficult to imagine how the panel could have functioned as well as it did without his inputs and assistance. Jack Van Steenburg and Scott Valentine of FMCSA also provided the panel with excellent presentations, and FMCSA staff in general, including Martin Walker and Albert Alvarez, provided us with several technical reports and answered a number of questions on various aspects of CSA/SMS. We would also like to call out Olu Ajayi of FMCSA's Research Division and Dee Williams of the John A. Volpe National Transportation Systems Center, who were very forthcoming in answering many queries related to MCMIS, including during a meeting at FMCSA in December 2016, all of which greatly facilitated our use of the MCMIS database.

The panel also wishes to thank the presenters during our three information-gathering meetings. They included: Steve Bryan (Vigillo), Chris Burroughs (Transportation Intermediaries Association), Cary Catapano (National School Transportation Association), Tom DiSalvi (Schneider National), Jacqueline Duley (Engility Corporation), James Edwards (National Association of Small Trucking Companies), Jean Gardner (Central Analysis Bureau), Jay Grimes (Owner Operator Independent Drivers Association), H. Brandon Haller (Government Accountability Office), Julie Heckman (American Pyrotechnic Association), Collin Mooney (Commercial Vehicle Safety Alliance), Daniel Murray (American Transportation Research Institute), John Lannen (Truck Safety Coalition), Don Osterberg (retired from Schneider National), Ken Presley (United Motorcoach Association), Andria Sequin (Schneider National), Irwin Shires (Panther Premium), Rudolph Supina (DATTCO), Eric Teoh (Insurance Institute for Highway Safety), Jeff M. Tessin (Government Accountability Office), Pat Thomas (United Parcel Service), William Voss (International Civil Aviation Organization), Tom Weakley (Owner Operator Independent Drivers Association), Ann Williamson (University of New South Wales), and Shuie Yankelewitz (Central Analysis Bureau). All of these individuals spent a great deal of their time preparing these presentations to advance the panel's work. We would also like to thank Becky Weber (Prime Policy), who helped identify stakeholders for our second panel meeting.

The panel also is indebted to staff of the National Academies outside of the Committee on National Statistics. They included staff in the Transportation Research Board, especially Steve Godwin, who provided extremely useful advice on potential panel members. We are also very grateful to Frank Porto of IT services, who was instrumental in hosting and extracting the MCMIS data and making it available to CNSTAT staff and panel members.

The panel is very grateful as well to Jacob Spertus and Andrew Yarger for their data analysis that supported the conclusions and findings in the report. We are also greatly in the debt of Rick Pain, who delayed his retirement to provide us with critical information and advice throughout all phases of this study.

Michael Siri and Agnes Gaskin dealt with the complicated administrative aspects of such a study with great patience and understanding. Genie Grohman and Paula Whitacre did an outstanding job of technical editing of the report, greatly improving its readability and helping to communicate our message.

Finally, we wish to thank the panel members. The panel worked extremely well together with a great collaborative spirit, writing the majority of the chapters and conscientiously reviewing the products of other panel members. They were a wonderful group of people to get to know.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We wish to thank the following individuals for their review of this report: Rebecca Brewster, President, American Transportation Research Institute, Arlington, VA; Stephen V. Burks, Associate Professor of Economics and Management, University of Minnesota, Morris; Thomas C. DiSalvi, Vice President, Safety and Loss Prevention, Schneider National, Inc.; Carol A. Flannagan, Research Associate Professor, University of Michigan Transportation Research Institute and Director, Center for the Management of Information for Safe and Sustainable Transportation; Robert D. Gibbons, Professor of Medicine and Health Sciences, The University of Chicago; Feng Guo, Associate Professor, Department of Statistics, Virginia Tech Transportation Institute; H. Brandon Haller, Assistant Director, Physical Infrastructure Issues, U.S. Government Accountability Office; Jeff Hickman, Research Scientist, Center for Truck and Bus Safety, Virginia Tech Transportation Institute; Susan Paddock, Senior Statistician, RAND Corporation, Santa Monica, CA; and Greer Woodruff, Senior Vice President, Safety, Security and Driver Personnel, J.B. Hunt Transport Inc.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the report's conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Hal S. Stern, Donald Bren School of Information and Computer Sciences, University of California, Irvine, and Henry G. Schwartz, Jr., Consultant, St. Louis, MO. They were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Joel Greenhouse and Sharon-Lise Normand, *Cochairs*
Michael Cohen and Esha Sinha, *Costudy Directors*
Panel on the Review of the Compliance, Safety, and Accountability (CSA) Program of the Federal Motor Carrier Safety Administration

Contents

Glossary

Average Number of Power Units (APU): The number of straight trucks, truck tractors, and buses owned or leased by a commercial motor vehicle (CMV) carrier. This information can be updated multiple times during a 2-year period, and so the Federal Motor Carrier Safety Administration (FMCSA) uses a weighted average of up to three inputs to arrive at the average number of power units.

American Transportation Research Institute (ATRI): The trucking industry's not-for-profit research organization, an independent and autonomous research organization whose primary mission is research focused on the industry's safety and productivity.

Behavior Analysis and Safety Improvement Categories (BASICs): Seven scores— which generate seven percentile ranks—six of which are weighted frequencies of particular collections of violations that pertain to various types of safety problems, and one of which is a weighted frequency of crashes.

Commercial Motor Vehicle (CMV): A truck or a bus, which obscures the enormous variety of vehicles referred to by those terms.

Compliance Review (CR): An on-site investigation of a carrier's operations to assess whether it is meeting the safety fitness standards. Compliance reviews address hours of service, maintenance, driver qualifications, commercial drivers' license requirements, finances, crash frequency, and other safety and transportation records.

Compliance, Safety, Accountability (CSA): A data-driven safety compliance and enforcement program intended to reduce the frequency of commercial motor vehicle crashes, injuries, and fatalities. CSA consists of the Safety Measurement System (SMS); resulting interventions; and a Safety Fitness Determination (SFD) rating system, which is used to determine motor carrier safety.

Fatality Analysis Reporting System (FARS): A nationwide census that provides the National Highway Transportation Safety Administration and Congress with annual data on fatal injuries resulting from motor vehicle traffic crashes.

Fixing America's Surface Transportation Act (FAST ACT): A federal law (Public Law No. 114-94) that provided funding for surface transportation infrastructure planning and investment, authorizing $305 billion from 2016 through 2020 for "highway, highway and motor vehicle safety, public transportation, motor carrier safety, hazardous materials safety, rail, and research, technology, and statistics programs."

Federal Motor Carrier Safety Administration (FMCSA): A federal agency established in 2000 within the U.S. Department of Transportation. Formerly a part of the Federal Highway Administration, its mission is to prevent commercial motor vehicle-related fatalities and injuries. This includes enforcement of safety regulations; identifying high-risk carriers and commercial motor vehicle drivers; improving safety information systems and commercial motor vehicle equipment and technologies; and increasing safety awareness. FMCSA works collaboratively on its mission with other federal, state, and local enforcement agencies, the motor carrier industry, and labor and safety interest groups.

Hazardous Materials (HM): Any item or agent that has the potential to cause harm to humans, animals, or the environment.

Hours of Service (HOS): Hours-of-service regulations for commercial motor vehicle drivers limit the number of hours that drivers can operate their vehicles and limit the number of hours they can be at work in a 24-hour period and in a work week.

Inspectors/ Investigators: *Inspectors* are CVSA-certified state and local officials who carry out roadside inspections of commercial motor vehicles. *Investigators* are FMCSA employees who carry out safety audits, which include compliance reviews and other special investigations, at a carrier's headquarters.

Inspection Levels: Roadside inspections conducted by trained inspectors who are overseen and guided by the Commercial Vehicle Safety Alliance. The inspections consist of the following:

Level I – Full inspection: Examination of the driver's license, medical examiner's certificate, skill performance evaluation certificate (if applicable), use of alcohol and drugs, driver's record of duty status, hours of service, seat belt use, vehicle inspection reports, checks of brake systems, coupling devices, exhaust systems, frames, fuel systems, lighting devices, securement of cargo, steering mechanisms, suspensions, tires, van and open-top trailer bodies, wheels, rims and hubs, windshield wipers, emergency exits and /or electrical cables and systems in engine and battery compartments for buses, and hazardous materials requirements as applicable;

Level II – Walk-Around Driver/Vehicle Inspection: A subset of a Level I inspection involving only those items that can be inspected without physically getting under the vehicle;

Level III – Driver/Credential Inspection: Examination of the driver's license; medical examiner's certificate and skill performance evaluation (SPE) certificate; driver's record of duty status; hours of service; seat belt; vehicle inspection report(s); and hazardous materials requirements;

Level IV – Special Inspections: One-time inspections of particular items, often in support of a study;

Level V – Vehicle-Only Inspection: Subset of a full inspection that can be done without a driver present;

Level VI – Inspection for Radioactive Material: Examination for select radiological shipments.

Intervention: Various ways in which FMCSA can inform a carrier that its safety performance is deficient. The primary types are notification letters, investigations, and various types of follow-on activities, which include at the extreme placing a carrier out of service.

Investigation: Safety audits that include compliance reviews and take place at a carrier's headquarters.

Million Vehicle Miles Traveled (MVMT): A measure of exposure useful as a denominator for crashes, due to their rarity.

Model Minimum Uniform Crash Criteria (MMUCC): A voluntary guideline for crash data element definitions suitable for states to use in designing police crash reports (PCRs) and statewide crash databases.

Motor Carrier Management Information System (MCMIS): A database containing information from states on commercial motor vehicle inspections, violations discovered during inspections, crashes, and registration data on all carriers. This is the dataset that supports SMS.

Motor Carrier Safety Assistance Program (MCSAP): A federal program that provides grants to states to assist in the reduction of crashes involving commercial motor vehicles, including

those involving hazardous materials, by supporting consistent, uniform, and effective safety programs.

Out-of-Service (OOS): If during a commercial motor vehicle's roadside inspection certain violations are discovered, the driver and/or the vehicle may be placed temporarily out-of-service.

Police Accident Reports (PARS): Reports from police officers arriving at the scene of an accident containing the information that they collected from the placement of the vehicles, skid marks, interviews of those involved and eye witnesses, and other information.

SafeStat: Predecessor that FMCSA employed prior to the institution of the Safety Measurement System (SMS).

Safety Measurement System (SMS): The current approach by the FMCSA to identify unsafe CMV carriers for the purpose of generation of interventions.

State Safety Data Quality (SSDQ): A variety of forms of technical assistance, grants, and dispute resolution to improve data on commercial motor vehicle crashes.

Straight Truck / Combination Truck: A *straight* truck is one where the body of the vehicle is in a single piece. A *combination* truck is typically a tractor-trailer where a cab is pulling a trailer.

Vehicle Miles Traveled (VMT): The number of miles all the trucks of a carrier have traveled during a specific period of time.

Summary

Commercial motor vehicle companies are responsible for moving freight and passengers in a safe manner over the nation's highways. To help ensure a high standard of safety, the Federal Motor Carrier Safety Administration (FMCSA), with its mission "to reduce crashes, injuries and fatalities involving large trucks and buses," makes use of the Compliance, Safety, and Accountability Program, and, in particular, the Safety Measurement System (SMS), to rank commercial motor vehicle (CMV) carriers by the degree to which they operate safely. This ranking is a function of the frequency of different groups of violations assessed during (mainly) roadside inspections, or is a function of the frequency of crashes that a carrier has, or both over the most recent 2-year period. SMS is used as a prioritization tool that allows the agency to target for interventions motor carriers that are operating unsafely and therefore are likely to be at higher risk for future crashes.

SMS partitions 899 possible violations that can arise from roadside inspections into six groups and forms a metric for each of these groups that are weighted frequencies of violations. SMS augments these six metrics with another metric, crash frequency (again weighted). These seven metrics are referred to as Behavioral Analysis and Safety Improvement Categories (BASICs). The noncrash BASICs make use of groups of violations that are associated with similar types of unsafe practices: Unsafe Driving, Hours of Service Compliance, Vehicle Maintenance, Controlled Substances/Alcohol, Hazardous Materials Compliance, and Driver Fitness. The seventh BASIC is referred to as Crash Indicator. For each carrier with enough inspections, violations, and crashes to meet FMCSA's data sufficiency standards, these seven measures are computed, and carriers are ranked within the peer groups to see which have the greatest frequency of crashes or violations. FMCSA has set thresholds within each of these groups, and carriers that rank above these thresholds are subject to a range of interventions, which include warning letters, on-site investigations, and more serious actions such as fines and suspension of operations.

Some stakeholders and outside reviewers, among others, have criticized the Safety Management System for, among other things: (1) making use of highly variable assessments, (2) not accounting for crashes where the motor carrier is not at fault, (3) including carriers that have very different tasks in the same peer groups, (4) using measures that are sensitive to effects from one or more individual states, (5) using measures that are not predictive of a carrier's future crash frequency, or (6) using measures that are not reflective of a carrier's efforts to improve its safety performance over time.

Given the stakes involved, it is important to examine SMS to see whether these and other criticisms are valid, and to examine the performance of this system to see if improvements can be made. The need for this review of SMS was written into the Fixing America's Surface Transportation (FAST) Act of 2015, in which it is recommended that FMCSA fund a study by the National Academies of Sciences, Engineering, and Medicine (NASEM) to evaluate SMS. Two units in NASEM, the Committee on National Statistics in collaboration with the Transportation Research Board, began work in March 2016, convening the Panel on the Review of the Compliance, Safety, and Accountability Program of the Federal Motor Carrier Safety Administration for this congressionally mandated study. The panel was charged with analyzing the ability of SMS measures to discriminate between low- and high-risk carriers, assess the

public usage of SMS, review the data and methodology used to calculate the measures, and provide advice on additional data collection and safety assessment methodologies. The panel met in June of 2016 and three additional times prior to issuing this report. These meetings, reinforced by additional research and analysis, resulted in six recommendations by the panel, which are presented in this Summary and explained in more detail throughout the full report.

The Federal Motor Carrier Safety Administration (FMCSA) has as its mission to prevent commercial motor vehicle-related injuries and fatalities. The SMS is a prioritization tool that allows the agency to identify motor carriers with safety compliance problems for intervention by FMCSA. The agency's Motor Carrier Management Information System (MCMIS), which contains data from commercial motor vehicle crashes, carrier registrations, commercial motor vehicle inspections, and inspection violations, is used for input into SMS. The state-based inspection system that feeds into MCMIS, using inspectors trained by the Commercial Vehicle Safety Alliance (CVSA), inspects more than 3 million commercial motor vehicles every year in the United States in order to determine whether truck and bus carriers are operating in violation of safety regulations. FMCSA deserves considerable credit for making use of these data in an attempt to discriminate between safe and unsafe motor carriers.

ASSESSMENT OF THE CURRENT SAFETY MEASUREMENT SYSTEM

Multiple factors contribute to crashes, many of which are not present in MCMIS. Given that, and the relative rarity of crashes for small carriers, the panel agrees that development of a crash prediction model based on carrier-level behavior using MCMIS data is not a productive way to approach the problem of discrimination between safe and unsafe commercial motor carriers. With SMS, FMCSA has instead adopted a sensible, related approach based on prevention rather than prediction. That is, SMS has the objective of identifying carriers that give too little priority to practices indicative of safety performance. By intervening with those carriers, the hope is to encourage them to modify their behavior and, by so doing, reduce future crashes. We believe that the general approach taken by SMS is sound, and shares much with similar programs in other areas of transportation safety. Further, we have examined, to the extent possible, the various issues that have been raised in criticism of SMS. We have found, for the most part, that the current SMS implementation is defendable as being fair and not overtly biased against various types of carriers, to the extent that data on MCMIS can be used for this purpose.

However, we believe some features of SMS implementation can be improved upon, and some of the details of the implementation are ad hoc and not fully supported by empirical studies. Many of these details of implementation would be easily addressed if the algorithm currently used were replaced by a statistical model that is natural to this sort of discrimination problem. Therefore, we reached the following conclusion:

Conclusion: SMS is structured in a reasonable way, and its method of identifying motor carriers for alert status is defendable. However, much of what is now done is ad hoc and based on subject-matter expertise that has not been sufficiently empirically validated. This argues for FMCSA adopting a more statistically principled approach that can include the expert opinion that is implicit in SMS in a natural way.

A MORE NATURAL STATISTICAL MODEL

The general approach taken by SMS can be shown to be related to item response theory (IRT) models that have been successfully applied in several similar contexts. These types of models accumulate zero-one responses to "tests," using the results to differentiate between "test takers" that have different latent traits. These are models that describe the relation between where an individual falls on the continuum of a given construct, such as depression, and the probability that he or she will give a particular response to a scale item designed to measure that construct. In IRT, such a construct is called a latent trait, because that trait is assumed to underlie and directly influence responses to items on the scale designed to measure that trait. Examples include assessment of elementary and high school teachers, and determination of hospital rankings. In these and other areas of application, IRT models have been shown to be very effective. Given various theoretical advantages, we recommend the following:

Recommendation: FMCSA should develop the suggested IRT model over the next 2 years. If it is then demonstrated to perform well in identifying motor carriers for alerts, FMCSA should use it to replace SMS in a manner akin to the way SMS replaced SafeStat. Specifically, IRT models would have the following specific advantages over SMS:

1) Instead of severity weights being based on expert opinion or dated empirical information, the item discrimination parameters are estimated based on a combination of current observed data and expert opinion, and ultimately on data alone;
2) IRT models can enhance the transparency of the evaluation system;
3) They support the direct estimation of variability of scores and ranks;
4) They can account for the probability of being selected for inspection;
5) They can provide a basis with which to evaluate how data insufficiency could impact safety ratings of carriers;
6) They can provide a basis to more rigorously evaluate the structure of the current BASICs, including which violations go into which BASIC;
7) They can provide for a natural way to examine the issue of further stratification;
8) They can provide for the possibility that safety is inherently multidimensional, which could inform how many BASICs are needed in the SMS model;
9) They can take account of time and thereby inform about the proper time weights in SMS;
10) They can allow for the addition of new safety measures as they become available, without having to start from scratch;
11) They can produce ranking ranges (by sampling from the posterior distribution of theta) to better understand overlap in the rankings (i.e., uncertainty);
12) They can adapt to changes in safety over time.

DATA IMPROVEMENTS

We considered data improvements in two respects. First, there are possible improvements to the variables collected in MCMIS that would not require major changes in what is currently

done. Second, there are variables that would benefit SMS that would need alternative sources for their collection.

Improvement of MCMIS Data

The two most important areas in which improvements could be made to the information that MCMIS collects are in exposure data and crash data. While updates are required for data on vehicle miles traveled (VMT) and the average number of power units (APU) every 2 years, the impact of flawed or out-of-date VMT and APU data on SMS percentile ranks may not be fully appreciated by the carriers. Therefore, increased efforts are needed to acquire better data. Also, a sizeable fraction of crash data is missing, and these data are collected in a nonstandard manner across states. Further, much of the information provided by police reports is not represented on MCMIS. Therefore, we make the following recommendation:

Recommendation: FMCSA should continue to collaborate with states and other agencies to improve the quality of MCMIS data in support of SMS. Two specific data elements require immediate attention: carrier exposure and crash data. The current exposure data are missing with high frequency, and data that are collected are likely of unsatisfactory quality. Further, to improve the exposure data collected involves not only collecting higher-quality VMT data, but also collecting this information by state and by month. This will enable SMS to (partially) accommodate existing heterogeneity in the environments where carriers travel. Crash data are also missing too often. Also, there is information available from police reports currently not represented on MCMIS that could be helpful in understanding the contributing factors in a crash. Such information could help to validate the assumptions linking violations to crash frequency. To address these issues, FMCSA should support the states in collecting more complete crash data, and in universal adoption of the Model Minimum Uniform Crash Criteria, as well as developing and supplying the code needed to automatically extract the data needed for the MCMIS crash file.

Improvement through Additional Variables and Possible New Sources

The information available on MCMIS is limited in terms of the ability to determine the factors that contributed to a crash. As a previous National Academies of Sciences, Engineering, and Medicine panel (2016) argued, we believe that it is reasonable to think of the causes of CMV crashes grouped into four categories, due to: (1) characteristics of the driver, (2) characteristics of the vehicle, (3) the driving environment, and (4) practices and procedures of the carrier. MCMIS is relatively good at capturing many characteristics of the vehicle and some aspects of the driver and the environment. However, it is incomplete regarding carrier operations. Since SMS is founded on the belief that a substantial fraction of crashes are due at least in part to carrier operations—and it is those operations that SMS is attempting to modify—it is clearly important to consider how to gain knowledge of carrier-related factors when making improvements to SMS. Further, in any study of which factors contributed to crashes, omission of any important (confounding) factors impairs the analysis.

Therefore, it would be extremely useful to know the carrier's cargo, the driver turnover rate, and the level of driver compensation. Then, should FMCSA issue an intervention to a carrier, it would be informative to see what aspects of carrier operations were modified to try to

address the intervention. Towards this end, we suggest that FMCSA look into how the following carrier characteristics might be collected externally to MCMIS:

- **Information on turnover rate:** This information could be very predictive of a company's treatment of its employees, which could be related to safety operations. In addition, a low turnover rate is likely associated with employment of drivers with longer tenures and hence greater experience.
- **Information on type of cargo carried:** Since current questions on type of business are producing lower-quality information, it might be preferable to ask a carrier about their typical cargos. The response to that question is nearly the same as type of business and might be easier to answer. (This could certainly be collected through the MCS-150.)
- **Information on compensation level and method:** It is known that drivers who are better compensated, and those not compensated as a function of miles traveled, have fewer crashes.
- **Better information on exposure:** We believe that state tax information is a possible source of high-quality VMT data, and therefore, we suggest that FMCSA interact with state taxing authorities to see whether an interagency agreement can be struck to share this information in support of SMS. In addition, at the end of 2017, electronic on-board recorders (EOBRs) will be required for most carriers, and results for all carriers could be reported to FMCSA. Having the number of vehicle miles traveled at the end of the year would be extremely easy to produce and would be definitive.

We, therefore, make the following recommendation:

Recommendation: FMCSA should investigate ways of collecting data that will likely benefit the recommended methodology for safety assessment. This includes data on carrier characteristics—including information on driver turnover rate, type of cargo, method and level of compensation, and better information on exposure. This additional data collection will likely require additional funds for research and development of the data collection instrument, and greater collaboration between FMCSA and the states as to how to undertake this new data collection effort so that it is standardized across the states. Protection and use of carrier specific data must be addressed as well.

TRANSPARENCY, REPRODUCIBILITY, AND PUBLIC DISCLOSURE OF SAFETY RANKINGS

SMS percentile ranks have very important implications for CMV carriers. Hence, it would be useful if the CMV community were able to reproduce the SMS measures and percentile ranks, that researchers have easier access to the SMS algorithm and the MCMIS database, that FMCSA better communicate with researchers about how SMS functions, and that carriers be able to know the implications of recent violations and crashes on their measures and percentile ranks. Therefore, we make the following recommendation:

Recommendation: FMCSA should structure a user-friendly version of the MCMIS data file used as input to SMS without any personally identifiable information to facilitate its use by external parties, such as researchers, and by carriers. In addition, FMCSA should make user-friendly computer code used to compute SMS elements available to individuals in accordance with reproducibility and transparency guidelines.

The panel was asked to comment on whether SMS percentile ranks should be made public. We are unable to recommend to FMCSA whether to make all SMS percentile ranks public. An understanding of the consequences of public consumption of the information requires a formal evaluation, possibly designed using randomization or controlled release of specific components of the SMS percentiles. In particular, what is needed to know is the current operating characteristics of SMS. That is, given that SMS can assess whether a carrier was one that should or should not have received an intervention given its safety behavior, what is the false positive and false negative rate of SMS? Given this, we make the following recommendation:

Recommendation: FMCSA should undertake a study to better understand the statistical operating characteristics of the percentile ranks to support decisions regarding the usability of public scores.

SMS percentile ranks are a relative metric, and so a motor carrier's efforts towards improving its safety performance will not be reflected in the percentile ranks if other carriers in its peer group have improved even more. On the other hand, a relative score has the advantage that if FMCSA sets an absolute standard of performance, since the entire industry is getting progressively safer, the standard will at some point become irrelevant. Having a relative metric permits FMCSA to keep pressing for better performance. Further, FMCSA operates on a fixed budget, and how it functions is consistent with a relative measure. Since there are advantages of both relative and absolute measures, we believe that FMCSA should strongly consider use of a two-dimensional measure that takes into consideration both the SMS score and the percentile rank, using some objective formula, to decide on which carriers will receive interventions.

Recommendation: Given that there are good reasons for both an absolute and a relative metric on safety performance, FMCSA should decide on the carriers that receive SMS alerts using both the SMS percentile ranks and the SMS measures, and the percentile ranks should be computed both conditionally within safety event groups and over all motor carriers.

1
Introduction

Every year roughly 100,000 fatal and injury crashes occur in the United States involving large trucks and buses (Federal Motor Carrier Safety Administration, 2015). The Federal Motor Carrier Safety Administration (FMCSA) in the U.S. Department of Transportation has as its mission: "… to reduce crashes, injuries, and fatalities involving large trucks and buses" (FMCSA, 2017). FMCSA believes that a sizeable fraction of these crashes could be prevented through better safety management. This enhanced emphasis could include instituting carrier policies that reduce fatigued driving in various ways, that motivate drivers to drive under the speed limit, or that dictate conscientious attention to maintenance. In addition, a carrier could institute better hiring practices; more regularly monitor the health of its drivers, including their use of drugs and alcohol; give greater priority to driver training; and acquire technology to assist in collision avoidance.

FMCSA uses information that is collected on the frequency of approximately 900 different violations of safety regulations discovered during (mainly) roadside inspections to assess motor carriers' compliance with Federal Motor Carrier Safety Regulations (FMCSRs), as well as to evaluate their compliance in comparison with their peers. Through use of this information, FMCSA identifies carriers to receive its available interventions in order to reduce the risk of crashes across all carriers. Approximately 3.5 million commercial motor vehicle (CMV) roadside inspections are conducted each year by specially trained inspectors, often state police and other enforcement officers, who are guided by the inspection standards and certification requirements of the Commercial Vehicle Safety Alliance (CVSA). The inspections, which encompass six different levels, involve checking a subset of all possible violations to assess various aspects of the compliance of the driver and vehicle. (See the Glossary for a description of the six inspection levels.) The identification of violations from these inspections and all associated data (including time and date of inspection) are input into the Motor Carrier Management Information System (MCMIS) and used by FMCSA to identify those carriers that are out of compliance with federal regulations and that are viewed the best candidates for interventions. MCMIS, primarily an administrative database that contains the results of carrier registrations, inspections, inspection violations, investigations, and crash involvements, provides all of the data input into these assessments. FMCSA deserves considerable credit for making use of this administrative dataset to discriminate between safe and unsafe motor carriers.

Carriers found to have frequent violations are subject to interventions from FMCSA. This can be in the form of warning letters (about 20,000 sent out a year), further monitoring, different types of investigations (about 15,000 a year), as well as suspension or revocation of the right to operate (around 800 a year). In addition, CMV crashes of a certain gravity are also input into MCMIS, and used by FMCSA both as an additional predictor for safety performance that can result in an intervention, and as a measure of the future crash risk of a carrier that can be used to validate the assumption that increases in safety performance can lead to reductions in crash risk. Since crash risk is dependent on the amount of driving done, crash *rate* (crashes divided by vehicle miles traveled [VMT]) is the metric that FMCSA uses to assess crash risk.

FMCSA's Safety Measurement System (SMS) identifies carriers for intervention. SMS uses MCMIS data to produce metrics to discriminate between carriers that are least compliant

with regulations and therefore in need of an intervention, and those carriers that are operating more safely by generally complying. SMS is used to evaluate about 550,000 active motor carriers (see Table 6-1 in Chapter 6), a population that is heavily skewed by carrier size, ranging from those with a single motor vehicle to those with tens of thousands of vehicles. Owner-operator companies account for 44 percent of the active motor carriers. Also, the largest 5 percent of motor carriers accounts for 67 percent of the CMV fleet.

The total of 550,000 active carriers entails a good deal of churn, as about 30,000 of these carriers go out of existence each year, to be replaced by about the same number of new carriers. Given the data sufficiency standards used in SMS, FMCSA generates SMS percentiles for roughly 200,000 of the 550,000 active carriers.

In SMS, the approximately 900 types of violations are grouped into six Behavior Analysis and Safety Improvement Categories (BASICs). The frequencies of violations in these categories, weighted by time and severity (which is an assessment of how relevant a violation is to future crash frequency), are divided either by the time-weighted total number of relevant inspections for that BASIC, or by an estimate of VMT for the carrier, to produce the six measures. The seventh BASIC looks at weighted (by crash severity) crash frequency over the previous 2-year period. Carriers are grouped essentially by size categories and the degree to which the carrier uses combination vehicles, so they are compared to other carriers of a similar size and type. Then, within these groupings, the BASIC measures for carriers are ranked from low to high, and each carrier is assigned its resulting percentile rank, which is the rank converted to a percentile between 0 and 100. High BASIC percentiles—indicating that a carrier was worse than a large percentage of similarly sized carriers—result in an intervention, with the overall goal being to incentivize carriers to adopt safe practices that will reduce their frequency of serious crashes in the future. Since some of these interventions are resource-intensive, SMS is a workforce prioritization tool for FMCSA.

THE FAST ACT AND THE CURRENT STUDY

Some stakeholders and outside reviewers have critiqued the SMS program concerning various aspects of its algorithms and structural assumptions. For instance, one issue that has been raised is that rather than percentile ranks, the measure used should be an absolute,[1] rather than a relative, assessment of safety behavior. Otherwise, a carrier could be improving its safety performance, but if its peers are all showing greater improvements, the carrier would have increasingly high percentile ranks suggesting that the carrier is becoming less safe over time. Further, little information is available for the safety performance of small carriers, especially owner-operators that typically involve a single driver operating a single truck. Such a carrier would likely have very few inspections over a 2-year period. As a result, the minimum amount of information needed to calculate BASIC measures using the SMS algorithm has been questioned. In addition, concerns have been put forward about the role of state enforcement priorities and the effects on the frequency with which violations are issued, whether the algorithm should stratify based on the types of vehicles the carrier operates (e.g., separate bus and truck strata), and whether crashes generally viewed as nonpreventable should be ignored or at least downweighted by the algorithm.

[1]Examples of absolute measures of safety are crash rate, number of crashes per mile, violation rate, and number of violations per inspection.

The overall concern raised by these various criticisms has motivated this review, which was written into Section 5221 of the Fixing America's Surface Transportation (FAST) Act of 2015, mandating that the U.S. Department of Transportation request that the National Academies of Sciences, Engineering, and Medicine conduct a study analyzing SMS. The NASEM units tasked to carry out this study were the Committee on National Statistics, as the lead, with the assistance of the Transportation Research Board. The resulting expert panel, the Panel on the Review of the Compliance, Safety, and Accountability Program of the Federal Motor Carrier Safety Administration, was convened in March of 2016, and was charged by this Congressional mandate to examine the effectiveness of the use of the percentile ranks produced by SMS for identifying high-risk carriers, and if not, what alternatives might be preferred. In addition, the panel was asked to evaluate the accuracy and sufficiency of the data used by SMS, to assess whether other approaches to identifying unsafe carriers would identify high-risk carriers more effectively, and to reflect on how members of the public use the SMS and what effect making the SMS information public has had on reducing crashes. The panel first met in June of 2016 and then three additional times prior to issuing this report. The agendas for the open parts of the first three meetings are given in Appendix A. In addition, FMCSA provided the panel with a version of the MCMIS data and the SMS algorithm, details of which are contained in Appendix B.

STATEMENT OF TASK

The following is the statement of task that the panel was charged to address:

An ad hoc panel will carry out a consensus study in response to Section 5211 of the Fixing America's Transportation (FAST) Act of 2015. The purpose of this study is to analyze:

a. The accuracy with which the Behavioral Analysis and Safety Improvement Category (BASIC) safety measures used by the Compliance, Safety, Accountability (CSA) Safety Management System (SMS):
    i. Identify high risk carriers.
    ii. Predict or are correlated with future crash risk, crash severity, or other safety indicators for motor carriers, including the highest risk carriers.
b. The methodology used to calculate BASIC percentiles and identify carriers for enforcement, including the weights assigned to particular violations, and the tie between crash risk and specific regulatory violations, with respect to accurately identifying and predicting future crash risk for motor carriers.
c. The relative value of inspection information and roadside enforcement data.
d. Any data collection gaps or data sufficiency problems that may exist and the impact of those gaps and problems on the efficacy of the CSA program.
e. The accuracy of safety data, including the use of crash data from crashes in which a motor carrier was free from fault.
f. Whether BASIC percentiles for motor carriers of passengers should be calculated separately than for motor carriers of freight.
g. The differences in the rates at which safety violations are reported to FMCSA for inclusion in the SMS by various enforcement authorities, including States, territories, and Federal inspectors.

h. How members of the public use the SMS and what effect making the SMS information public has had on reducing crashes and eliminating unsafe motor carriers from the industry.

The study should also consider:

1. Whether the SMS provides comparable precision and confidence, through SMS alerts and percentiles, for the relative crash risk of individual large and small motor carriers.
2. Whether alternatives to the SMS would identify high risk carriers more accurately.
3. The recommendations and findings of the Comptroller General of the United States and the Inspector General of the Department of Transportation, and independent review team reports, issued before the date of the act.

The panel will issue a report with findings and recommendations at the end of the study.

For clarification, the Compliance, Safety, Accountability (CSA) program has components in addition to SMS. CSA is an overall agency, data-driven, safety compliance and enforcement program. Its objectives are to assess and intervene with a large segment of the industry, maximize the impact on large truck and bus safety, respond early to unsafe operation using a broad array of interventions, and make more effective use of resources. The components of CSA are (1) SMS, (2) the interventions process, and (3) safety fitness determinations that identify carriers that are not fit to operate CMVs. CSA is based on carrier's performance in seven BASICs and investigation results. The idea is a carrier's significant pattern of noncompliance would be documented. SMS, which produces the BASIC percentiles, is itself comprised of a Carrier Safety Measurement System (CSMS) and a Driver Safety Measurement System (DSMS). This study is not concerned with non-SMS aspects of CSA, and it is concerned only with CSMS, not with DSMS, but we will refer to our topic as SMS in the remainder of this report.

HOW SMS WORKS

SMS partitions nearly 900 possible violations that can arise from primarily roadside inspections into six groups, the Behavior Analysis and Safety Improvement Categories (BASICs). The violations in each BASIC are grouped together because they are associated with similar types of unsafe practices: Unsafe Driving, Hours of Service Compliance, Vehicle Maintenance, Controlled Substances/Alcohol, Hazardous Materials Compliance, and Driver Fitness. For each carrier that has enough data to meet FMCSA's data sufficiency standards for each BASIC, the SMS algorithm produces measures which, for five of these six noncrash BASICs, are weighted violation frequencies divided by weighted counts of inspections, for any violations within a given BASIC. Unsafe Driving uses a different denominator, which is essentially an estimate of vehicle miles traveled. The seventh BASIC, referred to as the Crash Indicator BASIC, is a weighted crash frequency, where the weights are time weights and crash severity weights. Then, within groups of similarly sized carriers, referred to as safety event groups, the carriers are ranked from low to high for each BASIC, and the percentile ranks (expressed between 0 and 100 percent) for each of the seven BASICs are computed for each

carrier. Separately for each BASIC, those carriers with percentile ranks above thresholds set by FMCSA may receive interventions, which range from a warning letter to an on-site investigation. Finally, FMCSA has, until recently, published the percentile ranks for five of the seven BASICs. There may be important implications in making the measures public, since those carriers that have measures that exceed the thresholds can lose business and have their insurance rates increased.

ISSUES RAISED IN CRITICISM OF SMS

SMS has been evaluated by the Government Accountability Office (2014), the American Transportation Research Institute (2012, 2014, 2015), the Independent Review Team (2014) of the U.S. Department of Transportation, Green and Blower (2011), and others. Before we present the issues that have been raised in these various reviews, we provide some perspective on the nature of what FMCSA is attempting to accomplish and a sense of what level of performance is likely.

We assert that a direct approach to the problem of predicting which motor carriers will have the highest future crash risk is extremely difficult. This is primarily because crashes are a rare phenomenon, each crash is associated with multiple contributing factors, and the data associated with many of these factors are either unrecorded or unidentifiable. With SMS, FMCSA has instead adopted a related approach focused on prevention and not prediction. FMCSA identifies carriers engaging in observable behaviors that have been shown to be associated with future crash risk, and it intervenes with those carriers to encourage them to adopt safer practices in the hopes of reducing future crashes. The objective of FMCSA's SMS is to identify carriers that are giving too little priority to behaviors and practices indicative of safety performance. FMCSA uses data on the frequency that inspections of CMVs are found to have violations of various safety regulations, investigations, along with the frequency of crashes, and thereby directly measuring important components of safety practices for carriers. By doing this, FMCSA's objective becomes an easier problem of discriminating between those carriers that do and do not emphasize various aspects of safe operations.

Specific Criticisms

A number of specific criticisms of SMS were documented in past reviews and repeated in presentations to the panel.

Not All BASICs Are Predictive: Evaluations of SMS have shown two of the seven BASICs, Driver Fitness and Controlled Substances /Alcohol, have low correlations with future crash rates, which raises a question about their utility as part of SMS.

Data Sufficiency Standards: As noted above, FMCSA does not apply SMS to carriers that have not had a sufficient number of inspections, violations, and crashes. There is no question that SMS measures and percentiles for carriers with just a few inspections would be extremely variable. However, FMCSA points out that if they adopt a more stringent standard, they will be excluding a very large fraction of the CMV population from their purview. This is the trade-off that must be considered.

Use of an Absolute Versus a Relative Metric: The use of percentile ranks to decide which carriers receive interventions is a relative scoring method, where each carrier has to do better than their peers to receive a lower (improved) ranking. In contrast, absolute measures inform users about a carrier's change in performance over time, not in comparison with any peers. Given the constant improvement in various aspects of CMV driving, including better technologies, a relative rank has the important advantage that such improvements do not make the metric irrelevant over time. Also, relative ranks are consistent with FMCSA's fixed budget. However, carriers that are improving against an absolute standard feel that the improvement should relieve them from receiving interventions.

Use of Data from Nonpreventable Crashes: The American Transportation Research Institute (ATRI, 2015) has raised the point that many crashes involving CMVs are not the fault of the CMV drivers, such as when they get rear-ended. It can be argued that such crashes do not help discriminate between unsafe and safe carriers and therefore should be removed from the SMS algorithm. The argument for the retention of such crash data is that based on effectiveness testing, all crashes are useful in discriminating between safe and unsafe carriers. Further, it is not always easy to determine preventability and doing so would require substantial additional resources.

State Differences in Rate of Inspections and Violations: There are differences from state to state in road type, congestion, and in prevalence of ice, degree of visibility, and other conditions. Since the driving environment varies state by state, this can have an impact on crash frequency. Further, ATRI (2014) has provided strong evidence that there are significant differences among states in administration of the CVSA inspection system. This raises the question as to whether SMS is unfair to carriers that operate in states that issue more frequent violations (or that issue a higher percentage of violations that have a higher severity weight).

Stratification of SMS in Addition to Safety Event Groups: Besides safety event groups, which as we have noted are essentially based on carrier size, the only formal stratification that SMS makes use of is the stratification of the carrier population into carriers where, for truck carriers, more than 70 percent of the trucks are "combination" as opposed to "straight," with a related stratification for motorcoach carriers. (See Glossary for explanation of these terms.) In addition, there are different thresholds for interventions for different types of carriers, including passenger carriers and hazardous material carriers, which is also a form of stratification. Given the heterogeneity of the trucking and motorcoach industry, a greater degree of stratification has been suggested by critics of SMS so that carriers are compared to peers who have similar operations. FMCSA counters such suggestions by pointing out that the greater the degree of stratification used, the fewer peers a carrier can be compared to and the less useful are the ranks in the tails of distributions.

Better Measures of Exposure: A relatively fair comparison between two carriers requires that the total number of violations for unsafe driving or the total number of crashes be standardized by the number of vehicle miles traveled. That is, the greater the miles traveled, the greater the likelihood of an inspection or violation, and the greater the likelihood for a crash. Therefore, the number of violations or the number of crashes is divided by the number of total vehicle miles traveled for a carrier, resulting in violations or crashes per mile. Unfortunately, the current data

from MCMIS on vehicle miles traveled for a carrier is often missing, and so FMCSA uses a more highly reported figure on the average number of power units, multiplied by a utilization factor (see Appendix B) that is a function of vehicle miles traveled, as a proxy for vehicle miles traveled. The resulting denominator is often based on the proxy rather than actual mileage, and therefore it is important to improve the quality of data on vehicle miles traveled so that this normalization is an accurate reflection of the motor carrier's operations.

Quality of MCMIS Crash Data: While crash data collection is not solely FMCSA's responsibility, evaluations of state reporting have shown significant underreporting of qualifying crashes by the states. The underreporting varies by crash severity (with fatal crashes less likely to be missed), truck type (where crashes with smaller qualifying trucks less likely to be reported), type of enforcement agency that covered the crash, and other factors.

Appropriateness of Severity Weights and Violation Coding: The severity weights were established by FMCSA using a combination of subject-matter expertise and empirical work. They are used to give additional weight to violations that are more closely associated with future crash risk. Severity weights and violation coding have been criticized, since essentially equivalent violations can be assigned to different violation codes, which can result in severity weights that can differ by two or more times, depending on the specific violation codes cited.

Currently Uncollected Variables That Might Substantially Improve SMS: MCMIS is relatively effective at capturing many characteristics of the commercial driver and the vehicle and some aspects of the environment. However, it is incomplete regarding carrier operations. SMS is in some sense based on the belief that some crashes are due in part to carrier operations. Hence, it is important to gain knowledge of those carrier factors when considering improvements to SMS, as these factors are what FMCSA is attempting to change. For example, it would be extremely useful, though currently not feasible, to have information on all 550,000 motor carriers as to their primary business, how they schedule drivers, and their driver turnover rate.

Sparsity of Some Violations: Many of the 899 violations are only occasionally cited on any inspections and so have little tangible impact on SMS for the great majority of carriers. It seems reasonable to believe that the greater the number of violations for which data are collected, the lower the quality of the information, which would advocate for the removal of such violations from data collection. On the other hand, such violations, for circumstances that are admittedly somewhat rare, might still be extremely predictive of unsafe carriers for those circumstances, and therefore very important information when they occur.

Selection Effects: The reason that some trucks are more frequently pulled over for inspections and others are not depends both on state guidance and the inspectors' discretion. As a result, there may be factors that cause some carriers to be inspected more than others, and these factors may or may not be closely related to safety performance.

Transparency of the SMS Algorithm: The panel did not carry out any formal surveys or even informal interviews, but the presentations suggested many carriers find SMS relatively complicated. As a result, they are uncertain whether their score or percentile ranks will increase or decrease based on their most recent months' pattern of violations and crashes. Some degree of

transparency would enhance the reproducibility of SMS measures and give carriers greater trust in the measures and resulting percentiles. This will make it clearer to the carriers what factors could reduce their future chances of having percentile ranks that generate an intervention. In addition, greater access to MCMIS data and the SMS algorithm would facilitate research on SMS and its alternatives by the academic research community.

Making Percentile Ranks Public: Until recently, the SMS measures for the BASICs except for Hazardous Materials and Crashes were made public, as were the percentile ranks. However, as part of the FAST Act, those percentile ranks are not released for property-carrying motor carriers. Doing so would have the benefit of increasing the incentives for carriers to improve their safety performance, since public reporting can result in lost business and in increased insurance rates. On the other hand, it can be argued that the worse SMS does in discriminating between low- and high-risk carriers, the weaker the argument for making the percentile ranks public. If the difference is not substantial, that would provide further justification for not publishing such ranks. FMCSA has examined the future crash risk of the carriers identified for interventions by SMS, and the same for the remaining carriers, and the differences (as discussed in Chapter 2) are substantial.

Given the stakes involved, it is important to examine SMS to see whether these criticisms are valid, and to see if improvements can be made to the input data or to the approach used to discriminate between safe and unsafe motor carriers. FMCSA has been forthcoming in incorporating changes in response to suggestions made by various stakeholders or in providing their reasons for not making changes. Since its implementation more than 6 years ago, SMS has had several revisions to finetune the methodology.

ORGANIZATION OF THE REPORT

The rest of this report is organized as follows. Chapter 2 provides a brief description of the SMS methodology, the evaluations of SMS that have been carried out, the reviews and critiques that have been presented, and our summary assessment of the evaluations and critiques. Chapter 3 describes the statistical modeling issues faced by FMCSA in developing SMS and the opportunity to use Item Response Theory (IRT) to assess safety in the trucking industry. Chapter 3 also describes previous uses of IRT to assess hospital performance and the effectiveness of teachers in elementary schools. Further, it discusses whether SMS percentile ranks should be made public, and the benefits of transparency of SMS or alternatives. Chapter 4 contains the details of the IRT model as it could apply to motor carrier safety, starting with its conceptual basis and continuing through its technical description. Chapter 5 contains some extensions to the model, including multivariate responses, as well as the ability of the model to accommodate exposure measures, provide absolute instead of relative metrics, represent the uncertainty of measures, and identify carriers for interventions. Chapter 6 describes what variables currently on MCMIS need to be improved for use in SMS and in the new model, and then what additional variables, if collected, could potentially improve the performance of the proposed model in the future. The panel's six recommendations are presented in the relevant chapters, specifically Chapters 3, 4, and 6. Appendix A summarizes the agendas of the panel's public meetings, Appendix B details the current SMS algorithm, Appendix C provides examples of simple IRT models applied to the MCMIS data, and Appendix D contains biographical sketches of the panel members.

2
Overview, Evaluations, and Criticisms of the Safety Management System

The primary objective of the Federal Motor Carrier Safety Administration (FMCSA) is to reduce the frequency and severity of commercial motor vehicle (CMV) crashes in the United States. One way of achieving this objective would be to predict, for each carrier, the number of reportable[1] crashes in which its vehicles would be involved in the near future. Such a prediction would take into consideration multiple factors such as the length and nature of the carrier's routes, schedule, cargo, and driver characteristics. Unfortunately, such a program is not feasible, in part, because in any given time period, the incidence of observed crashes is small even though the risk of a crash for a carrier might still be high, and because FMCSA does not have access to motor carrier, vehicle, and driver data that could be used to calculate crash risk. FMCSA has instead adopted a different approach, known as the Safety Measurement System (SMS).

GENERAL OBJECTIVES OF SMS

FMCSA's SMS uses Motor Carrier Management Information System (MCMIS) data to produce metrics to identify, for interventions, carriers that have patterns of noncompliance with FMCSA's safety regulations greater than their peers. FMCSA has data on the frequency that various violations of safety regulations are found during roadside inspections of CMVs. FMCSA uses the identification of violations from these inspections to identify carriers that are systematically out of compliance with federal regulations and therefore are assessed to be at higher risk of future crashes than their peers. FMCSA argues that the presence of such patterns of violations is an indicator of the degree to which a carrier gives priority to safety. Therefore FMCSA's objective becomes a problem of discrimination between those carriers that do and do not act in accordance with regulations for safe operation.

FMCSA intervenes with those carriers that are frequently in violation to encourage them to adopt practices that will result in fewer violations. FMCSA has evaluations that indicate that doing so prevents many future crashes by notifying carriers when they are engaging in unsafe practices and also hopes to minimize the targeting of carriers unlikely to have violations. Until recently, information on which carriers had percentiles above established intervention thresholds were made public, which incentivized carriers to make changes since a carrier could lose business and/or have its insurance rates raised as a result.

There is an important difference between the two objectives of predicting carriers with high future crash risk versus identifying carriers with current high frequency of violations. To predict future crash involvement, FMCSA would not want to find *all* violations, but rather just violations closely linked with future crash risk. However, if the objective is that of prevention, there does not need to be a direct causal link between the frequency of occurrence of many of the

[1]Throughout this report, to be included in the MCMIS, a "reportable crash" has to involve a fatality, a person transported from the scene for immediate medical attention, or a vehicle towed from the scene due to disabling damage suffered in the crash. Also, the vehicle factors for reportability are a vehicle greater than 10,000 lbs., or a bus with seats for 9 or more passengers including the driver.

violations and future crash risk, since their productive use in SMS only depends on the assumptions that carriers that are frequent violators engage in unsafe practices, and carriers that engage in unsafe practices are also likely to have a high frequency of future crashes. So, for example, while a truck or bus often cited for "minor lights out" may not as a direct result be involved in crashes, carriers that are not meticulous about such things may have more crashes due to a general poor approach to vehicle maintenance that impacts crash risk.

It is true that the statistical models would be very similar for models of both prevention and prediction. However, for prediction, more weight would be given to current and previous crash rates, using only violation rates established as strongly predictive of future crash rate. In contrast, for prevention, less weight could be given to current and previous crash rates, with a broader collection of violation rates used that were found to be effective at identifying carriers that have a lack of emphasis on safe practices, not necessarily predictive for crash rate.

The panel understands the reason for the approach taken by FMCSA in trying to prevent future crashes, rather than predict them. It is a common approach taken in other transportation industries. Later in this chapter, we describe FMCSA's evaluations of the degree to which the assumption of the linkage between the frequency of inspection violations and future crash risk obtains.

The data used in SMS are those primarily collected during inspections by Motor Carrier Safety Assistance Program (MCSAP) officials (usually state or local law enforcement) who are certified by the Commercial Vehicle Safety Alliance (CVSA). Depending on the level of inspection, trucks and buses are checked for up to 899 possible violations. The violations recorded during inspections are transferred to the MCMIS database. MCMIS also includes records of crashes involving commercial motor vehicles and records of investigations of motor carriers. In addition, MCMIS contains a census file of all motor carrier companies with self-reported data on the number of power units they have and total mileage traveled, and the results of any investigations. It is to FMCSA's credit that it used this administrative database to produce metrics to monitor well over 500,000 active commercial motor vehicle carriers engaged in interstate commerce or in the intrastate transportation of hazardous materials.

As mentioned, the inputs to MCMIS are crashes, inspections, and violations, mostly reported by the states, investigation information from FMCSA or states, and basic self-reported carrier information. In particular, since crash data are collected by the states and the District of Columbia, there are 51 different crash report forms, 51 different coding manuals, and different training protocols. Crash reports are issued by between 650,000 to 700,000 police officers, ranging from state police to village sheriffs. Therefore, the data are not collected by people trained in CMV data collection nor supervised by data users, but instead by police officers whose main job is enforcing the law and protecting lives and property. FMCSA has access only to these limited data, which were collected for administrative purposes.

Another complication is the diverse world of CMV carriers, including carriers transporting freight cross-country, custom harvesters that transport crops, and buses that take people to church outings, among many others. In addition, carriers come in sizes that differ by many orders of magnitude, from single vehicle owner-operator carriers to carriers that own tens of thousands of vehicles.

Brief History

The program that uses MCMIS data to identify potentially unsafe carriers is referred to as Compliance, Safety, Accountability (CSA), which contains the Safety Measurement System (SMS) as the specific tool to identify carriers for intervention. The CSA/SMS system replaced SafeStat (Federal Register, 2010), which was FMCSA's initial attempt to use MCMIS data to evaluate motor carriers' safety performance. SafeStat, implemented in 1997, comprised four Safety Evaluation Areas (SEAs): Accident, Driver, Vehicle, and Safety Management. These four SEA numbers were combined into an overall assessment, referred to as a SafeStat score. The scores became public in 1999 until 2004, when the Accident SEA was made confidential due to problems with completeness of the data, and because there was no attempt to distinguish between crashes that were and were not preventable from the point of view of the involved motor carrier.

FMCSA's goal of reducing the frequency and severity of CMV crashes implies a set of objective measures that are in line with those Congress mandated in the Moving Ahead for Progress in the 21st Century legislation (MAP-21). In particular, safety in surface transportation was to be measured based on four performance measures: (1) total fatalities, (2) total injuries, (3) fatality rate, and (4) serious injury rate. SMS contributes to responses by FMCSA, the U.S. Department of Transportation, and the states to MAP-21 requirements by focusing regulatory and enforcement attention on factors in motor carrier operations that are related to serious crash risk. Related strongly to this objective is the statement, from FMCSA, that the SMS exists to "change unsafe behavior" (Federal Register, 2010). Not only are carriers with poor safety data subject to intervention (of various types), but also the measures of carriers have, until recently, been made public so that the motor carrier industry and other safety stakeholders would have access to comprehensive and regularly updated safety performance data. In addition, the hope is that by doing this, motor carriers will have an incentive to improve their SMS measures (relative to their peers), and, in the process, safety will improve. Industry stakeholders informed the Government Accountability Office (GAO) that SMS has contributed to a greater awareness of safety and safety performance data by motivating carriers to improve their safety scores to gain an advantage over their competitors (GAO, 2014).

How SMS Operates

SMS produces percentile ranks for each carrier along seven different dimensions: Unsafe Driving, Hours of Service Compliance, Vehicle Maintenance, Controlled Substances/Alcohol, Hazardous Materials Compliance, Driver Fitness, and Crash Indicator (Table 2-1). These seven dimensions are referred to as BASICs (an acronym for Behavior Analysis and Safety Improvement Categories). Besides Crash Indicator—which is an assessment of the current rate of crashes—they are areas of related violations identified in roadside (and other) inspections. Again leaving aside Crash Indicator, the six other BASICs are aggregates of different subsets of 899 separate possible violations that are associated with each BASIC's purpose. Each carrier with sufficient data, therefore, is assigned seven possible measures; carriers with few or no crashes, inspections, or violations are not given SMS measures.

TABLE 2-1 Definition of SMS BASICs

Unsafe Driving: "Operation of commercial motor vehicles (CMVs) in a dangerous or careless manner. *Example violations include: speeding, reckless driving, improper lane change, texting while operating a CMV, not wearing safety belts."* The measure used is: $$\textit{Unsafe Driving Measure} = \frac{\textit{Total of time and severity weighted violations}}{\textit{Average PUs x Utilization Factor}}$$
Hours of service (HOS) Compliance: "Operation of CMVs by drivers who are ill, fatigued, or in noncompliance with the HOS regulations. This BASIC includes violations of regulations pertaining to records of duty status (RODS) as they relate to HOS requirements and the management of CMV driver fatigue. *Example violations include: operating a CMV while ill or fatigued, requiring or permitting a property-carrying CMV driver more than 11 hours, failing to preserve RODS for 6 months/failing to preserve supporting documents."* The measure used is: $$\textit{HOS Compliance} = \frac{\textit{Total of time and severity weighted violations}}{\textit{Total time weight of relevant inspections}}$$
Vehicle Maintenance: "Failure to properly maintain a CMV and prevent shifting loads, spilled or dropped cargo, and overloading of a CMV. *Example violations include: inoperative brakes, lights, and other mechanical defects, improper load securement, failure to make required repairs."* The measure used is: $$\textit{Vehicle Maintenance} = \frac{\textit{Total of time and severity weighted violations}}{\textit{Total time weight of relevant inspections}}$$
Controlled Substances/Alcohol: "Operation of CMVs by drivers who are impaired due to alcohol, illegal drugs, and misuse of prescription or over-the-counter medications. *Example violations include: use of possession of controlled substances or alcohol, failing to implement an alcohol and/or controlled substance testing program."* The measure used is: $$\textit{Controlled Substances} = \frac{\textit{Total of time and severity weighted violations}}{\textit{Total time weight of relevant inspections}}$$
Hazardous Materials (HM) Compliance: "Unsafe handling of HM on a CMV. *Example violations include: failing to mark, label, or placard in accordance with the regulations, not properly securing a package containing HM, leaking containers, failing to conduct a test or inspection on a cargo tank when required by the United States Department of Transportation (U.S. DOT)."* The measure used is: $$\textit{HM Compliance} = \frac{\textit{Total of time and severity weighted violations}}{\textit{Total time weight of relevant inspections}}$$
Driver Fitness: "Operation of CMVs by drivers who are unfit to operate a CMV due to lack of training, experience, or medical qualifications. *Example violations include: failing to have a valid and appropriate commercial driver's license (CDL), being medically unqualified to operate a CMV, failing to maintain driver qualification files.*" The measure used is: $$\textit{Driver Fitness} = \frac{\textit{Total of time and severity weighted violations}}{\textit{Total time weight of relevant inspections}}$$

Crash Indicator: "Historical pattern of crash involvement, including frequency and severity. The BASIC is based on information from state-reported crashes that meet reportable crash standards. All reportable crashes are used regardless of the carrier's or driver's role in the crash. This BASIC uses crash history that is not specifically a behavior but instead the consequence of a behavior or a set of behaviors." The measure used is:

$$Crash\ Indicator\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ crashes}{Average\ PUs\ x\ Utilization\ Factor}$$

SOURCE: FMCSA (2016a).

These seven ratios are referred to as a carrier's SMS measures. For the six noncrash BASICs, the numerators are the sums of the product of two weights for each violation relevant to that BASIC that a carrier has received during the past 2 years. There is a time weight, with more recent violations receiving a higher weight, and a severity weight, with violations viewed as being more critical to safety being given a higher weight. The denominators of these ratios are, with one exception, also weighted sums, but of the number of relevant inspections, where the weights only involve time weighting. The exception is that for the Unsafe Driving BASIC, the denominator is essentially an estimate of vehicle miles traveled, which is arrived at by multiplying the number of power units a carrier has by a utilization factor (see Chapter 6 for more detail). This calculation accounts for the fact that some carriers operate for more miles a year than other carriers, and therefore have more exposure to violations, such as speeding.

The Crash Indicator BASIC is also a ratio. The numerator is a weighted sum of the crashes that a carrier has experienced in the past 2 years, with time and crash severity weights (which give higher weight to crashes with a fatality or injury). The denominator is again a measure of the number of power units a carrier has, with some allowance for more vehicle miles traveled.

Commercial motor vehicle carriers, for both buses and trucks, are characterized as belonging to one of two strata, either straight (vehicles with permanently mounted bodies) or combination (vehicles that pull trailers). The strata were determined by FMCSA by observing how patterns of vehicle miles traveled vary by fleet composition. If a carrier is comprised of greater than 70 percent combination trucks or motorcoaches, it is placed in the combination stratum, with the remainder of carriers placed in the straight stratum. Truck and bus carriers are not separate in SMS. Carriers within these two strata are then placed into peer groups defined roughly by a measure of the size of the carrier based on the number of inspections, which are referred to as safety event groups. Within a safety event group, the measures for each BASIC for all carriers in the group are sorted from low to high. The ranks for each carrier, divided by the number of carriers in the group, are then associated with each carrier and referred to as percentile ranks. For example, if a carrier has the 112th lowest score for a particular BASIC out of 400 carriers in an event group, it is given the percentile rank of 100 x (112/400), or 28 percent, meaning that 72 percent of the carriers had higher measures for that BASIC, or worse safety performance.

For each BASIC, and for each combination and straight segment, FMCSA has derived thresholds for use with the associated percentile ranks that are the same across safety event groups, which comes to a total of 7 x 2 =14 thresholds.[2] If a carrier has a percentile rank above

[2]Given the fact that combination and straight segmentation is only applied to Crash Indicator and Unsafe Driving categories, there are only 9 different thresholds.

the threshold for an individual BASIC, it may receive an intervention ranging from a warning letter, to an investigation, to further monitoring. If its percentile rank is below the threshold, there is no intervention based on the SMS information. FMCSA may still intervene with the carrier due to a crash, complaint, or other nonsafety violation of the agency's regulations.

EVALUATION OF THE EFFECTIVENESS OF SMS BY FMCSA AND VOLPE

A number of organizations have reviewed various aspects of FMCSA's SMS. They include the Volpe National Transportation Center (2014), American Transportation Research Institute (2012, 2014, 2015), Government Accountability Office (2014), Independent Review Team (2014), and Green and Blower (2011). The remainder of this chapter summarizes evaluations of the SMS carried out by FMCSA. We then examine various issues raised by various external agencies and researchers to (1) provide our own synthesis and evaluation of the issues, and (2) assess how the current SMS successfully addresses the concerns or how proposed modifications to SMS could address those concerns. Before proceeding, we provide some thoughts on how one might compare SMS with an alternative in Box 2-1.

Box 2-1 Comparing SMS with an Alternative

SMS is used to identify a specific number of carriers to receive notifications within each safety event group. Suppose there is an alternative to SMS, called SMS*. One way of comparing SMS to SMS* is to examine, across all safety event groups, the number of future crashes for the carriers that receive such notifications under each system with the algorithm having the greatest number of future crashes the winner. There would be some carriers that are identified by both methods and some by neither method, and there would be carriers identified by SMS but not by SMS* and vice versa. Therefore, to decide which was preferred, one could add the number of crashes occurring in some future time period involving carriers in the SMS but not the SMS* group, and compare that to the number of crashes involving carriers in the SMS* but not the SMS group.

This comparison would not be acceptable for the following reason. Very likely, under such an evaluation, the best algorithm would be one that selected the largest carriers for interventions, assuming that they would have the greatest number of future crashes. The problem is that this leaves small carriers safe from scrutiny. Consider a similar problem in a completely different context: which individual tax returns to audit. One wants to focus audit efforts on those returns for which there is the potential for large underpayments, which is obviously those returns where the incomes are highest. However, individuals with smaller incomes should not be immune from an audit, since that would eliminate a large group of people from any scrutiny.

Similarly, FMCSA does not want small carriers to feel immune from scrutiny with SMS. So there is an interest in stratifying the problem so that everyone is subject to scrutiny. An excellent way to do this is to constrain the algorithms being compared to intervene with a set percent of the carriers within each safety event group.

Safety Measurement System (CSMS) Effectiveness Test

Starting with FMCSA / Volpe's evaluations of SMS,[3] the input to compute SMS for this analysis was MCMIS data from 2009 and 2010 (with a few exceptions), and the SMS results were compared to future crash data from January 2011 to June 2012. In this evaluation, FMCSA quantified the effectiveness of SMS for the identification of carriers for interventions.

Crash Rates for Carriers with Alert Status for Various BASICs

The national average for crash rates for the subset of CMV carriers that were active and that had APU and VMT data (which are likely to be larger than average) was 3.43 crashes per 100 power units over a 2-year period. (Concerns about the use of power units as a primary component of a measure of exposure have been raised, which we discuss later in this report.) In comparison, carriers with at least one BASIC in alert status (where "alert status" is identical with being above the FMCSA threshold and therefore issued an intervention for the associated BASIC) had crash rates greater than the national average. Table 2-2 shows that the more BASICs in alert status a carrier has, the greater the carrier's current crash rate.

TABLE 2-2 Crash Rates for Carriers with Different Number of Alerts

Number of Alerts	0	1	2	3-4	5+
Crash Rate	2.69	4.26	5.77	6.24	7.17

SOURCE: Carrier Safety Measurement System (CSMS) Effectiveness Test by Behavior Analysis and Safety Improvement Categories (BASICs) (January 2014).

Further, Table 2-3 shows that separately for each BASIC, the crash rate for carriers with alerts is considerably higher than the national average, with the exception of Driver Fitness, where the crash rate for those carriers with alert status is lower than the national average. Though the measure of total crashes per 100 power units could be improved upon, these data suggest that SMS is identifying an appropriate group of carriers for interventions, and the data support use of six of the seven BASICs.

[3]Carrier Safety Measurement System (CSMS) Effectiveness Tests by Behavior Analysis and Safety Improvement Categories (BASICs), January 2014, Federal Motor Carrier Safety Administration, and CSA Effectiveness Measures. U.S. Department of Transportation, June 30, 2016.

TABLE 2-3 Crash Rates for Carriers with Alert Status in the Seven BASICs

BASIC Identified for Interventions	Number of Carriers Identified	Total Power Units	Total Crashes	Crash Rate (per 100 power units)	% Increase in Crash Rate Compared to National Average (3.43)
Unsafe Driving	9,594	194,756	12,888	6.62	93%
Crash Indicator	4,662	246,463	15,638	6.34	85%
HOS Compliance	22,558	343,114	21,462	6.26	83%
Vehicle Maintenance	15,734	234,895	13,261	5.65	65%
Controlled Substance / Alcohol	2,914	44,945	2,070	4.61	34%
HM Compliance	746	250,892	11,266	4.49	31%
Driver Fitness	5,067	323,038	10,047	3.11	-9%

SOURCE: Carrier Safety Measurement System (CSMS) Effectiveness Test by Behavior Analysis and Safety Improvement Categories (BASICs) (January 2014).

It is important to point out that Tables 2-2 and 2-3 are affected by the fact that one BASIC is the crash rate. Having an alert for that BASIC is equivalent to having a high historical crash rate, though not a future crash rate, which is what is examined here. Also, a concern is that the inference from such analyses (and other findings in these evaluations) is likely affected by the selection effect of which vehicles are chosen for inspection. Even so, this result provides support linking the frequency of inspection violations to crash rate (using 100 power units as the denominator). One improvement that should be considered is to carry this analysis out separately within each safety event group to show that *within safety event groups*, SMS identifies those carriers that have a higher crash rate for interventions than the remainder. By doing that, the aggregate crash rate (with power units as the measure of exposure) for those carriers with interventions would be compared with that for those carriers without interventions within safety event groups. In such an analysis, the comparison would not be the national statistic, it would be the statistic for the remaining carriers in the safety event group.

A surprising result is that carriers with alert status in the Driver Fitness BASIC have lower crash rates than the national average. This result can be better understood if attention is restricted to For-Hire Combination carriers only. FMCSA shows that when doing so, a higher crash rate is obtained (Table 2-4).

TABLE 2-4 Crash Rates for Carriers with Alert Status in the Seven BASICs—For-Hire Combination Carriers Only

BASIC	Unsafe Driving	Crash Indicator	Controlled Substances/Alcohol	HOS Compliance	Driver Fitness	Vehicle Maintenance	HM Compliance
Number of Carriers	6,245	2.826	1,587	17,684	2,329	10,528	276
Crash Rate (crashes per 100 PUs)	8.34	8.02	8.00	7.42	7.21	6.97	5.87

SOURCE: CSA Effectiveness Measures (June 30, 2106).

The crash rate for Driver Fitness for this subgroup of carriers is now 7.21, much larger than the national combination segment average of 5.20 crashes per year per 100 power units. An

argument to condition on For-Hire Combination carriers is that doing so eliminates carriers that carry out many short trips, which are unlikely to result in serious crashes. However, the fact that conditioning can change directions of relationships illustrates a larger point identified by the panel as important. Since there are often several contributory factors for crashes, the value of an individual factor may only be evident in conjunction with other factors as a sort of interactive effect. A full analysis, including all of the important contributory factors simultaneously, is necessary before making decisions on which violations and which BASICs are or are not playing an important role in identifying unsafe carriers. This is a point that we will discuss in more detail in Chapter 5.

Data Sufficiency

FMCSA has established "data sufficiency" standards, which are minimum numbers of crashes, inspections, and violations necessary to support the calculation of SMS measures. (The specific definition of these standards is provided later in this chapter.) To better understand the implications of data sufficiency standards, FMCSA examined the percentage of carriers that were later involved in crashes that had sufficient data to compute BASIC percentiles (Table 2-5).

TABLE 2-5 Percentage of Carriers that Have Sufficient Data to Support SMS

Carrier Group	Number of Carriers	Percent of Carriers	Number of Crashes in Month after Snapshot	Number of Power Units (PUs)	Percent of Power Units (PUs)	Percent of Crashes in Month after Snapshot
All Carriers with Recent Activity	521,952	100.00%	7,890	4,410,068	100%	100.00%
Carriers with Sufficient Data to Be Assessed in SMS	204,651	39.2%	7,217	3,627,065	82.2%	91.5%
Carriers with at Least One BASIC above Threshold	54,674	10.5%	3,645	2,099,101	24.7%	46.2%

SOURCE: CSA Effectiveness Measures (June 30, 2106).
NOTE: Data used from MCMIS, March 2016 snapshot.

We can see from Table 2-5 that about 39 percent of active carriers have sufficient data for SMS; of those, about one-fourth have at least one BASIC above the threshold. However, the number of power units associated with the carriers with sufficient data for SMS measures is 82 percent. Therefore, the coverage of the industry using the current data sufficiency standards is reasonably high, though it would be useful to see what this coverage was for different data sufficiency standards. Further, the carriers satisfying data sufficiency standards represent 92 percent of the crashes the month after the SMS percentile ranks were produced.

Size of Carrier

FMCSA has to apply SMS to a very large and disparate set of carriers, one dimension of which is the substantial difference in the size of those carriers**.** As shown in Table 2-6, crash rates differ substantially by carrier size, with smaller carriers appearing to have a higher crash risk.

TABLE 2-6 Carriers Identified in One or More BASIC Alerts, Size of Carrier, and Crash Rates

Carriers and Power Units (PUs)	# of Carriers Prioritized	% Carriers with at least 1 BASIC Prioritized	Total Power Units	Total Crashes	Crash Rate (per 100 PUs)	% Increase in Crash Rate
5 or fewer Pus	24,647	12%	56,731	4,336	7.64	137% (7.64-4.82) /7.64
5<PUs<15	10,253	24%	92,965	6,173	6.64	149%
15<PUs<50	5,514	30%	145,894	8,693	5.96	117%
50<PUs<500	2,359	35%	308,120	15,110	4.90	84%
More than 500 Pus	269	49%	469,384	17,451	3.72	60%
All Carriers	43,042	15%	1,073,093	51,763	4.82	79%

SOURCE: Carrier Safety Measurement System (CSMS) Effectiveness Test by Behavior Analysis and Safety Improvement Categories (BASICs) (January 2014).

Decrease in Percent of SMS Violations per Inspection, 2009–2016

There is evidence that the percent of inspections with violations has decreased over the time that SMS has been in use. FMCSA argues that this is due, at least in part, to the introduction of SMS (Table 2-7).

TABLE 2-7 Reductions in Violations per Inspection, 2009–2016

Fiscal Year	SMS Violations	Percent Reduction
2009	1.35	
2010	1.35	0.4%
2011	1.26	6.7%
2012	1.18	6.2%
2013	1.17	0.8%
2014	1.18	-0.8%
2015	1.11	5.9%
2016	1.10	0.9%
Percent reduction in violations, 2009-2016		18.5%

SOURCE: CSA Effectiveness Measures (June 30, 2106).
NOTE: Data used from MCMIS, over the period from 2009 through June 2016.

It is important to point out that this evidence is not completely compelling, since during that time period there may have been other confounding factors that were also changing. Having said that, there are no confounding factors that easily come to mind other than a general easing of investigator standards across states.

As shown in Table 2-8, FMCSA also looked at the effectiveness of interventions over time by calculating the crashes prevented by assuming that the observed crash reductions for carriers receiving interventions were due to receiving interventions and not due to other factors (which could include general improvements in safer operations either through management changes, generally safer conditions on the road, or use of technology).

TABLE 2-8 Carriers Receiving Warning Letters and Their Reduction in Crashes[4]

FY	Carriers	Crashes Prevented
2009	9,650	-
2010	13,203	-
2011	40,780	4,232
2012	23,899	3,354

SOURCE: CSA Effectiveness Measures (June 30, 2106).
NOTE: An alternative reason for a decrease in the crash rate for carriers that have alerts for the Crash Indicator BASIC is regression to the mean.

Intervention Effectiveness over Time

Investigations (which are a more intensive form of intervention than a warning letter) seem to have an impact on violation rate, as shown in Table 2-9.

TABLE 2-9 Reduction in Violation Rate Following an Investigation

FY	Number of Carriers	Violation Rate, 1 Year prior to Investigation	Violation Rate, 1 Year after Investigation	Average Violation Rate in the 1 Year after Investigation (2-Year Average)
2009	15,627	1.39	1.27	1.35
2011	14,591	1.41	1.17	1.23
2012	15,925	1.24	1.06	1.17
2013	14,470	1.21	1.09	1.17
2014	11,505	1.24	1.06	1.15

SOURCE: CSA Effectiveness Measures (June 30, 2106).

It is encouraging to see that carriers that have been investigated have endeavored to reduce their violation rate after the year of the inspection, though it is possible that some of this decline is a result of regression to the mean.

DISCUSSION OF CONCERNS RAISED ABOUT SMS

General View of the Approach Taken in SMS

Given the assumption that carriers that violate safety provisions more frequently are also those that have a higher future crash risk, it is reasonable to use the percent of (weighted) inspections that have violations that are associated with a specific type of safety deficiency, as metrics for each carrier. This is, in fact, what the six noncrash BASICs are. The seventh BASIC is weighted crash frequency, which is obviously relevant to future crash frequency. The weights used are severity weights and time weights for the six noncrash BASICs, and time and crash severity weights for the Crash Indicator BASIC.

[4] It should be noted that an alternative reason why there might be a decrease in the crash rate for carriers that have alerts for the BASIC of Crash Indicator is regression to the mean.

We find that carriers are generally supportive of a system that reliably discriminates between safe and unsafe carriers, and that motivates unsafe carriers to improve their safety practices, leaving safe carriers alone. The goal of SMS was not only to reliably identify unsafe carriers, but also to reliably identify safe carriers and thereby not subject those carriers to interventions. This is important since the intention is to make SMS percentile ranks public. While there are no studies that we are aware of that have attempted to measure the size of the economic impact from making SMS ranks public, it is reasonable to expect that the impact would be substantial. Poor percentile ranks presumably would hamper a carrier's ability to attract business. FMCSA interventions typically begin with warning letters, but can progress to investigations, and at the extreme FMCSA can place an unsafe carrier out of business. Therefore, the accuracy and fairness of the inferences based on SMS are of great importance, not only to identify unsafe carriers, but also to ensure the process does not harm safe carriers.

While sensible, this approach does raise some questions, some of which have been discussed in the major critiques of SMS. The issues raised can be viewed as coming from two primary questions: (1) Does SMS do a good job of discriminating between unsafe and safe carriers, and (2) In doing so, is SMS fair to identifiable subgroups of carriers, such as buses, small-sized carriers, or carriers that travel predominantly in given states? To answer those two central questions, we addressed issues such as the following: (1) Does SMS account for state differences in the administration of commercial vehicle inspections; (2) How effective are the data sufficiency standards; (3) Are large and small carriers treated fairly; and (4) Should "nonpreventable" crashes be included in SMS computations? In addressing these and other issues, we at times suggest actions that FMCSA might consider taking to reduce the concerns expressed.

In the following, we review and synthesize the major concerns of FMCSA's SMS raised in other recent reviews, as well as concerns based on our own analysis. These concerns are summarized in Table 2-10.

TABLE 2-10 Summary of Critiques of SMS

Issue	Comments in Literature and by Speakers to Panel
Most, but Not All, BASICs Are Predictive	*ATRI (a) found the Driver Fitness and Controlled Substances BASICs were not very predictive *GAO made the same point *Green and Blower (2011) made the same point
Data Sufficiency Standards	*ATRI (a) argued this is a serious problem that can be partially addressed through the use of wireless roadside (partial) inspection data *ATRI (b) pointed out that many roadside inspections with zero violations were not reported. *GAO argued that it is hard to compare metrics that are so highly variable, providing research that showed that if the data sufficiency standards were raised, SMS percentiles would be better at discriminating between safe and unsafe carriers
Absolute Vs. Relative Measure	*Independent Review Team supported use of an absolute rather than a relative measure for SMS

Issue	Comments in Literature and by Speakers to Panel
Use of Data from Nonpreventable Crashes	*ATRI (c) argued that nonpreventable crashes should not be used for SMS, stating that doing so makes substantial differences and that raters can have high reliability in assessing preventability
Differences in State-Specific Rates of Inspections and Violations	*ATRI (b) showed strong differences in inspection frequency and violation frequency by state *GAO showed similar findings
Stratification of SMS	*The stratification of SMS was raised by several of the speakers during meetings of the panel
Better Measures of Exposure	*ATRI (b) showed substantial state differences in crash rates, which suggests that some states are riskier to drive in, which should be taken into account when forming the denominators for Crash Indicator and possibly for Unsafe Driving
Quality of Existing MCMIS Data	*GAO noted delays in reporting crash data, and the quality of vehicle miles traveled and APU data are low due to misresponse and nonresponse *Independent Review Team was concerned with the degree of incomplete and missing data in MCMIS *Green and Blower (2011) said that crash data in MCMIS were substantially underreported *However, many praised FMCSA for its efforts, including State Safety Data Quality assistance (see Chapter 6 for further discussion)
Appropriateness of Severity Weights	*ATRI (a) tried to validate severity weights through use of a logistic regression model of crash risk as a function of data on violations, crashes, and carrier characteristics. They then examined which violations were most important in modeling crash risk and looked to see if those were the violations that had the highest severity weights. The fit of the resulting model was poor. *GAO, to validate severity weights, also used a logistic regression of crash risk as a function of violations, crashes, and carrier characteristics
Currently Uncollected Variables that Might Substantially Improve SMS	*Independent Review Team said that FMCSA should try to identify carrier variables related to carrier management (see Chapter 6 for further discussion.)

Issue	Comments in Literature and by Speakers to Panel
Sparsity of Some Violations and Utility of Individual Violations; Aggregation of Violations in BASICs	*ATRI (b) showed that certain violations are not very predictive of crash risk *GAO pointed out that 593 of 750 violations occurred for less than 1 percent of carriers, and only 13 of the 750 had a clear association with future crash risk *Independent Review Team said FMCSA should distinguish between violations that are causal for crashes and those that are indicative of management behaviors that may or may not lead to high crash risk
Clean Inspection Reports	*ATRI (a) discovered that a carrier not having a percentile rank due to having sufficient inspections to be scored, but not a sufficient number of violations, had fewer crashes than those without a sufficient number of inspections. Therefore, clean inspections are worth including. *ATRI (b) pointed out that a large percentage of clean inspections are not reported.
Selection Effects	*The panel is concerned that biases on the part of the inspectors in selecting vehicles for inspections might therefore bias SMS
Transparency of SMS Algorithm	*Independent Review Team argued for the importance of greater transparency in SMS (see Chapter 3 for further discussion)
Making Percentile Ranks Public	*Independent Review Team supported the publication of percentile ranks. This was also true for several of the presenters to the panel, though other presenters argued for the percentile ranks to remain private (see Chapter 3 for further discussion)
Comparing Carriers of Different Sizes	*GAO argued that the smallest carriers in each safety event group are most likely to fall into alert status due to the variability of their measures.

NOTE: In the righthand column of the table, ATRI (a): Compliance, Safety, Accountability: Analyzing the Relationship of Scores to Crash Risk, Micah D. Lueck, October 2012; ATRI (b): Evaluating the Impact of Commercial Motor Vehicle Enforcement Disparities on Carrier Safety Performance; July 2014. Amanda Weber and Dan Murray; ATRI (c): Assessing the Impact of Non-Preventable Crashes on CSA Scores, November 2015. Caroline Boris and Dan Murray.

GAO: Federal Motor Carrier Safety: Modifying the Compliance, Safety, Accountability Program Would Improve the Ability to Identify High Risk Carriers; GAO-14-114; Susan Fleming, February 2014.

Independent Review Team: Blueprint for Safety Leadership: Aligning Enforcement and Risk. William R. Voss (chair), Jacqueline Dudley, Neil R. Eisner, Lynne B. Judd, William O. McCabe, and Charles C. B. Raley; 2014.

SOURCE: Green and Blower: P.E. Green and D. Blower, Evaluation of the CSA 2010 Operational Model Test, 2011.

Panel Consideration of Concerns Raised about the Functioning of SMS in Practice

The research carried out by Green and Blower (2011), ATRI (2012, 2014, 2015), and GAO (2014) raised concerns about such issues as the impact of state effects, data sufficiency limitations, and the use of nonpreventable crashes along with preventable crashes. FMCSA has done an excellent job of issuing responses to these concerns, which is important since they need to be addressed in order for SMS to maintain trust as a reliable discriminator between safe and unsafe motor carriers. Here, we provide our own views about these issues, discussing the degree to which the above concerns are justified. As part of this, we will examine whether SMS is fair, by which we mean that even though motor carriers differ in various respects—as a result of the risk environment that they operate in, the nature and size of their business, and the areas in which they operate—SMS should, to the extent feasible, compare carriers in a way that takes such differences into account. We now proceed to discuss issues that the panel itself or other observers have raised about the performance of SMS, as summarized above in Table 2-10.

Most, but Not All, BASICs Are Predictive

Most of the BASIC percentile ranks have been found to correlate strongly to future crash risk. However, one or two BASICs have been shown to have weak or negative correlations. Specifically, ATRI (2012), GAO (2014), and Green and Blower (2011) have shown some of the BASICs, especially Crash Indicator and Unsafe Driving, have very strong correlations with future crash risk, and three more BASICs have moderately strong correlations. However, Driver Fitness has been shown to have a negative correlation with future crash frequency. Given that, some critics have suggested that Driver Fitness be considered for removal or refinement.

ATRI (2012) studied a subset of 471,306 motor carriers from a sample of 772,281 registered interstate and intrastate hazardous material carriers that had evidence of recent activity in the 24 months from April 12, 2010, to April 11, 2012. They focused on the five BASICs then available to the public: Unsafe Driving, HOS Compliance, Vehicle Maintenance, Driver Fitness, and Controlled Substances/Alcohol. To determine whether percentile ranks were related to crash frequency, ATRI (2012) fit a log-linear negative binomial regression model with dependent variable crash frequency and predictors percentile ranks for each BASIC. It should be noted that the crash data was contemporaneous, so ATRI (2012) was not evaluating SMS in a predictive environment. Their results, provided in Table 2-11, showed a strong correlation between crash frequency and SMS percentile ranks for the Unsafe Driving, Hours of Service, and Vehicle Maintenance BASICs, respectively. However, ATRI (2012) found a negative relationship for both Driver Fitness and for Controlled Substances / Alcohol BASICs. That is, in those two cases, higher (worse) percentile ranks were associated with lower crash frequencies. ATRI (2012) raised the concern that this was due to inclusion of violations that were not associated with safety deficiencies that contributed to crashes.

TABLE 2-11 Log-Linear Negative Binomial Regressions Models for BASICs

BASIC	Parameter Estimate (x 100)	Std. Error (x 100)
Unsafe Driving	1.1	0.02
Vehicle Maintenance	0.8	0.01
HOS Compliance	0.8	0.02
Driver Fitness	-0.9	0.06
Controlled Substances/Alcohol	-1.0	0.08

SOURCE: ATRI (2012), Adapted from Tables 4-8.

Further, Green and Blower (2011) looked at scatterplots of Crash Rates by BASIC percentile ranks. These plots showed strong positive relationships for Unsafe Driving, Fatigued Driving, Vehicle Maintenance, and Controlled Substance / Alcohol. However, the plots for Driver Fitness and for Loading/Cargo of the association between crash rate and percentiles showed a negative association. (Note that SMS has defined slightly different BASICs over time.)

A related question is whether those receiving alerts from SMS have higher crash frequencies than those not receiving alerts. To answer this question, ATRI (2012) classified carriers into one of two groups depending on whether they received an alert for each BASIC. They computed the average crash rates for each group (alerts and nonalerts) and took the ratio. Values greater than 1.0 correspond to carriers that received an alert having a higher average crash risk, and vice versa. The results are given in Table 2-12.

TABLE 2-12 Relative Crash Risk for Carriers Above vs. Below Alert Threshold in Each BASIC

BASIC	Crash Risk
Unsafe Driving	1.74
Vehicle Maintenance	1.42
HOS Compliance	1.34
Controlled Substance/Alcohol	1.32
Driver Fitness	0.87

SOURCE: ATRI (2012), Adapted from Table ES-1.

Table 2-12 suggests that, with the exception of Driver Fitness, intervening with carriers with alerts is sensible. ATRI (2012) posited that one problem with Driver Fitness might be that the associated severity weights might not all be assigned such that violations that are more associated with future crash frequency are given higher weights.

ATRI (2012) also developed negative binomial regression models to compare carriers with alerts against carriers without alerts. The coefficients for the indicator variable for alert status are given in Table 2-13. This analysis includes standard errors with the regression coefficients, which helped the panel assess the magnitude of the effect. Again we see that Driver Fitness is the only BASIC where the indicator variable for alert status is inversely related to crash risk.

TABLE 2-13 Log-linear Negative Binomial Regressions Comparing Carriers with Alerts Against Carriers Without Alerts

BASIC	Parameter Estimate	Standard Error
Unsafe Driving	.554	.012
Vehicle Maintenance	.348	.009
HOS Compliance	.293	.010
Controlled Substances/Alcohol	.276	.03
Driver Fitness	-.140	.023

SOURCE: ATRI (2012), Adapted from tables 10, 12, 14, 16, and 18.

The study carried out by GAO (2014) used data from December 2007 through June 2011, the first 2 years to fit models and the last 18 months for purposes of evaluation. GAO (2014) pointed out that for Driver Fitness and for Controlled Substances/Alcohol, the association with crash risk is not very strongly positive.

Green and Blower (2011) evaluated the effectiveness of SMS by comparing the crash rates for the carriers with BASIC percentile ranks that exceeded the SMS thresholds to the carriers with BASICs not exceeding the SMS thresholds. The results, given in Table 2-14, showed that the carriers SMS selected for interventions had higher crash rates than those that SMS did not select, though for Driver Fitness and Improper Loading the evidence was weaker.

TABLE 2-14 Crash Rates for Carriers Identified by SMS Compared to Those Not Identified

BASIC Threshold Exceeded	Carriers	Crashes	Power Units	Crash Rate per 100 PU	Ratio to Not Identified
Unsafe Driving	9,245	33,532	450,874	7.44	3.56
Fatigued Driving	17,959	15,525	248,862	6.24	2.99
Driver Fitness	3,981	11,539	379,009	3.04	1.46
Controlled Substance and Alcohol	1,013	6,860	104,799	6.55	3.14
Vehicle Maintenance	18,280	13,643	278,198	4.90	2.34
Improper Loading/Cargo Securement	9,409	16,747	421,670	3.97	1.90
Crash Indicator	5,077	33,946	463,766	7.32	3.51

SOURCE: Green and Blower (2011).

We on the panel believe that while the evidence against the Driver Fitness BASIC is worrying, eliminating it based on the currently available information would be premature. As an example, FMCSA has shown that when focusing on for-hire, combination truck carriers, Driver Fitness became a much better predictor. Further, ATRI has shown that by using the number of total alerts as a metric, there was a monotonic, positive relationship with crash risk. This suggests that all of the BASICs have unique contributions to assessments of safe operations. As CMV safety is multidimensional (many factors contribute to making a carrier safe), a BASIC percentile rank in one category that is not strongly correlated with future crash risk may still be predictive when combined with other BASIC percentile ranks. For example, regressing crash rate on BASIC

percentile ranks with interaction terms can reveal underlying relationships between the BASICs. We are not necessarily arguing for retention of Driver Fitness in its current form. However, the consensus of our panel is that the evaluations carried out by FMCSA supports the judgement that six of the seven BASICs are positively (sometimes very strongly) associated with future crash frequency, and that the unconditional correlation of Driver Fitness's percentile ranks with future crash frequency is insufficient to remove it from SMS. We describe a new approach to SMS in Chapter 4 that is more natural to the problem and the dataset used and can naturally address the question of modification of BASICs to enhance their predictive strength.

Data Sufficiency Standards

Data sufficiency standards must trade off the reliability of SMS measures and percentile ranks with the percentage coverage of the carrier population that is given SMS measures. As data sufficiency standards are relaxed, resulting in less reliable SMS measures, it is possible to provide SMS percentile ranks for a larger fraction of the active CMV carriers. (Appendix B provides the data sufficiency standards for each BASIC.) Also, since most carriers operate at most a few vehicles that are therefore inspected infrequently, it is difficult to compare small carriers against each other because their measures are so variable. GAO has carried out research that demonstrates that if FMCSA raised its data sufficiency standards, SMS would better discriminate between carriers that have lower and higher future crash risk, though GAO acknowledges that would result in a small number of carriers for which SMS can provide percentiles.

There is no getting around the point that providing BASIC measures to carriers that have very infrequent inspections will result in highly variable assessments of such carriers. This is simply because not much is known about the frequency of violations for small carriers. Such high variance measures can result in mischaracterizing the nature of a carrier—the high variability could result in the carrier being given alerts more or less often than what would be warranted given its behavior. On the other hand, the industry is highly skewed, being comprised of a very large number of small carriers. If the data sufficiency standards were raised, a high percentage of the industry would be excluded from measurement by SMS and therefore monitoring by FMCSA. We believe that this issue should be further investigated. Our preferred model, described in Chapter 4, will have some ability to reduce the variance of these measures through use of smoothing with the measures of a carrier's peers. Ultimately, this is a policy decision for FMCSA to make, but one that can be informed by additional research.

Use of Absolute versus Relative Measure

The Independent Review Team (2014; p. 9) recommended that FMCSA: "Continue to identify and implement methods for emphasizing absolute rather than relative individual motor carrier rankings so that it does not undermine industry's willingness to innovate and share best practices." The team based its recommendation on the following conclusion:

> The relative SMS percentile ranks motor carriers based on their SMS scores relative to their peers. In this system, it is possible for a motor carrier's rating to rise or fall based on the actions of its peer carriers and may be unrelated to any action by the rated carrier. For the investigators, the relative nature of the BASIC scores makes it difficult for them to discern if changes in percentile ranks are occurring because of: (a) aging of violations, (b) changes in the peer group's performance with no change in operator performance, (c)

> real changes in a carrier's operating performance. For the motor carrier, the Independent Review Team found that the relative scoring actually can discourage the sharing of leading safety practices because any increase in the score of a peer may result in a reduction in the relative rating of the motor carrier that shares it. It is possible the competitor subsequently achieves a better percentile score while the first carrier's own relative rating decreases without any actual change in safety performance.

As the Independent Review Team pointed out, the use of percentile ranks, rather than the SMS measures, is a relative metric, which has the following disadvantage. It is possible for a carrier to lower its SMS score from one time period to the next and still have its percentile rank increase as a result of larger improvements on the part of the remaining carriers in its safety event group. On the other hand, using an absolute standard of performance, as the entire industry gets progressively safer, the standard will at some point become irrelevant. Having a relative metric enables FMCSA to keep pressing for better performance. Also, a relative metric is natural since CSA/SMS operates on a fixed budget. The program can only support a fixed number of interventions of various types, which is consistent with looking for the worst percentiles of carriers for interventions. Since there are advantages to both relative and absolute measures, we believe that FMCSA should strongly consider use of a two-dimensional metric that takes into consideration both the SMS score and the percentile rank, using some objective formula, to decide on which carriers will receive interventions. Further, given that a safety event group could be a subset of the active carriers that are very safe performers, there might be an advantage in seeing how a carrier ranks over all active carriers. Lastly, given that the only reason for safety event groups is to compare measures with similar variances, it might be beneficial to see how a carrier's measures compare to the entire population of carriers with SMS measures. This is because, while a relatively small safety event group could presumably have widespread improvement, it would be more difficult for this improvement to occur across the industry.

Recommendation: Given that there are good reasons for both an absolute and a relative metric on safety performance, FMCSA should decide on the carriers that receive SMS alerts using both the SMS percentile ranks and the SMS measures, and the percentile ranks should be computed both conditionally within safety event groups and over all motor carriers.

Use of Data from Nonpreventable Crashes

There are crashes considered not preventable by many in the CMV community. Examples include colliding with an animal in the roadway, being hit while legally parked, being struck by another driver who ran a red light or a stop sign, being hit by another driver who was under the influence of drugs or alcohol, or a truck-assisted suicide by a pedestrian or driver. ATRI (2015) showed large changes to the Crash Indicator BASIC when removing the contribution to MCMIS from such crashes. The study that ATRI carried out focuses some of its analyses on the above five situations, and is therefore conservative since there are, of course, many other types of nonpreventable crashes. The suggestion is that SMS will be more effective at identifying unsafe motor carriers if such crashes are removed from calculation of the Crash BASIC.

Put another way, the suggestion is that nonpreventable crashes should not be included in SMS because any carrier placed in that same circumstance would have been involved in a crash,

and so including them in the Crash BASIC does not help in discriminating between safe and unsafe carriers. This is an important issue, especially for small carriers, since such events can be extremely damaging, possibly putting some small carriers out of business.

However, some considerations complicate the proposal that such crashes be set aside. First, a large percentage of such crashes might have been prevented by drivers who took a more defensive approach to operating their vehicles. For example, they might have given themselves a larger distance from a swerving driver, deaccelerated slower when approaching a stoplight or a crash scene, parked in a more well-lit location, and so on. This is supported by the high correlation between the Crash Indicator BASIC and future crash risk, making it likely that there is some predictive value from data on most crashes, not just the preventable ones. It might also be the case that clearly nonpreventable crashes will make up such a small percentage of overall crashes that removing them would make little difference in the utility of the Crash Indicator BASIC.

Second, it would be difficult to create an algorithm that would take as input the evidence at the scene of a crash and determine which crashes were and were not preventable. If an algorithm could not be created, a subjective element of the determination would have to be part of the decision rule at the state level. Also, additional data (beyond the FMCSA required data elements) are recorded in different states for crashes that meet the reporting criteria based on the contents of the state's own standard crash report form or crash reporting software. (National guidelines—the Model Minimum Uniform Crash Criteria [MMUCC] and the American National Standards Institute ANSI D16.1 standard—are not mandatory but provide guidance that states may choose to use in designing their own data requirements for crash reporting.) Even in the most obvious cases (a single-vehicle crash), the causal attribution may not be simple enough that it could be assigned using a software algorithm. Expert investigation, including postcrash investigation, is required in order to be reasonably certain that all of the causal contributing factors are accounted for, and fault is apportioned as accurately as possible. This process is difficult to manage even in a single state. Doing so across all states, with multiple datasets, would require extraordinary, sustained, and costly efforts. In addition, FMCSA has no authority to ensure uniform application of such a guideline. Further, the lack of a uniform dataset standard adopted by all states means that expanding the data available to examine other vehicles and drivers involved in the crash, as well as to potentially assess the crash circumstances to reliably apportion fault, is not feasible or practical at this time.

Third, there is the question of the reliability of such evaluators. In a 2012 FMCSA report (Craft, 2012), researchers using police accident report data coded 1,221 crash records across five severity categories, with 93.2 percent agreement with assessments provided by researchers using data from the large truck crash causation study (Blower and Campbell, 2002), which was considered to be a reasonable surrogate for truth. This study is certainly encouraging for the position that such assessments would be reliable. However, ATRI (2015), based on findings in FMCSA (2012), stated that, "The reliability of PARS [police accident reports] was tested by comparing them to FARS [Fatal Analysis Reporting System] records. There were significant inconsistencies between PAR and FARS data for areas critical to determining culpability; 82 percent of the PARS were missing driver contributory factors and 47.5 percent of the PARS were missing the first harmful event." So the reliability remains unclear.

FMCSA has initiated a research project to look into the costs and benefits of setting aside nonpreventable crashes. We believe that such research is of interest. While we are skeptical about setting such data aside, there might very well be schemes in which downweight crashes

judged to be nonpreventable—even if the method for arriving at such a determination is error-prone—could result in an SMS percentile rank for Crash Indicator that is preferable to the current version. One way of doing this would be to downweight the vehicle struck in a collision relative to the striking vehicle. However, we do not believe that additional research along the same lines should be given a high priority since we do not believe that such an appreciable change will make a large difference in percentile ranks. There is the separate question of whether to use such a metric as the dependent variable for objective functions evaluating the other BASICs, and this is also worthy of study.[5]

State Differences in Rate of Inspections and Violations

ATRI (2014) and GAO (2014) showed that there are strong differences by state with respect to the frequency of inspections and the frequency of particular violations. With respect to differences in the frequency of inspections, ATRI (2014) found that: "In 2011, on average, CMV enforcement personnel conducted 12.2 RIs [roadside inspections] per MVMT [million vehicle miles traveled] and issued 22.8 violations per MVMT. … Maryland had the highest inspection rate with 27.9 RIs per MVMT, which was 128.7 percent greater than the national average. In comparison, Oklahoma conducted the fewest RIs with 3.7 per MVMT, which was 69.7 percent less than the national average."

Second, with respect to differences in the frequency of violations, ATRI (2014) found that: "among all driver violations reported in 2010, the share of violations for speeding varied significantly from state to state, representing 31.7 percent of all driver violations in Indiana, 16.9 percent in Ohio and 4.2 percent in Arizona." Further, "While the national average was 11.97 light violations for every speeding violation, the ratio varied from a low of 1.91 in Indiana to 321.02 in Texas." Also, the "national average for 'windshield wipers inoperative or defective' … violations per 100 relevant RIs was 2.0… Texas issued 12.2 windshield violations per 100 relevant RIs, which was 510.0 percent greater than the national average and ranked first nationally. In comparison, North Dakota issued 0.19 windshield violations per 100 relevant RIs, which was 90.5 percent lower than the national average…" Further, ATRI showed that if these state differences were eliminated, SMS percentile ranks would change appreciably.

ATRI (2014) makes it clear that this is not something that FMCSA can unilaterally change.

> "While FMCSA sets guidelines on the adoption and enforcement of Federal Motor Carrier Safety Regulations (FMCSRs), each state enforcement agency has the discretion to emphasize specific enforcement foci and activities in order to accomplish FMCSA's overall safety goals, with this privilege extending even further to local jurisdictions. For example, FCMSA acknowledges that different enforcement jurisdictions may utilize differing methods to select or screen a commercial motor vehicle (CMV) for inspection. Likewise, it is the decision of the enforcement officer to issue a citation, violation, or both during a roadside inspection (RI). Finally states have the discretion to vary their enforcement foci, for instance, taking a close look at driver issues as opposed to vehicle

[5] Some have proposed that carriers' internal data be used to assess preventability of crashes through appeals to FMCSA. This idea has several major disadvantages: 1) carriers crash records are subject to great variability in quality and detail; 2) FMCSA would have almost no influence on the quality and completeness of such data; 3) it is unlikely that a sufficient number of carriers will have appropriate data; and 4) assembling and managing the data would be extremely complicated and costly. Further, some carriers may be incentivized to bias their classification.

defects, or focusing more attention on certain failures (e.g., brakes) versus behaviors (e.g., speeding).

The report also provides statistics that support the assertion that the driving environment is more challenging in some states than others. The reports states that "… the national average for large truck crashes per MVMT in 2011 was 0.26. Wyoming had 0.52 large truck crashes per MVMT, which was twice the national average and ranked first nationally. Conversely, New Mexico had the lowest rate with 0.08 large truck crashes per MVMT, which was 69.2 percent less than the U.S. average."

To summarize, there are state effects in how frequent crashes are from state to state. Also, the frequency that inspections take place and various violations are cited depends on the state in which a truck or bus is traveling. Let us first discuss crash frequency, which is essentially one of the seven BASICs. There are many known causal factors for crashes, including two-lane highways as opposed to interstate highways, and frequency of ice and snow on the roads, congestion, visibility, etc. As a result, carriers that travel more often in states with a greater frequency of these and other causal factors will likely have higher crash frequencies than other carriers, assuming that their propensity for crashes is otherwise identical. Therefore, the SMS measures for carriers that travel more often in those states will likely be higher as a result. To eliminate this bias, FMCSA would have to develop a model for exposure that took into consideration not only vehicle miles traveled but also the relative risk of the environment traveled through. Unfortunately, the information needed for input for such a model, which would include the time and location for all trips a carrier makes, is not available.

The situation is similar for violations. It is reasonable to believe that there are violations that are obvious and there are violations that are more borderline. For instance, depth of tire tread can be slightly past what the regulations require, and as a result whether a violation is issued can depend on whether the inspector is focused on such violations, and this might depend on the environments encountered when traveling on that state's roads. The current exposure measure for most violations (besides Unsafe Driving) is the number of inspections. One could argue, analogously to crash frequency, that carriers that travel more often in states that issue borderline violations with greater frequency will have a greater propensity for violations, independent of their safety performance. Again, one could develop a model that could correct for this difference in propensity for issuing violations by developing an exposure measure. However, here again, the inputs necessary to develop such a model, including the time and location for all trips, does not exist. Given that, there will be a bias for carriers whose CMVs travel to states with greater propensity for issuing violations. If such data were available in the future, the new approach to SMS described in Chapter 4 could accommodate the new information to better understand state effects.

Stratification of Types of Carriers

In order to provide for a fair comparison of carriers, SMS would benefit from stratification that formed peer groups of carriers that were undertaking trips of similar risks. This is somewhat taken care of by standardizing by vehicle miles traveled for the Crash Indicator BASIC, but, as discussed above, trips through some states and some roads at some times of the year and at some times of the day are riskier than others. More importantly, some truck and bus tasks are much riskier than others, such as transporting logs. This is related to the discussions of measures of exposure for crashes below, and in the above discussion of state differences in

inspections and violations. A similar argument can be made for the noncrash BASICs, which are normalized by the weighted number of inspections.

Currently, SMS makes use of two factors to stratify carriers: (1) carriers that have a certain percentage of combined versus straight trucks (and a similar stratification for buses), and (2) safety event groups, which is essentially stratification by size. Combination trucks tend to have substantially higher annual vehicle miles traveled than straight trucks. Since trucks with more vehicle miles traveled are more exposed to crashes, it makes sense to treat carriers that operate primarily combination trucks separately from those that operate primarily straight. However, that argument is somewhat offset by the use of denominators that are similar to vehicle miles traveled. A better argument points to studies that support the view that combination trucks are riskier to drive than straight trucks (National Highway Traffic Safety Administration, 2013: Table 48).

There are currently only a few additional variables that FMCSA could stratify on if desired. The primary ones potentially of use are business/operation data, type of cargo carried, and hazardous materials shipped. The key suggestion that is heard is to stratify SMS by truck or bus carriers, then by type of business within the truck and bus strata. For instance, one might wish to divide buses up into school buses and other local transportation, and other, or divide trucks up into interstate and local carriers. It might also be interesting to consider stratification of hazmat vehicles and non-hazmat vehicles, and to consider more delineations in the separation of straight and combined carriers. The general advantage of stratification in SMS would be to form peer groups where the risk of travel is more comparable for the members of the stratum or peer group. Otherwise, crashes might be occurring simply as a result of having a riskier set of trips. Unfortunately, the quality of the responses to type of business is suspect since the carriers themselves identify their types of business and types of cargo they carry. These responses are often fairly wide ranging, possibly to support any possible business opportunity.

In addition to the quality of the characteristic on which to stratify, two considerations need to be traded off in deciding whether additional stratification is desirable. First, there is no reason to employ additional stratification if the resulting cells are not clearly more homogeneous with respect to risk. Therefore, empirical work needs to be carried out to determine whether the better carriers (for example, eliminating those outside of the thresholds) with those differing characteristics have clearly different frequencies of crashes. (It is useful to look at the better carriers since those carriers that need to improve their operations would be excluded.) The result would be better discrimination between safe and unsafe carriers. On the other hand, further stratification results in having fewer peers, and the fewer peers that a carrier has, the more difficult it is to be certain that a carrier's performance is atypical by being among the highest by a certain percent.

Trading off these two considerations is difficult, assuming the desire to retain the same number of safety event groups throughout the stratification. A collection of carriers with very high-risk businesses should be given their own cell even if very few carriers have the characteristics. Sometimes, such situations are handled well by statistical models of the risks involved. Absent use of such an approach, FMCSA needs to examine whether some additional stratification results in a collection of carriers for intervention that is preferable to the current stratification.

Better Measures of Exposure

The goal of SMS should be to sum the risks of crashing for each trip that all the CMVs make for a carrier; to compare the expected number of crashes, assuming reasonable efforts to operate safely, with the observed number; and to intervene with those carriers that have a large multiplicative or additive positive difference between observed and expected crashes. Unfortunately, as noted several times in this report, this objective is not attainable given the current data that are available.

A measure of exposure sums the risks of all trips and creates a way of standardizing that allows numbers of crashes (or numbers of violations) for carriers to be comparable. A first attempt would be to divide the number of crashes by total vehicle miles traveled (VMT) by a carrier to arrive at crashes per miles traveled (or violations per mile traveled) as comparable statistics. The problem with using total vehicle miles traveled is that this quantity in MCMIS is self-reported with substantial nonresponse and likely substantial misresponse. The reported VMT in MCMIS is a poor quality measure of exposure. FMCSA instead uses the average number of power units (APU), averaging over three possible responses in a 2-year period, for a carrier, multiplied by a utilization factor, which accommodates carriers that travel more and less than others, but truncates very low and very high values for reported vehicle miles traveled. However, APU is itself of uncertain quality, and likely goes out of date for some carriers fairly quickly due to growth or diminishment of business, mergers, or other factors. APU also has some nonresponse, though much less than for VMT. Given that this factor can clearly be off by a substantial amount and has a direct impact on the Crash Indicator BASIC and the Unsafe Driving BASIC, it is vitally important for FMCSA and CMV associations to work collaboratively on an improvement. SMS cannot be any better than the data that are input into it.

Further, as ATRI (2014) showed, there are substantial state differences in crash rates, which suggests that some states are riskier to drive in than others. This could be due to quality of roads, congestion, terrain, placement of rest stops, types of weather, and other factors that differ by state. Trucks used for long-distance transport tend to operate on the safest roads, such as interstate-quality highways. In comparison, logging operations may use unpaved roads, with variable and uncertain loading, while delivery operations to highly urbanized areas encounter substantial congestion. Similar safety-relevant operational distinctions between different types of bus operators could be made.

If possible, the most important of these factors, in addition to total VMT, should be taken into account when comparing the number of crashes for the Crash Indicator BASIC. With the required information, such factors could be accounted for in a number of ways. For instance, a statistical model of crash risk as a function of these factors could be developed and, given the model, either weight miles traveled given the estimated risk of each additional mile, or stratify by the total exposure risk. If the input data on vehicle miles traveled could be improved, the method for exposure that FMCSA currently uses would be satisfactory until the above information is available. In addition, such improvements will not only improve the Crash Indicator BASIC as an indicator of which carriers receive interventions, but also as an improved measure of future crash risk.

Appropriateness of Severity Weights

ATRI (2012) argued “certain violations suggested to have a stronger relationship to safety events may not be the best predictors of crash risk and that CMV enforcement strategies may need to shift focus to address other violations that may have stronger relationships to crash

risk." ATRI developed a list of 10 violations that it referred to as Crash Predictor (CP) violations. These were the violations most strongly associated with future crash risk. ATRI then examined the state differences in issuing these violations, finding considerable differences. States with high frequencies of these 10 violations had higher crash rates than states with high frequencies of what FMCSA calls red flag violations that it viewed as the most indicative of a lack of safety behavior.

Further, the Independent Review Team (2014) included the following recommendation relevant to SMS: "Recommendation 2.3.1: FMCSA should expand its work with industry and stakeholders to develop SMS enhancements. These enhancements should enable FMCSA to better discern motor carrier management actions that lead to crashes and to allow more timely and appropriate investigation and enforcement actions."

ATRI (2012) and GAO (2014) tried to validate severity weights by modeling crash frequency as a function of violations, previous crashes, and carrier characteristics. The hope was to find that the violations that were most predictive were generally the violations that had the highest severity weights. Unfortunately, the fit of the resulting models was poor, which is not that surprising. As mentioned before, crashes have many causes, some of which are very particular to the situation, and some of which a carrier has no control over. These factors add noise to such a model and, as a result, the lack of fit of crash frequency will likely be substantial. We believe that this information is of interest but insufficient to change severity weights at this time. The severity weights were derived starting with subject-matter expertise refined using empirical methods (Volpe, 2010). This research relied on sub-BASIC grouping, which again utilized subject-matter expertise. We do not believe that research into the relationship between severity weights and future crash risk should be a high priority for FMCSA since the algorithm is not extremely dependent on such weights. Also, such research is similar to the models that were argued to be too difficult to development earlier in this report. (For further discussion, see the sections below on predictive strength and sparsity of violations, and on violations and severity weights in Chapter 6.) In addition, a feature of the approach described in Chapter 4 has a natural way of adjusting such weights over time.

Predictive Strength and Sparsity of Violations: Aggregation of Violations into BASICS

Two questions raised by the data on violations are: (1) whether all 899 violations are useful in discriminating between safe and unsafe carriers, and (2) whether aggregating them into the current six noncrash BASICs (the BASIC on Crash Indicator is an obvious measure) makes sense, or whether different or more or fewer groups be considered. ATRI (2014) demonstrated that most violations are not individually very predictive of crash frequency. The GAO (2014) and the Independent Review Team (2014) reports also raised questions about the importance of all of the violations, with GAO determining that 593 out of the then 750 violations occurred for less than 1 percent of carriers, and only 13 of the 750 violations had a clear bivariate association with future crash risk.

FMCSA is well aware that individual violations have modest associations with future crash risk. That is the primary reason for aggregation into BASICs. Further, as pointed out by safety experts currently and formerly from Schneider National in presentations to the panel, some of the violations used in the BASICS are not clearly related to operational safety. For example, the paperwork accompanying a hazardous materials shipment can result in a violation when that paperwork is primarily the shipper's responsibility. Also, whether a hazmat placard mounted at the correct angle contributes to safety can certainly be questioned. There are also

violations for the various detail lights on the frame of a truck being out, and it is unclear how relevant such violations are to safety assessment.

It is important to point out violations could appear to be uncorrelated with crash risk but be extremely predictive in particular situations. An example might be a CMV with tires with soft tread, which might not be generally predictive but might be very predictive when roads are icy. Confounding factors that affect the relationship between a violation and crash risk are referred to as moderators, factors that have important interaction effects with individual violations.

Also, the assumption on which SMS is based is not that violations are predictive of crash frequency, but rather that violations are indicative of carriers with poor safety operations. In addition, violations that only rarely occur could be very predictive of crashes conditional on less common circumstances. Therefore, the question of whether individual violations should be discarded is quite complex, It is difficult to lay out a decision rule for their inclusion or exclusion.

To add to this, Collin Mooney, executive director of the Commercial Vehicle Safety Alliance, provided examples to the panel whereby almost identical situations could result in different violations with substantially different severity weights, which would have substantially different impacts on SMS measures (see Chapter 6 for examples). Related to this, different software packages help road inspectors translate their observations into MCMIS violations, and these packages are not required to satisfy any national standards. There are currently efforts to mandate that these software packages be standardized to some extent to map the observations of inspectors to violations in the same way. Clearly, the fact that identical situations can result in very different violation codes with different severity weights and the lack of any national standardization of the coding that is done adds additional variability to the percentile ranks, which makes it harder to identify the carriers operating less safely than others.

Based on these considerations, it is undeniable that some of the current collection of violations, especially those that are not the responsibility of the driver, should not play a role in SMS. It is plausible that they may still be useful in identifying carriers that are giving insufficient attention to safety, but it seems more likely that they are not playing a beneficial role in SMS and should be dropped. In support of this, FMCSA should consider examining the current group of 899 violations and remove any that clearly are not indicative of a carrier operation that prioritizes safe operations. Data quality would benefit, since the collection of detailed information on all possible violations is a costly exercise; once it is determined which set of violations are indicative of safe operations, FMCSA can concentrate on collecting quality information on that subset. In addition, FMCSA should consider setting up the violations in a manner that reduces the possibility of alternate scoring of identical circumstances, which might include setting standards for how the software tools function, but which also may be related to redundancies in the current way the violations are formatted. Finally, the new approach to CSA/SMS we describe in Chapter 4 has an empirical method for analyzing the utility of individual violations.

The second issue raised above concerns the bundling of the violations into groups, which has up to now relied on a strong subject-matter understanding of which violations are indicative of related behaviors by carriers. For instance, it makes sense to have a separate Vehicle Maintenance BASIC because there is a prima facie case that the mechanical condition of trucks affects crash risk (though this does not defend all of the violations that make up the Vehicle Maintenance BASIC). Having said this, a large number of violations could be considered for deletion or for movement to other BASICs, and there is the question of whether to form new BASICs out of the components of existing ones. FMCSA is clearly open to the modification of

the grouping of violations into BASICs, as regular changes have been made since the inception of SMS in 2010. An advantage of modestly increasing the number of BASICs is that it could provide more targeted information as to what safety issues a carrier needs to address. Also, it would be beneficial if the BASICs identified relatively separate sets of carriers for interventions, as otherwise the use of different BASICs offers little benefits. The panel analyzed this issue by forming 15 two-by-two tables of carriers, with columns defined by two different noncrash BASICs ($15 = \binom{6}{2}$), and the rows defined by not exceeding or exceeding the alert threshold. This analysis is flawed because the stratification by safety event groups complicates things, but it still provides some idea of the redundancy of BASICs. One measure of agreement is the percentage of those with alert status for a given BASIC also getting alert status for the BASIC that is being compared. This percentage is as high as 55 percent for Driver Fitness compared to Vehicle Maintenance, and 34 percent for Hours of Service and Unsafe Driving. These values do not support worries about the redundancy of the BASICs, though this analysis is very preliminary.

Succeeding in setting up somewhat different BASICs better at defining separate areas in which a carrier may need an intervention is a challenge. This is because the great majority of violations are rarely cited, and crashes are also rare, making such empirical research difficult. (The related question as to whether entire BASICs should be retained is addressed elsewhere in this report.) The new modeling approach described in Chapter 4 provides a natural way of examining the question of which violations are grouped into BASICs and whether the number of BASICs should be changed. In particular, the later discussion of a multidimensional item response theory (IRT) model is relevant to this question.

Clean Inspection Reports

ATRI (2014) pointed to "a 2012 study that found that only 10.4 percent of roadside inspectors 'almost always' completed a RI report when no violations were issued, while 6.8 percent 'never' completed a RI report [with no violations, presumably]." If this practice involving "clean" inspections is widespread, this is an important source of bias since such inspections will reduce the estimated frequency of violations. A clean inspection provides important information about the extent to which a carrier prioritizes safe operations. It would be helpful to report as many of the clean inspections as possible. One remedy is to make reporting of clean inspections mandatory.

Further, carriers that have only clean inspections for the relevant 2-year period are not given a score in SMS. We understand that a carrier with only clean inspections is an extremely safe carrier and not going to be issued an intervention. But such carriers are peers. Their performance as peers is relevant to understanding what is feasible, and to the relationship between SMS percentile ranks and future crash frequency, and therefore the information is critical to know. In addition, as ATRI (2014) pointed out, this group is different from the insufficient-data group because their future crash risk is considerably smaller. Therefore, FMCSA should change its data sufficiency standards to accept, once a set number of inspections is carried out, carriers with a sufficient number of only clean inspections.

Selection Effects

The carriers pulled over for inspections are often chosen based on such observations as swerving, heavy braking, or speeding, or the trucks appear to be in disrepair. CVSA officials

have no stated set of considerations on which to base their decision as to whether to pull a truck over for inspection; the decision is viewed as a matter of on-the-job expertise acquired over time. These selection effects are not viewed as a concern since it is thought that CVSA officials are generally inspecting the right trucks (although it is known that most inspections occur on limited access roads, so CMVs operating elsewhere could receive some modest "protection" from inspection). There would be a problem if the CVSA officials were using the wrong indications for identifying which trucks to inspect. Then, the SMS measures and percentile ranks for those carriers could be better or worse than those of their peers for reasons other than their safety performance. The best way to learn about such selection factors and their correlation with violation frequency would be to randomly inspect trucks that would not have been selected for inspections. Such a research proposal has been made in the past but is difficult to get approved. We support such a project, since doing so would provide important information on which carriers are and are not inspected, and which indications are more and less important. Finally, we point out that the new approach to SMS described in Chapter 4 jointly models inspections and violations and therefore can make allowance for such effects.

Comparing Carriers of Different Sizes

It should be pointed out that safety event groups are not defined by the size of the carriers. The variables that define safety event groups are the number of relevant inspections for five BASICs, the number of crashes for the Crash Indicator BASIC, and the number of inspections with an unsafe driving violation for the Unsafe Driving BASIC. However, the difference between the current definition of safety event groups and size is fairly modest, so our analysis proceeded as if the categories were defined by size. Further, we see no advantage to defining safety event groups through use of number of crashes or number of inspections in comparison to defining safety event groups by the current APU.

However, if the stratification of carriers by size is carried out using APU, this does run into the increasingly important complication that carriers are using contractor drivers, and renting out vehicles to other carriers and the overall "Uberization" of the industry, but we have not considered how FMCSA should deal with this future problem. The usual justification for the use of safety event groups is that the size of a carrier can imply various aspects of its operation that represent different challenges and opportunities for ensuring safe operations. For instance, very large carriers can afford technology and fatigue management programs that can provide more assistance in developing safe practices. We do not feel that this is a reasonable justification for this peer group stratification. After all, the public has an interest in safe operations regardless of the size of the carrier.

However, as mentioned by GAO (2014), there is a justification of peer grouping by size. While the carriers in a safety event group are intended to be roughly of the same size, there remain substantial size differentials within some safety event groups, especially the group of largest carriers. In those cases, for carriers that have about the same frequency of violations below the threshold, the carrier with the measures with the highest variability will end up above the threshold more often due to the randomness of being selected for inspections and of finding violations. Therefore, formation of safety event groups that are as homogeneous by size as possible (and therefore homogeneous with respect to the variability of the measures) helps to promote fairness.

The problem is fully homogeneous peer groups are not possible. Some additional size heterogeneity will always remain. To avoid harming the smaller carriers in each safety event

group, instead of only publishing the percentile ranks, it might be preferable to include some type of confidence interval with the percentile ranks. In that way, insurance companies, shippers, and the public will be able to observe that a high percentile rank could have been due to random factors. Further, it would also be desirable for FMCSA to take the natural variability of the percentile ranks into consideration in the determination of which carriers receive interventions, rather than just treating the percentile ranks as fixed quantities. (A natural assessment of variability is one of the key advantages of the proposed alternative approach described in Chapter 4.)

We are aware of one other problem raised by safety event groups. We were told by the representative of a carrier that between 2 months, its BASICs did not appreciably change but its percentile ranks increased substantially. The company had grown slightly in size and, as a result, been placed in the next highest safety event group, which was generally a safer group. To avoid this discontinuity at the boundaries of safety event groups, the groups can instead be defined dynamically as the closest so many carriers (which can be a function of the size of each carrier). (See FMCSA, 2014.) Such dynamically defined safety event groups would mostly eliminate boundary problems and would be easy to implement.

While interesting questions about the optimal number of safety event groups and where the boundaries should be set remain, we do not feel this issue is important to research at this time. However, it should be mentioned that in discussions about further stratification of SMS, there is the opportunity to have larger numbers of peers in safety event groups by combining some of them. The trade-off to evaluate is size heterogeneity versus heterogeneity of the driving risks associated with the type of CMV operation.

SUMMARY STATEMENT ABOUT SMS

Conceptually, SMS is structured reasonably. Using the number of violations found during inspections, and the number of crashes, with violations bundled into groups that represent related areas of safe operations, weighting these frequencies by severity and time weights, properly standardizing these counts, stratifying carriers into similarly sized peer groups, and then seeing which carriers are doing worse than the others, is a reasonable approach to the identification of unsafe carriers. However, too much of the detail of what is done is ad hoc. Instead, it would be better to make use of the appropriate statistical model, which will help address many of the issues that have been raised in a natural way and interpretable way.

Conclusion: SMS is structured in a reasonable way, and its method of identifying motor carriers for alert status is defendable. However, much of what is now done is ad hoc and based on subject-matter expertise that has not been sufficiently empirically validated. This argues for FMCSA adopting a more statistically principled approach that can include the expert opinion that is implicit in SMS in a natural way.

3
Application of IRT Models, and Public Release and Transparency

The use of data to measure the performance of individual *units* has a long history in the United States, cutting across a wide range of industries. For example, the U.S. Department of Transportation, along with state and local agencies, collects and reports data on crashes for identifiable subgroups of the population of motor carriers, by specific locations, and associated with various vehicle types, makes, models, and equipment. This approach to assessing safety performance is ubiquitous throughout the transportation agencies that report on safety. In commercial air, rail, and waterway transportation, as with the Federal Motor Carrier Safety Administration (FMCSA), safety reporting is aggregated for segments of the respective industries and specific to individual companies.

To pick one example, the Federal Aviation Administration (FAA) collects data from airlines on accidents and incidents (https://www.faa.gov/data_research/accident_incident/) and on a variety of reported hazards and safety discrepancies, including several voluntary reporting systems. The FAA's safety management process offers an interesting parallel to the FMCSA's Safety Measurement System (SMS). The Aviation Safety Information Analysis and Sharing (ASIAS) program works to monitor known risk, evaluates the effectiveness of deployed mitigations, and detects emerging risk.[1] Components of the ASIAS are designed to monitor safety-related events (including self-reported events) to identify and mitigate risks *before* they result in crashes. The FMCSA's monitoring of safety violations to characterize a carrier's safety culture is also designed to identify risky carriers and hopefully mitigate future crashes.

As another example, the Federal Rail Administration (FRA) keeps records of passenger and freight rail accidents and produces ordered lists by frequency of occurrence that can be queried by the public.[2] This is similar to the way that FMCSA receives reports of crashes and assigns them to the responsible motor carrier. In 2016, the FRA promulgated a rule requiring passenger rail operators to develop and implement a system safety program (SSP) for risk-based hazard management addressing maintenance, inspection, repair, rules compliance and procedures review, employee and contractor training, and outreach.[3] FRA maintains a cadre of inspectors who review rail carriers' SSPs for sufficiency and who perform inspections and issue violations as required. These and other programs in the FRA parallel FMCSA's SMS in fulfilling regulatory requirements and managing safety through monitoring of safety-related behaviors and violations as they relate to risk of accident.

Outside of transportation, the medical and educational sectors have been particularly active in measurement. The World Health Organization proposed several measures of health, based on five dimensions, to compare countries: overall health, inequality in health, fairness of financing, overall health system responsiveness, and inequality in health system responsiveness (Almeida et al., 2000; Coyne et al., 2002). In the United States, the quality of health care

[1]For more information, see ASIAS Fact Sheet, April 2016 (https://www.faa.gov/news/fact_sheets/news_story.cfm?newsId=18195).

[2]For more information, see "Train Accidents by Rail Groups" (http://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Query/inctally3.aspx).

[3]For more information, see FRA-2011-0060, Notice Number 3, July 29, 2016 (https://www.fra.dot.gov/eLib/Details/L18294).

provided by specific delivery systems, such as hospitals, health plans, ambulatory centers, and even individual physicians, is quantified using a variety of data sources (Courtney et al., 2002). In education within the United States, school and teacher performance are assessed with the goal of improving the quality of education in school districts (Rothstein, 2000). These two sectors have a history of measurement and public reporting, with many similar measurement, modeling, and public reporting challenges as those involved in assessing motor carrier safety. Countries, hospitals, physicians, schools, and teachers play the role of motor carrier—the accountable units—while outcomes associated with whole populations, patient encounters with the health system, and students are analogous to safety violations.

ASSESSING THE QUALITY OF MEDICAL CARE

The conceptual foundation upon which medical care delivery is assessed traces back to the work of Avedis Donabedian (1980), who identified three general dimensions of health care associated with a medical provider, whether the provider is a health plan or an individual clinician: structure, process of care, and outcomes of care. Structural measures refer to characteristics of the provider that enable the capacity to deliver high quality of care and include physical attributes such as the existence of computer order entry systems. Process measures refer to what providers do to and for patients, often linked to adherence to established best practices and medical guidelines, for example, if a patient admitted for a heart attack who is eligible for beta-blocker therapy receives a prescription for beta-blocker therapy upon discharge. Finally, outcome measures refer to responses that characterize the patient's health status, such as death within 30 days of admission for a heart attack. Advantages and disadvantages of the specific measures associated with the three dimensions have been discussed at length (see, for example, Birkmeyer, Dimick, and Birkmeyer, 2004). The choice of structural, process, or outcome measures depends on numerous factors including the population size (e.g., the number of patients eligible for a particular therapy), the potential for serious adverse consequences (e.g., the delay in getting a patient having a heart attack thrombolyzed), the frequency of the outcome (e.g., the percentage of patients dying within 30 days of a hospital admission), and the duration of assessment period (e.g., hospitals are assessed using 3-years of data), among others (Birkmeyer, Dimick, Birkmeyer 2004).

Accountability in the Health Measurements

The Centers for Medicare and Medicaid Services (CMS) measures the quality of health care at multiple different levels of accountability for its beneficiaries: hospital inpatient, hospital outpatient, nursing homes, health plan, and physician. For example, several process-based patient safety measures are determined at the physician level, such as the percentage of a physician's older patients screened for a future fall. While screening (a process measure) does not characterize injuries as a consequence of a fall (an outcome measure), it can help identify patients who are at risk of falling. The analogy with highway safety is that safety violations may be viewed as process measures, while crashes are outcome measures. Process measures are more directly actionable to improve outcomes because process measures can suggest specific procedures to improve outcomes. For hospitals, several process and outcome measures are publicly reported (www.medicare.gov/hospitalcompare). Moreover, in addition to publicly reporting measurements, CMS uses some measures to reward health care providers with

incentive payments for the quality of care they deliver to their beneficiaries. In 2012, CMS adopted an approach to calculate excess 30-day all-cause readmission ratios associated with patients hospitalized for a heart attack, heart failure, and pneumonia, and uses this measure in part to calculate a readmission payment adjustment.[4]

Professional societies, such as the Society of Thoracic Surgeons, American College of Cardiology, American Heart Association, and American Board of Internal Medicine, have been involved with quality assessment and reporting as well. These societies provide information to specific clinicians and to delivery systems characterizing their performance at an absolute and at a relative level.

The CMS approach to modeling quality measures varies across and within type of measure. Hierarchical generalized linear models are estimated for hospital outcome measures for specific diseases in which patients are clustered withinhospitals. Because patients are not randomized to hospitals, admission characteristics of the patients are used to adjust for differences in patients across hospitals. Indirect standardization based on the hierarchical model is used to classify hospitals into three categories: those having higher than expected outcomes, those having lower than expected outcomes, and those having outcomes no different from expected.

Combining Health Quality Measures

Given the numerous provider-specific measures utilized by CMS, conveying information to beneficiaries regarding provider quality or using the information to improve quality became complicated. For example, CMS's pay-for-performance project required a valid and reliable methodology for combining the individual measures into a single quality score (composite) for specific hospital conditions. A univariate summary could then be used to categorize the distribution of the condition-specific scores into deciles to identify high-quality hospitals. In a demonstration project (the Hospital Quality Improvement Demonstration Project), hospitals in the top 20 percent were given a financial bonus. The specific measures and conditions were chosen on the basis of the seriousness of the conditions, the size of the populations impacted by the conditions, and the costs associated with the conditions. The methodology required aggregating individual measures to create a summary measure.

ASSESSING STUDENTS, TEACHERS, AND SCHOOLS

Recent years witnessed dramatic changes in the evaluation of the performance of public schools and public school educators. Although student scores on standardized tests have been used by education reformers to promote improvement of what they saw as poor-performing schools since the mid-19th century (United States Congress, Office of Technology Assessment, 1992), test-based accountability for schools, principals, and teachers rose to new dimensions in the 1990s and the early years of the 21st century. States such as Kentucky, North Carolina, and Texas, among others (Grissmer and Flannagan, 1998; Koretz and Barron, 1998), wrote laws and regulations requiring school performance be evaluated annually on the basis of students' average scores or performance levels on standardized tests. These regulations also specified

[4]For more information, see https://www.cms.gov/Medicare/Medicare- Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html.

sanctions, such as possible closure or restructuring for schools where students failed to meet performance targets and financial rewards for schools that performed exceptionally well. With the 2002 re-authorization of the Elementary and Secondary Education Act, commonly known as No Child Left Behind (NCLB), test-based accountability became federal law. The law required public elementary and secondary schools be evaluated on the proportion of their students who scored proficient or better on standardized tests selected and administered by the states. Schools with performance measures that failed to meet federally established targets faced sanctions including possible restructuring or closure. To support the annual calculation of the performance measures, NCLB mandated annual testing for all public school students in grades 3 to 8 in mathematics and reading and in one grade in secondary school in the United States.

Value-Added Models in Teacher Evaluations

Performance measures for education took another turn in the early 2000s with the introduction of value-added modeling and its use for teacher evaluations. Value-added models are statistical models for students' current standardized achievement test scores as functions of prior years' test scores, student background variables, and educational inputs such as schools or teachers. They are closely related to Item Response Theory models. These models are meant to isolate the value of a particular input (e.g., a year of schooling or a teacher) on students' achievement. Unlike the average proficiency measures required by NCLB, which relied only on students' scores at a point in time, value-added models use multiple years of test score data on individual students in their efforts to identify the contributions of individual teachers or schools to student learning (National Research Council and National Academy of Education, 2010).

In the late 1990s, multiple researchers used value-added models to argue that teachers were the most important schooling input to student achievement and that large differences existed in the effectiveness of teachers to support student learning as measured by standardized achievement tests (Sanders and Rivers, 1996; Wright, Horn, and Sanders, 1997). By the early 2000s, this research—along with research demonstrating that the traditional methods of teacher evaluation, which relied primarily on supervisor evaluations—found almost all teachers receiving satisfactory or better evaluations (Weisberg et al., 2009), which gained the attention of policy makers. States and school districts began exploring the use of value-added for teacher and principal evaluations. The federal government again accelerated accountability, this time value-added-based accountability, through the Race to the Top competitive grant program for states The government included in its criteria for award the requirement that states develop revised teacher and principal evaluation systems that rated educators on performance indicators that combined value-added measures and other measures, such as scores from observation of practices (e.g., classroom teaching for teachers) on standardized protocols. These revised evaluation methods were further promoted through waivers as part of the Principal Flexibility Provisions in the Elementary and Secondary Education Act, which, starting in 2011, included requirements for revised educator evaluation systems.

The rapid implementation of value-added models in teacher evaluations generated large concerns in parts of the research community and stakeholder groups about the accuracy of the models for measuring effectiveness of teachers and the quality of their teaching. Critics of value-added measures pointed out that because students were not randomly assigned to classes, there remained potential for value-added measures to reflect student backgrounds in addition to teacher

inputs into learning. Critics also noted value-added measures could contain large statistical errors that could result in year-to-year instability in the measures or misclassification of teachers. A large body of literature on both positive and negative attributes of value-added models quickly developed, and the debate even gained coverage in the popular press after the new teacher evaluation systems went operational.

In 2016 Congress again reauthorized the Elementary and Secondary Education Act. Although the reauthorization retained the annual testing requirement, it removed federally mandated specific penalties based on the scores. In addition, it did not retain any of the requirements for individual teacher and educator evaluations that were part of the waiver program. The law gave states and local school districts greater control over their educator evaluation and school accountability systems than the previous law.

Value-Added Modeling for Student Achievement

Value-added models describe student achievement as a function of multiple factors including teachers or schools. Estimation of the model parameters then provides estimates of the teacher's contribution to achievement test scores. However, because achievement is only one desired outcome of schooling and any given achievement test measures only a subset of the content students are expected to master, value-added is an incomplete measure of teacher effectiveness. In addition, value-added measures are not tied to any specific teaching practices so they provide limited guidance to teachers on how to improve their performance. Consequently, evaluation systems generally relied on performance indicators that combined value-added measures with other measures of teachers' performance including classroom observations. Researchers have explored Item Response Theory and latent variable models to combine multiple measures of teaching (Lockwood, Savitsky, and McCaffrey, 2015); however, state evaluation system rules for combining the various component measures have tended to be ad hoc and often based on multiway classification tables, in part, because the Race to the Top required educators be classified in three or more performance categories.

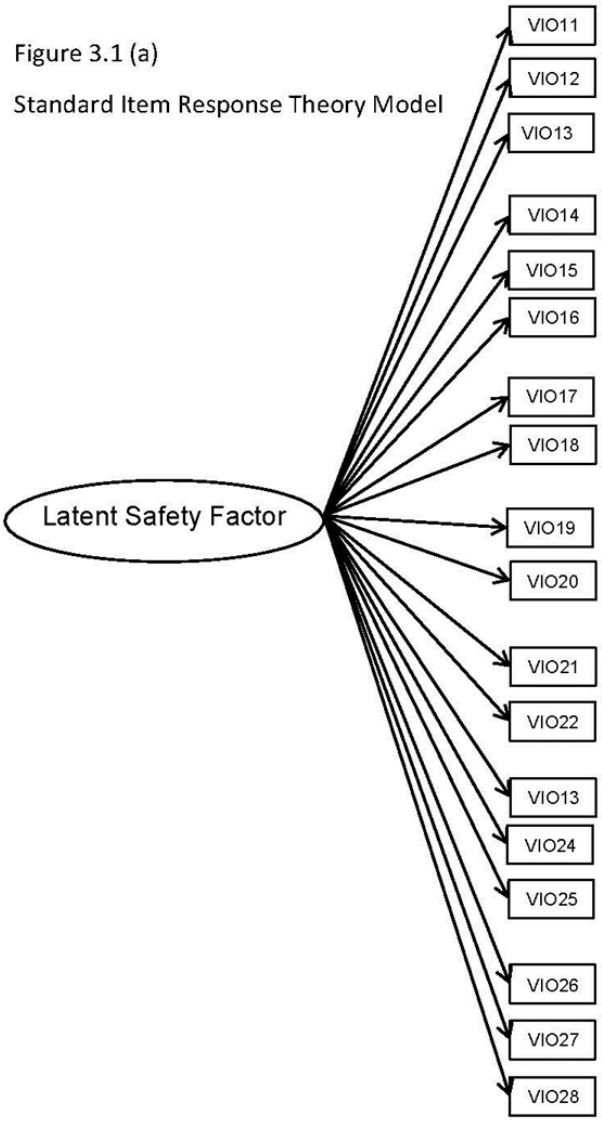
ASSESSING MOTOR CARRIER SAFETY

As highlighted throughout this report, the mission of the FMCSA is to reduce crashes, injuries and fatalities involving large trucks and buses operationalized through safety measurement. Simply stated, safer commercial motor vehicles should have fewer crashes, all else being equal. The panel agrees with a focus on safety; crashes are rare, unlikely to provide information for the majority of commercial vehicles operating in the United States, and thus yield little knowledge to prevent future crashes. FMCSA uses violations, binary-valued variables, as a basis to measure carrier safety culture, assigning violations to carriers as the accountable units. Thus, violations provide indicators of carrier safety under the assumption that safety impacts crashes.

Like measurement in health and education, carrier safety is a complex multidimensional construct requiring multiple measurements. Overall, the rationale for combining the subdomains to create a composite safety score includes providing an overall summary of carrier safety, creating a fairer mechanism to measure carrier safety because multiple violations are needed to get an unsafe score, and improving the statistical properties of estimates of carrier safety.

A psychometric approach to carrier safety culture measurement requires the completion

of several steps to produce valid and reliable summaries (Table 3-1). The measures included in the construct should have a plausible relationship with the carrier's underlying safety culture. The key idea relates to the notion that multiple violations are caused by the safety culture of the carrier, reflected in Figure 3.1(a) by arrows from the unobserved construct to each violation. The strengths of the relationships between each violation and the latent construct (depicted by arrows in the figure) may vary by violation, with some violations having little or no relationship to the carrier's latent safety construct and others having strong relationships. Changes in the underlying safety culture of the carrier would change the risk of each violation—thus, the safety culture construct is reflective rather than formative. Because the violations are binary-valued variables, the strengths of the relationships can be estimated using an Item Response Theory model and their relationship with crash risk directly assessed (Figure 3.1(a)). Item Response Theory models are basically factor analysis models for dichotomous (Y/N) data (i.e., test items), which is the case for the violations that are coded as present or not.



NOTE: Figure 3.1(a) panel generated.

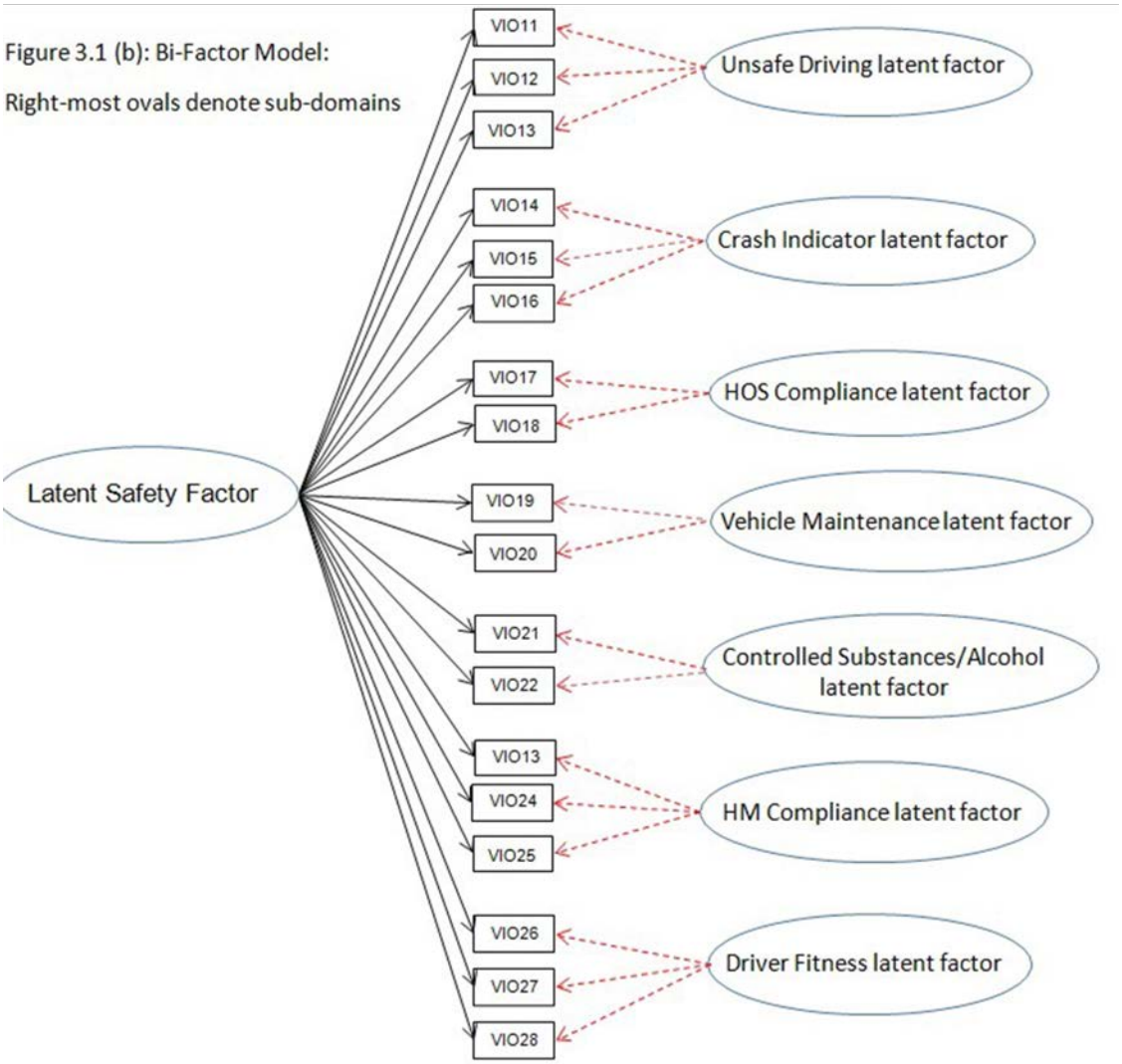
TABLE 3-1 Steps Required to Empirically Derive a Composite Safety Measure

Task	**Example**
Specify safety subdomains	Unsafe driving, fatigued driving
Derive content for each subdomain	Violations assigned to specific latent subdomains
Assess relationship between violations and subdomain score	Item Response Theory (IRT) model
Derive subdomain scores	Estimated from IRT model
Combine subdomain scores	Composite score via aggregation of estimates

TABLE 3-2 Summary of Features Characterizing National Assessments

	Setting		
Feature	Health Delivery	Education	FMCSA
Primary Units	Health plans, hospitals	Schools, teachers	Motor carriers
No. of Primary Units	5000 hospitals	13,600 school districts, 98,500 schools, 3,100,000 teachers	500,000 active carriers
Range of Number of Secondary Units	1 to 3000 patients	About 50.5 million students	1 to several thousand power units
Unit of observation	Patient discharge	Student exam	Violation-inspections
Measure Type	Outcomes: process, structural	Student outcomes, teacher practices	Process
Period of Assessment	3 years	1 year	2 years
Frequency of Assessment	Annually	Annually	Monthly
Stratification	Medical condition, measure type	None	Carrier size, straight or combination
Public Reporting	Hospitals point and interval estimates	Schools point estimates	Carrier point estimates
Data Sufficiency Standards	Minimum 25 for public reporting, minimum 1 for calculation	Varies by state	See Appendix C

Given the broad and encompassing nature of carrier safety culture, the creation of several subdomains (denoted as specific BASICs by the right-most ovals in Figure 3.1(b)) by FMCSA is sensible. However, the subdimensions and the overall score should be more thoroughly empirically tested to assess the reliability and other operating characteristics, such as construct validity—the degree to which each BASIC measures the types of unsafe operation it is intended to, and discriminant validity—the degree to which the BASICs measure separate things. Finally, determining a meaningful difference would enhance the interpretation of the summary score.



NOTE: Figure 3.1(b) panel generated.

There are additional considerations, however, that complicate the creation of a safety construct using violation data. First, violation information is collected only when carriers are inspected. Second, the risk of inspection depends on systematic factors, some but not all measured, as well as random factors. Third, both the risk of inspection and the risk of a violation once inspected vary by the number of active trucks associated with a carrier, in addition to systematic and random factors. Fourth, the hierarchical structure of the data is deep—the data are reported at the inspection-violation level repeatedly over time and, theoretically the inspections are nested within driver.

These considerations are summarized in Table 3-1 above. Table 3-2 provides a summary of features characterizing national assessments.

Figure 3-2 (explained more fully in Appendix C) illustrates the relationship between each of the violations comprising the Unsafe Driving Basic obtained from estimating a 2-parameter logistic Item Response Theory model. The x-axis represents the underlying safety construct for each carrier while the y-axis represents the probability of a specific violation. More discriminating violations are identified by steep slopes while more frequent violations

have functions that start further to the left.

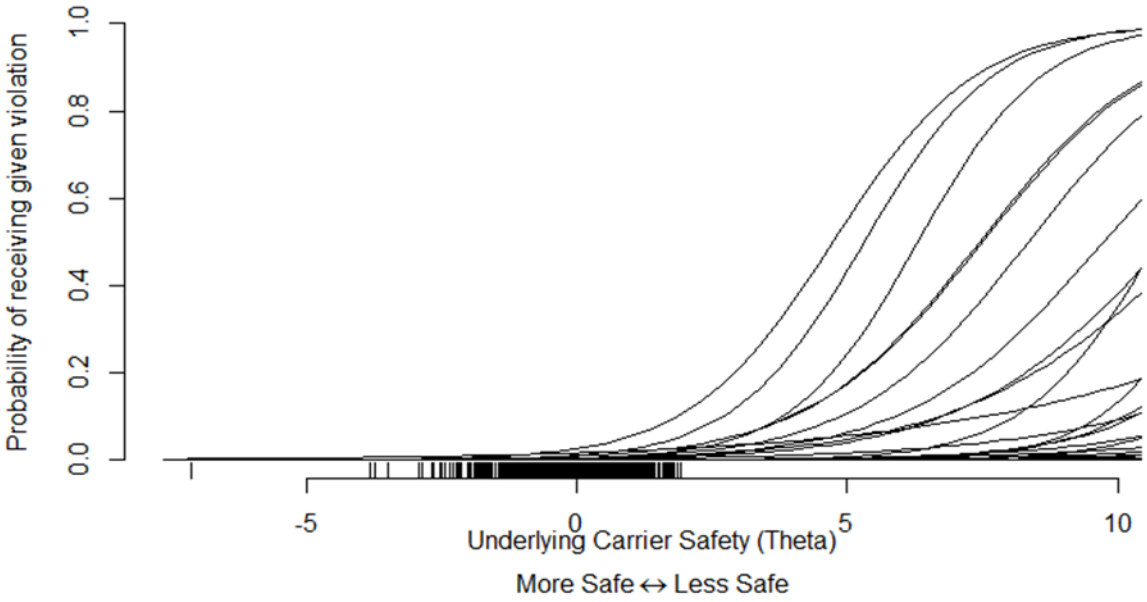


FIGURE 3-2 The Relationship between Carrier Safety Construct (Theta) and the Probability of Receiving Given Violations.
SOURCE: Panel's analysis as presented in Appendix C.

PUBLIC DISCLOSURE OF SMS MEASURES AND PERCENTILE RANKS

From 2010 to 2015, FMCSA made the measures and the associated percentile ranks for five of the seven BASICs publicly available on its website, with the exceptions of the Crash Indicator and Hazardous Materials Compliance BASICs. As a provision of the Fixing America's Surface Transportation (FAST) Act of 2015, FMCSA was precluded from publishing the SMS percentile ranks for any BASICs, though it was still permitted to release the BASIC measures for the five BASICs previously available. This policy remained in place while the National Academy of Sciences (through this study panel) examined the quality of SMS. (For those five BASICs, the FMCSA website also includes information on crashes, roadside inspections, and violations resulting from roadside inspections; the prohibition from publication only pertains to property carriers, and not to passenger carriers; further, all of SMS is available to law enforcement.)

There are advantages and disadvantages to disclosing the SMS percentile ranks to the public. The advantage is primarily that publicity is an important motivator. The goal of SMS is to encourage poor safety performers in the commercial motor vehicle (CMV) industry to devote greater attention to safe operations, and this is accomplished by intervening with those carriers that have frequent violations and crashes. The hope is that the initial interventions, typically notifications, motivate carriers to make changes so that their BASIC percentile ranks become

more representative of typical carriers. Making the percentile ranks public creates pressure on motor carriers to make these changes because they then face the prospect of losing business to their rivals or having their insurance rates increased. In this way, FMCSA is employing competition to incentivize safer behavior. (As noted throughout this report, FMCSA also provides incentives through other increasingly serious forms of intervention.)

The disadvantages of public reporting stem from the fact that SMS is not a perfect discriminator between the carriers that need to improve their safety performance and those that do not. It may be impossible to precisely define what could be meant in this situation by a false negative or a false positive intervention, since there is no objective definition of what it means for a carrier to be an unsafe carrier. But hypothetically assuming achievement of a thorough investigation of what happens during maintenance, scheduling, loading cargo, hours driven per day, and other activities for all 550,000 active carriers, every carrier could be judged to be or not be operating safely. In that case, SMS almost certainly identifies some carriers for interventions that would have been judged to be operating safely, and it fails to identify some carriers for interventions that would have been judged to be operating unsafely. Both are problematic, and if they become too frequent, could result in a program whose costs outweigh its benefits. Our impression is that SMS has a lot of true positives and a lot of true negatives: that is, SMS does not issue interventions to many carriers that are operating with a high priority for safety while it does issue interventions to many carriers that need to give safe operations a higher priority. However, there is a particular concern about false negatives and false positives among smaller carriers, which results from not having much data with which to judge them. It should be stated that crashes are rare, so it is possible to operate for many years without a serious crash, even if a carrier is operating in a risky manner. Therefore the Crash Indicator BASIC could never be a perfect discriminator of small carriers. For example, the National Highway Traffic Safety Administration (NHTSA) estimated that in 2012, the fatal crash rate for large trucks was 1.42 per 100 million miles. The injury crash rate was 29 per 100 million miles. If a truck averages 150,000 miles per year, on average it would go 470 years between fatal crashes and 23 years between injury crashes. If the truck were operated twice as riskier as average, it would still average 235 years between fatal crashes and 11.5 years between injury crashes. Thus, because crashes are rare, it is necessary to use alternative measures of safe behavior, which are the BASICs.

In addition, though the following effects are generally small and therefore their impact infrequent, there are a number of ways in which a borderline carrier could have its percentile ranks go from the nonintervention side to the intervention side, or vice versa, from one month to the next, which may also be a source of some false negatives or positives. These effects include the following:

- A carrier's measure of size or average power units (APUs) decreased or increased, and was updated, and the carrier is now in a different safety event group whose general performance is worse or better. Further, even staying within the same safety event group, if, for a carrier, the denominator for the Crash Indicator BASIC score increased or decreased due to a larger or smaller utilization factor, or due to fewer or more power units, its percentile ranks could increase or decrease.
- If a carrier had a month just over 2 years ago in which it had a very low or very high frequency of violations (now outside of the 2-year SMS window), the data for that month are now replaced with data of the most recent tallied month in which the safety

performance was more typical of the 2-year period. It should be noted, ignoring time weights, that a change in a percentile rank from one month to the next is the result of changes to 1/24th or about 4 percent of the data for a carrier over a 2-year period, so it is difficult for a carrier to make large changes in its measures or ranks in a short amount of time.

- As noted in Chapter 2, percentile ranks could go up or down if a carrier's peers in its safety event group have all improved or worsened more than the carrier in question.
- Random variation occurs because there is a random element in who gets inspected.

The panel heard from various stakeholders and representatives from carriers both encouraging and discouraging the release of the percentile ranks. John Lannen of the Truck Safety Coalition told the panel that due to many of its features, CSA/SMS is used by many in the industry, and it has brought about a positive change in the general safety culture. He worried that not disclosing the measures and percentiles will deprive the public, shippers, and insurance brokers from learning about the comparative safety of motor carriers. According to Shuie Yankelewicz and Jean Gardner of the Central Analysis Bureau, SMS percentile ranks create a common language among CMV industry partners who are focused on monitoring and risk assessment of CMV operations, such as insurance companies, insurance agents and brokers, shippers, safety and risk management companies, and loss control companies. SMS percentile ranks fulfill the need of the CMV industry to have a measure of risk assessment so that they know what type of insurance options to offer, the premium to charge, and what kind of coverage to offer. Data-based risk assessment measures help the industry to comply with regulations. Due to these advantages, there is a strong interest by many in the industry that FMCSA continue to produce and disclose SMS percentile ranks.

An argument raised by William Voss against making the BASIC percentile ranks public is that it might lessen the motivation for carriers to collaborate on techniques that have been found to be effective in bringing about safe operations. Such collaboration has occurred in other transportation industries as reported by the Independent Review Team (2014). The National Transportation Safety Board (NTSB) has recommended that FMCSA, the industry, and other stakeholders develop a mechanism that allows for cooperative development and coordinated implementation of voluntary safety programs. Also, Mark Burroughs of Transportation Intermediaries Association pointed out to the panel that the disclaimer language regarding SMS measures is confusing to shippers and brokers, motor carriers, and insurers and can lead to increased litigation costs. Due to the confusion stemming from SMS results for motor carriers, shippers and brokers are being brought into lawsuits and named as defendants for negligent selection of a motor carrier when an accident occurs. The situation is complicated further because state courts take different approaches to interpreting the disclaimer language on FMCSA's website. Burroughs recommended that as a result, FMCSA should not disclose SMS measures and percentile ranks and use them for internal prioritization only.

Based on these considerations, the panel is unable to recommend to FMCSA whether or not to make SMS percentiles public. What is needed is a greater understanding of the consequences of public consumption of the information, which would necessitate a formal evaluation using randomized or controlled release of specific components of the SMS percentiles. The panel would also need to better understand the statistical operating characteristics of the measures and percentile ranks to judge decisions regarding their usability.

Recommendation: FMCSA should undertake a study to better understand the statistical operating characteristics of the percentile ranks to support decisions regarding the usability of public scores.

While we have discussed assessing the operating characteristics of SMS with regards to the issue of public release of SMS percentiles, doing this is also an important component of a full validation of SMS. This is because the evaluation summarized in Chapter 2 is carried out on an aggregate basis, which does not directly assess the extent to which carriers that have frequent violations are those that have policies and procedures that contribute to unsafe driving. A full validation of a model in this context should have three components: (1) model fit, (2) the degree to which frequent violators, as identified by the algorithm, are those that operate unsafely, and (3) the degree to which those that operate unsafely have frequent crashes. Assessment of model fit is relatively trivial for SMS, since what is used is essentially weighted averages. However, there remains the evaluation of the model for its stated goal. This means determining the degree to which carriers that have high violation rates are those that are operating unsafely. FMCSA has not directly validated this because it is expensive to do and because they look at aggregate crash rates for those selected and not selected for interventions as an overall evaluation. Although the panel is not recommending this since we believe resources should be invested in a principled scientific approach such as the proposed IRT model, if FMCSA chooses to retain SMS, we believe that it is important for the agency to carry out such an evaluation.

Finally, we point out that the improved approach to SMS outlined in Chapter 4 of this report would very likely improve the "false negative" and "false positive" performance of SMS and thereby provide greater justification for release of the percentile ranks.

TRANSPARENCY NEEDED FOR SMS

As with any release of measures and percentile ranks judging performance with consequences, there is a benefit to transparency. Transparency alleviates worries about fairness of application that can persist in its absence. Should a carrier's rank move it into alert status after a period of time in which it was not in alert status, a carrier would first check to see if its frequency of violations in one of the six BASICs or frequency of crashes had increased considerably during the previous month. If this were not the case, a carrier would worry about a mistake. Unfortunately, both the existing SMS and approach recommended in Chapter 4 are complicated, which makes such assessments difficult.

Further, motor carriers have difficulty monitoring their measures or progress on their own, which will continue if FMCSA adopts the new approach. They operate without knowing, day to day, the impact of their inspection results and crash experiences on their BASIC measures because they cannot easily run the relative comparisons on a continuously updated basis. Therefore, the plan to have scores affect behavior is not functioning as well as it might if the motor carriers could make their own computations of their BASIC measures on a demand basis. This ties back to the overall goals and objectives of SMS and, if addressed, would make the system more effective than it is now in improving safety. Motor carriers could actively manage their business in ways that they know would improve their BASIC measures. A complication is that to see the impact of a carrier's recent performance or to produce indications of alert status for the carrier, it is necessary not only to view the carrier's data, but also the input data for all carriers in the same safety event group (which is almost certainly different for each BASIC).

The Independent Review Team (2014, p. 25) commented on SMS transparency, noting the following:

> There is concern in the industry about the transparency of the BASICs information and its effect on the reputations of their businesses. Throughout society, however, consumer and public protections are being enhanced in many other areas through increased use of public transparency and dissemination of safety ratings and assessments. Availability of such data informs the public and helps them make better choices. Poor performers suffer in the marketplace, and better performers gain market share. The government—if it is the party releasing the data—has the obligation to ensure data quality. Safety ratings should obviously be a fair reflection of a motor carrier's operation; and the more accurate they become the more useful they will be in informing public choice and enhancing safety. Because many stakeholders (e.g., shippers, insurers, and litigants) assume SMS data reflects safe versus unsafe operations, FMCSA should take steps to clearly identify for the public the information that can be tied reliably to safety; and to distinguish it from other information that may be useful for other reasons but does not relate to crash risk.

A first step in this direction would be if carriers could better understand how the inputs are used to produce the current SMS meaures and alerts. For this purpose, FMCSA could provide an accessible description about how SMS functions. In addition, FMCSA could explore development of software, possibly a mobile application (app), that would assist carriers in determining why their percentile ranks changed from one month to the next. Such an app could also be a mechanism for collecting more timely data on such things as vehicle miles traveled, number of power units, and type of operation.

Transparency also applies in a somewhat different way to the research community. Researchers would greatly benefit if FMCSA, once the approach presented in Chapter 4 is implemented, could make well-commented code more widely available and if FMCSA could restructure the MCMIS data in a way that would facilitate research by those not adept at complicated ORACLE database structures. This could be done in a way that protects personally identifiable information.

Recommendation: FMCSA should structure a user-friendly version of the MCMIS data file used as input to SMS without any personally identifiable information to facilitate its use by external parties, such as researchers, and by carriers. In addition, FMCSA should make user-friendly computer code used to compute SMS elements available to individuals in accordance with reproducibility and transparency guidelines.

WIRELESS ROADSIDE INSPECTIONS

FMCSA is currently field-testing technologies in three states that can identify CMVs and violations concerning registration, hours-of-service, licensing compliance, and some safety violations. A wireless inspection report (WRI) would be sent to inspectors to enhance their ability to identify noncompliant CMVs. The field test includes testing of multiple vehicles from multiple motor carriers in multistate corridors. A wireless inspection will require modifying existing telematics systems in CMVs. The goal of using the wireless inspection technology is to carry out inspections as vehicles pass at speed so that compliant trucks would not have to stop for an inspection. Provided such telematics are fitted on to all trucks, WRIs will enable collection

of some inspection data on all CMVs instead of the current system where a selected set of CMVs are pulled up for inspection, thereby reducing selection effects. As GAO (2014) pointed out, this could help a great deal in the assessment of small carriers for which there are currently little inspection data generally available.

4
Item Response Theory Models Applied to Highway Safety

In this chapter we describe a formal statistical model for estimating a carrier's overall safety using item response theory (IRT) models introduced in Chapter 3. The current BASIC methodology, which uses data regarding violations collected from various types of truck and carrier inspections, has many features in common with IRT models. In an inspection, numerous violations are assessed and each violation is associated with a single Behavior Analysis and Safety Improvement Category (BASIC). If a violation is similar to a "test item," then this approach of assigning violations to the BASICs essentially corresponds to a multidimensional confirmatory IRT model (see Figure 3-1). That is, the various BASICs are the multiple factors of the IRT model, and the assignment of violations to these factors is known, and so it is a confirmatory model. An IRT model is basically a factor analysis model for dichotomous (Y/N) data (e.g., test items), which is the case for the violations that are coded as present or not. Also, in obtaining a score on a given BASIC, each violation has a severity weight attached to it. These are akin to the role of the item difficulty and discrimination parameters of an IRT model, which treat items differently depending on their frequency and how related they are to the underlying factor(s). The main difference is that the BASIC methodology uses severity weights that are dependent on expert opinion and empirical observations in a less empirical and static manner, whereas the item difficulty and discrimination parameters are estimated based on a formal combination of the observed data and expert opinion through the use of priors that are updated dynamically as more data are collected. Thus, the severity weights for both SMS and in an IRT model are intended to be related to the underlying factors for the outcomes, but in the case of an IRT model, the data will ultimately refine the weights to do so should expert opinion be wrong. To summarize, the BASIC methodology is essentially a multidimensional IRT model that uses expert opinion and empirical data together in a scientifically prinicipled way.

Despite the conceptual similarities, the IRT model offers several advantages over the existing BASIC methodology. The IRT modeling framework can do the following:

- account for the probability of being selected for inspection by explicitly modeling the likelihood a carrier is inspected, as a function of carrier characteristics (e.g., so if it were known that carriers of type X were inspected more often than carriers of type Y, everything else being equal, including their approach to safe operations, then increases in percentile ranks for carriers of type X over type Y should be reduced);
- provide a basis with which to evaluate how data insufficiency could impact safety ratings of carriers (e.g., impact of inaccuracy of vehicle miles traveled [VMT] on safety ratings) since it yields measures of uncertainty of the latent safety score for each carrier, unlike the current deterministic algorithm;
- provide a basis to more rigorously and empirically evaluate the utility of individual violations since it yields measures of relevance of individual violations—in other words, the extent to which each violation is related to the latent safety construct;
- allow severity weights to change over time (e.g., as violations become more or less prevalent);
- determine empirically whether severity weights should be different for trucks versus passenger carriers;

- enable adjustment for factors that may be outside a carrier's direct control or differ systematically (e.g., state-level and seasonal effects),
- accommodate new violations over time easily, which is not the case for other test theory models that might have been proposed for use.

Thus, overall, the IRT modeling approach enhances the transparency of the safety evaluation system and provides approaches to evaluate it.

A NEW MODEL

We use the following notation to lay out the modeling strategy. The index i $(i = 1,...,n)$ denotes carriers. The total number of inspections for carrier *i* is denoted as N_i. The index *j* (*j*=1,…6) denotes the inspection level/type (see Glossary for explanation of the six inspection levels/types). Based on inspection type, *j*, a subset of the 899 possible items are actually inspected; these items are represented by the index *k*. The set of possible items for inspection type *j* is denoted by S_j, $j = 1,2,...,6$. N_{ij} is the number of inspections of type *j* for carrier *i*; thus the total number of inspections for carrier *i*, $N_i = \sum_{j=1}^{6} N_{ij}$. N_{ijk} is the number of inspections for carrier i of type j where item k was inspected , i.e., $N_{ijk} = N_{ij} I\{k \in S_j\}$. Finally, $Y_{ijk} = 0,1,..., N_{ijk}$ is the number of times item *k* was found to be in violation among the N_{ijk} inspections of type *j*. Note that the grand total of inspections where a particular item *k* was inspected would be given by $N_{i.k} = \sum_{j=1}^{6} N_{ij} I(k \in S_j)$, and the total number of times the item was found in violation is $Y_{i.k} = \sum_{j:k \in S_j} Y_{ijk}$.

The model is specified in stages. For the first stage, the total number of inspections of type *j* for carrier *i*, N_{ij}, is assumed to follow a Poisson distribution (a common distribution used for count data) with parameter χ_{ij} The parameter represents the average number of inspections for carrier *i* of type *j*. Here we decompose χ_{ij} as $\chi_{ij} = E_i \lambda_{ij}$, where λ_{ij} is the rate at which carrier *i* receives inspections of type *j*, and E_i is the exposure for carrier *i* (e.g., VMT). So the rate, λ_{ij}, is based on the average number of inspections normalized by the 'exposure' (the exposure can be thought of like a denominator). The rate, λ_{ij}, is allowed to depend on carrier-level characteristics, X_i, that may influence the rate at which a carrier receives inspections. Formally, the first stage of the model , for each inspection type *j*=1,2,…, 6 is given by

$$N_{ij} \sim \text{Poisson}\left(E_i \lambda_{ij}\right) \qquad (4\text{-}1)$$

$$\log(\lambda_{ij}) = \gamma_{0j} + X_i \gamma_{1j} \qquad (4\text{-}2)$$

The covariates (carrier level characteristics) included here should not be inherently related to safety (such as driver training programs), but rather reflect the operational characteristics of the carrier (e.g., where they travel the most or register, cargo type, and type of roads traveled).

The parameters γ_{0j} and γ_{1j} are coefficients corresponding to carrier-level characteristics; the values of the coefficients may depend on the inspection type *j*.

While a simpler model on N_i instead of (4-1) could be posited, the model above explicitly allows for the possibility that different types of inspections are more frequent, and that different carriers receive certain types of inspections more frequently than others. These possibilities are embedded in the model by allowing the γ_{0j} and γ_{1j} to be different for different inspection types. For instance, carriers that have a higher percentage of their operations in cold-weather states may receive a lower number of undercarriage inspections during the winter. Finally, there may be many carriers without inspections and some with a very large number of inspections. Hence a zero-inflated Poisson (ZIP) model will likely be needed in lieu of Equation (4-1) (Lambert, 1992). Such a model allows for more zeros (i.e., carriers with no inspections of type *j*) than is implicity assumed by the Poisson model.

The next stage is to model the number of violations for item *k*, Y_{ijk}, for each inspection type *j*=1,2,…,6 where the *kth* item is inspected (i.e., if Ind ($k \cdot S_j$)=1)). So we assume for $N_{ijk} > 0$, i.e., the number of inspections of type *j* where item *k* could be assessed is nonzero, that Y_{ijk} follows a Binomial distribution with sample size N_{ijk} and probability (of a violation) that follows a generalized linear model. In particular, a logit transformation of the probability is a function of three 'parameters': θ_i, an "overall" unobserved (or latent) safety measure for carrier *i*; α_k, parameters that capture how strongly item *k* is related to the latent safety measure θ_i for carrier *i*, which take the place of severity weights, and which determine how well that violation discriminates safe versus less safe carriers, and β_k, which when transformed as $\frac{\exp(\beta_k)}{1+\exp(\beta_k)}$ represents the prevalence of violation *k* for carriers with an average safety measure $(\theta_i = 0)$. As such, these parameters help us determine in a data-driven way how different violations help us differentiate the safety levels of carriers. Expert opinion can be incorporated into these models in terms of different violations ability to "discriminate," which is not quite the same as the expert opinion that was used to determine severity weights in the SMS.
More formally, the model is written as follows:

$$Y_{ijk}|N_{ijk}>0 \sim Binomial(N_{ijk}, p_{ik}) \qquad (4\text{-}3)$$

$$logit(p_{ik})^1 = \beta_k - \alpha_k\theta_i \qquad (4\text{-}4)$$

$$\theta_i \sim N(0,1) \qquad (4\text{-}5)$$

Large, positive values of the latent safety measure θ_i indicate a safe carrier; large negative values of θ_i indicate a less safe carrier, and $\theta_i = 0$ represents "average" carriers. Chapter 5 elaborates on how θ_i's can be used to identify carriers that need interventions from the Federal Motor Carrier Safety Administration (FMCSA).

[1] Logit(p) is log(p/(1-p)).

The number and type of violations for each carrier allow us to identify and estimate both the 'regression coefficients', α_k, β_k, and the safety score θ_i. The α_k is estimated based on different violations occurring together (or not) and β_k (as discussed earlier) is a function of the prevalence of violation k for a safety score of 0. For the safety score θ_i, the probability of a violation increases (or decreases) based on the sign of α_k as θ_i increases. So the complete set of violations for a carrier along with the α_k and β_k (which are estimated across all the carriers) provides the information needed to estimate θ_i (along with the assumption about the shape of the distribution of the θ_i's which is assumed to follow a standard normal distribution (bell-shaped curve)).

Modeling the number of violations of item *k* for each inspection type *j* separately reduces biases related to aggregating counts of violation *k*, i.e., $Y_{i.k}$ that may arise by instead modeling $Y_{i.k} \mid N_{i.k}$. For instance, aggregating could artificially attenuate (or induce) correlation among violations that are (or are not) part of the same inspection type.

We note that some might feel it more natural to specify the unit of analysis at a less aggregate level. While the majority of IRT applications involve modeling hierarchically structured data, there are applications using the aggregated (level 2) units. (See, e.g, Camilli and Fox, 2015). The panel felt that the computational efficiency of the aggregated analysis would be more practical for implementation purposes to the FMCSA.

MODEL ESTIMATION

We anticipate the model will be refitted and updated on a monthly basis, similar to the current practice of updating the BASIC measures to incorporate new data. Re-estimating the model on a monthly basis has the advantage of allowing the severity weights to be data-adaptive in the long run. That is, as certain violations become rare, the information they provide regarding the latent safety scores diminishes, which would be reflected in the parameters α_k becoming unimportant.

We advocate a Bayesian approach for model estimation for several reasons. First, the Bayesian approach is the most natural way to incorporate expert opinion into the model. One area where this could be particularly beneficial is to enable the use of information that led to the development of severity weights in the current Safety Measurement System (SMS) (and/or more refined information of that type) in order to specify informative priors for the discrimination parameters/factor loadings $(\alpha_k's)$. That is, if violation *k* is a priori thought to be more related to safety, then its prior could be centered at a higher positive value. Note that when we use such informative priors, the ultimate estimates will be a combination of the informative prior and the data. Eventually with enough data, the data will "swamp" the prior.

Second, in creating a scoring system for carriers, it is also important to convey a sense of the degree of uncertainty in the rankings or scores. The Bayesian framework also provides the most natural way to accurately quantify this uncertainty. In other words, this estimation method can produce credible intervals (analogous to confidence intervals in frequentist statistics) for the safety measure of each carrier, θ_i. This is particularly important for carriers that have received few inspections. This approach can therefore produce a safety measure for all carriers that have received at least one inspection, and the credible interval can convey that given the amount of

available data, there is little precision in the estimated safety measure. Posterior probabilities of being in the lower and higher percentile can also be easily computed; details can be found in Chapter 5. Third, this estimation approach enables the use of "shrinkage"-type priors, which will allow more precise estimation in the presence of sparsity. In particular, for low-prevalence violations, such priors will allow corresponding parameters to be estimated with moderate precision. Fourth, the proposed models could be fit within standard open source software.[2] Finally, it is important to point out that if there are violations that FMCSA believes are important to retain regardless of the empirical evidence, they can always do so through the use of priors that are either very focused on specific values or are deterministic. For instance, if some hazmat violation rates were found not to be strongly associated with safe operations, but FMCSA thought important to retain, instead of downweighting those violation rates through updating of a prior, FMCSA could specify a prior distribution that assigns high a priori probability to the severity weights of those violations.

ITEM RESPONSE THEORY MODEL MODIFICATIONS

In this section we propose a number of extensions and modifications to the model described above. These extensions incorporate some of the unique features of the BASIC data including: data sparsity/sufficiency given that some carriers are never inspected; the longitudinal nature of the violations and approaches to more heavily weight recent violations; and the multidimensional nature of the violations versus assuming a single dimension for safety. Chapter 5 addresses additional considerations such as whether motor carriers and freight can and should be combined and state-level effects.

Joint Modeling of the Number of Inspections and Frequency of Violations

The motivation for this extension to the model is that the total number of inspections for carrier *i*, N_i, and/or the number of each type of inspection, N_{ij}, can provide information on the safety of the carrier. If unsafe carriers tend to get inspected more or less than safe carriers, then this relationship can be modeled by including additional effects in the model that link the rate of inspections and the probability that violations are found during an inspection. For example, unsafe driving violations (e.g., speeding) often precipitate an inspection, such that there may be a correlation between number of inspections and safety. The information that a carrier of more than a few vehicles had no inspections during a month could be indicative of a very safely operating carrier, and could be helpful in identifying such carriers. However, whether this should result in IRT percentiles in specific instances in this application is unclear and would need some research.

Specifically, the model for the number and rate at which carrier *i* receives inspections of type *j*, Equation (4-1) - (4-2) can be augmented to include an effect b_i:

$$N_{ij} \sim Poisson\left(E_i \lambda_{ij}\right) \quad (4\text{-}6)$$

$$log(\lambda_{ij}) = \gamma_{0j} + X_i \gamma_{1j} + b_i \quad (4\text{-}7)$$

[2]For example, JAGS or winBUGS could be called from R. However, with large amounts of data, specialized software might be more computationally efficient.

This inspection effect of each carrier, b_i, is then linked to the latent safety measure θ_i in Equation (4-4) through

$$b_i \mid \theta_i \sim N\left(\eta\theta_i, \sigma_{b|\theta}\right)$$

Of partcular interest here is the association between the carrier's safety θ_i and inspection effects b_i, which is expressed by the regression coefficient η. A negative value of η suggests that safe carriers have a lower rate of inspections than unsafe carriers, while a positive value suggests that safe carriers have a higher rate of inspections. When η is equal to zero, then there is no benefit to this joint modeling approach because the number of inspections N_{ij} provides no information on the safety of the carrier.

To allow more flexibility in the relationship between "safety," θ_i, and carrier-level heterogeneity in the violation rate, b_i, we can consider more flexible distributions than a bivariate normal, e.g., using variations on Dirichlet process mixtures of normal distributions (which is a way of putting probabilities on different normal distributions that can be updated to reflect new data), that might also allow nonlinear relationship between θ_i and b_i.

Models that Downweight Violations Further Back in the Past

An additional extension of the model can downweight violations that occurred least recently, similar to time weights used in computing BASICs. However, we advocate letting the weights change smoothly over time, rather than using time severity weights that change at discrete time windows, so that changes in the safety scores occur smoothly.

We assume safety scores are estimated based on a 24-month window and monthly counts of inspections and number of violations. The index *t* will take values, *t*=-24,…,0 (months). The model construction provides an overall safety measure for the 24-month window and downweights violations from earlier in the window.

Let Y_{ijkt} indicate the number of type *k* violations received at inspection of type *j* by carrier *i*, in month *t*. The variable N_{ijkt} is the number of inspections carrier *i* received of type *j* with violation *k* in month *t*. Also, E_{it} is the exposure (e.g, VMT) of carrier *i* in month *t*. The key changes from the previous model include the number of inspections and violations now being measured each month (so an index of *t* has been added), the exposure and rate parameters for the number of inspections for each carrier now being indexed by *t* as well, and the IRT parameters, β_{kt} and α_{kt} now being estimated as a function of the month. The overall safety parameter, θ_i, does not change with the month, which we discuss below. In particular, we propose the following model:

Error! Bookmark not defined. $$N_{ijt} \sim Poisson\left(E_{it}\lambda_{ijt}\right) \quad (4\text{-}8)$$

Error! Bookmark not defined. $$log\left(\lambda_{ijt}\right) = \gamma_0 + \gamma_1 X_{it} + b_i \quad (4\text{-}9)$$

Error! Bookmark not defined. $$Y_{ijkt} \mid N_{ijkt} > 0 \sim Binomial\left(N_{ijkt}, p_{ikt}\right) \quad (4\text{-}10)$$

Error! Bookmark not defined. $$logit\left(p_{ikt}\right) = \beta_{kt} - \alpha_{kt}\theta_i \quad (4\text{-}11)$$

where:

$$\begin{pmatrix} b_i \\ \theta_i \end{pmatrix} \sim N_2\left(0, \Sigma\right)$$

and where $N_{ijkt}=N_{ijt}I\{k \in S_j\}$ and Z_{ijt} can include time-varying covariates (e.g., seasonal variables) in addition to static variables specific to carrier and inspection. As above, θ_i represents an "overall" safety measure over the current time window. However, as time evolves, θ_i will change smoothly given that if safety measures are updated monthly, there is a 23-month overlap for safety measures computed in consecutive months.

We agree with the premise that the overall safety measure "downweights" violations further away from the current time (i.e., the end of the time window), rather than discrete time intervals. Hence, in this model, we suggest to model the α_{kt} using a monotone function of month *t*. However, given that there is likely little (if any) information in the data to identify the form of the α_{kt} over time, we suggest a parametric function (e.g., a cdf) be used to constrain the shape of the α_{kt} as a function of time. For instance, we could set $\alpha_{kt} = \alpha_k \Phi\left(t^*\right)$, where $\Phi(\cdot)$ is a normal cdf and $t^\star = -2 + \frac{(t-1)}{5.75}$. This function, shown in Figure 4-1, sets the parameter α_{kt} for the most current month to be α_k (essentially), and zero (essentially) for violations that occurred 24 months ago (month $t = -24$). Expert opinion can again be used to determine informative priors for the parameters α_k. Analogous to the model above, β_{kt} is linked to the prevalence of a particular violation, but now depends on time which allows the violation "prevalence" to vary over time. Due to likely sparsity when we consider 1-month time windows, a shrinkage prior on β_{kt} could be used that shrinks the β_{kt}towards their average—in other words, analogous to the average prevalence of the violation over the 24-month time window with some month-to-month variation.

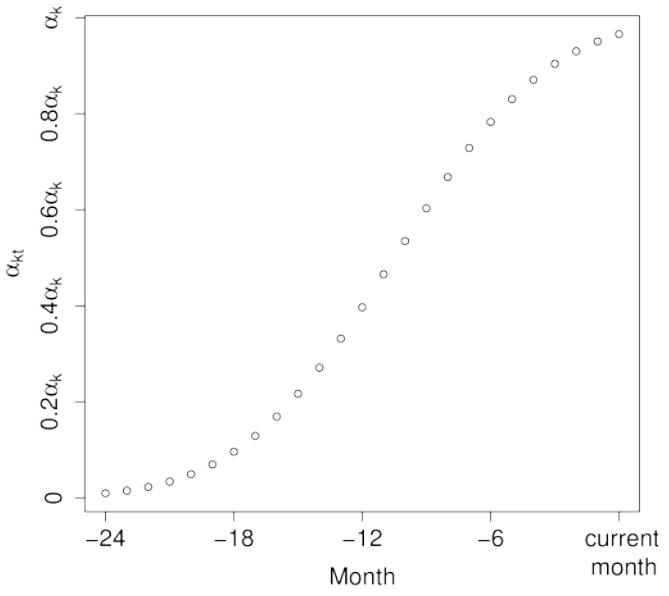


FIGURE 4-1 Parameters α_{kt} as a function of *t* and α_k.
NOTE: This figure shows how violations of type *k* further in the past have less impact on the safety parameter, θ_i, than those in the current month via the feature that α_{kt} decreases as one moves further away from the current month.
SOURCE: Panel-generated.

Multidimensional IRT Model

The model in Section 2, and depicted in Figure 4-2(a), assumes unidimensionality: that is, all of the violations measure a single underlying latent trait (safety). In some settings, if the data are multidimensional, the application of a unidimensional IRT models may result in distorted estimates of the latent dimension. As mentioned earlier, the current BASICs methodology assigns each violation to a distinct BASIC, and so assumes a multidimensional approach as there are multiple BASICs. Figure 4-2(b) depicts a multidimensional model with separate, uncorrelated dimensions of safety. In reality, it seems plausible that all safety dimensions are correlated to one another, a model that is represented graphically in Figure 4-2(c). The multidimensional models extend the model in above section to allow for a vector valued θ_i, and extend the model for the distribution of θ_i to be multivariate, such as multivariate normal with a diagonal (or unstructured) covariance matrix in the case of independent (correlated) safety dimensions.

A special version of the multidimensional IRT model is the bifactor model (Gibbons and Hedeker, 1992; Reise, Moore, and Haviland, 2010). In the bifactor model all of the violations load on a primary latent trait (overall safety), and then each item additionally loads on one of several secondary latent traits (e.g., unsafe driving). The bifactor model is shown in Figure 4-2(d). The secondary latent traits represent associations in the violations that are not fully captured by the primary overall latent trait. In this way, the overall latent trait reflects the general safety measured by an inspection, and each secondary latent trait indicates the unique contribution of, for example, unsafe driving over and above the general safety latent trait. Relative to a fully multidimensional IRT model, the secondary factors in the bifactor model represent variation attributable to the violations that are beyond the overall primary latent trait. Hence, an advantage

of the bifactor model over the fully multidimensional IRT is that it retains a single "summary" measure of safety, which can be used to prioritize interventions. The single measure could be more reliable than separate measures (dimensions) as it pools information from all violations.

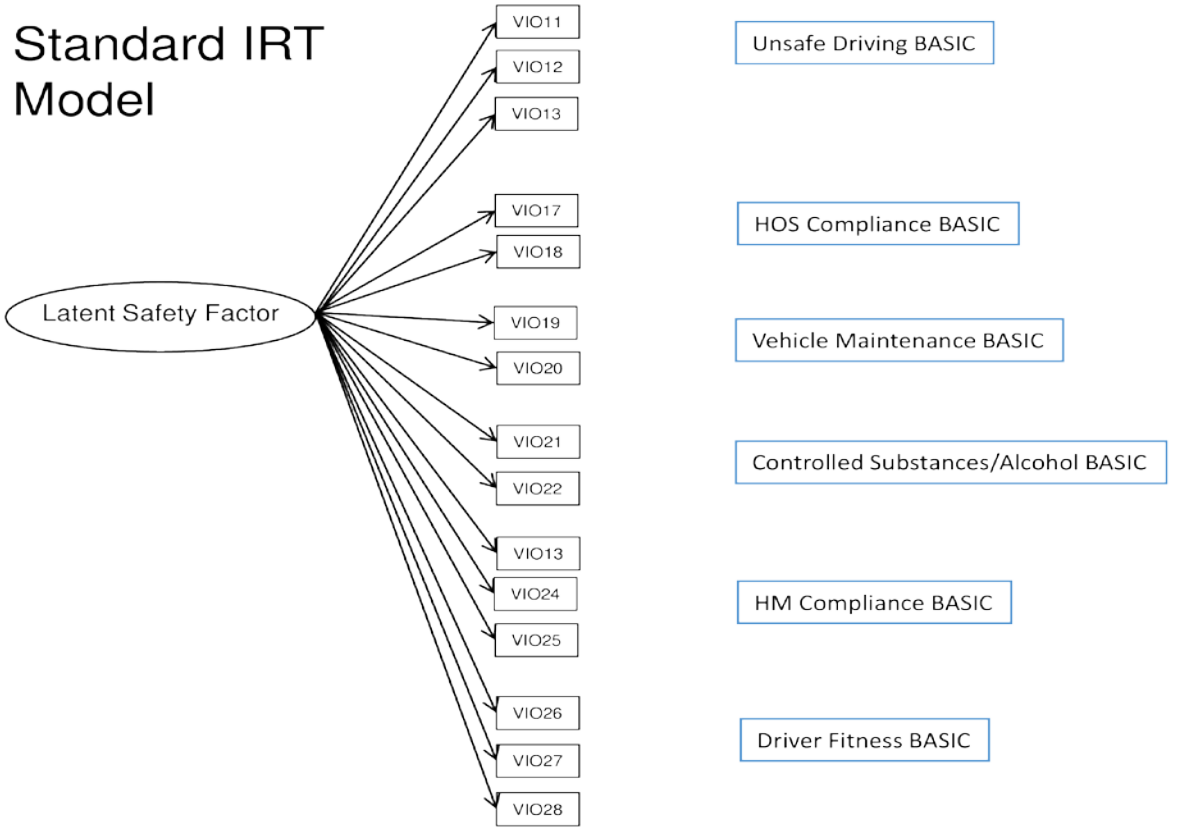


FIGURE 4-2 (a)
See note after Figure 4-2 (d)

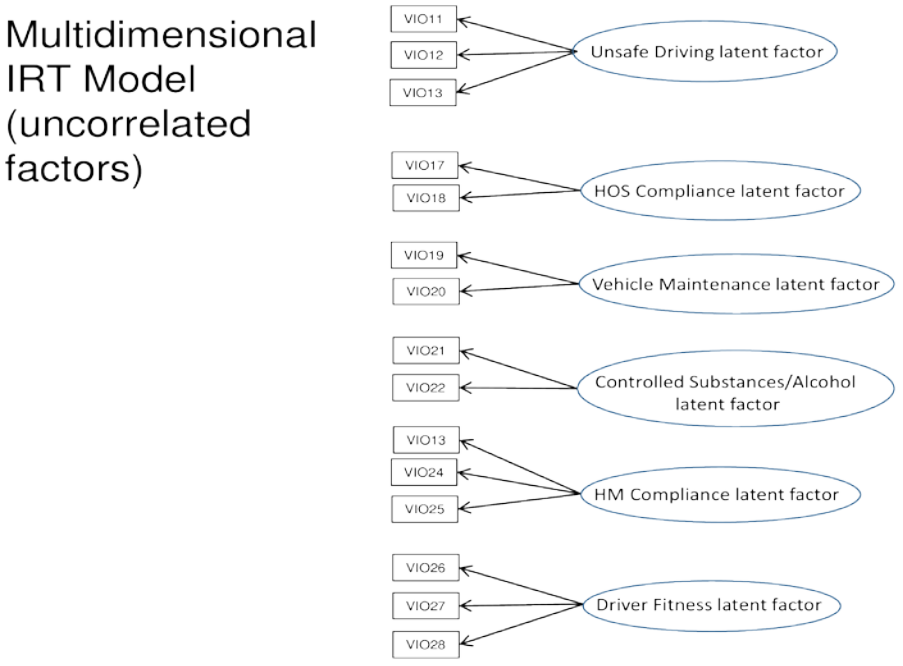


FIGURE 4-2 (b)

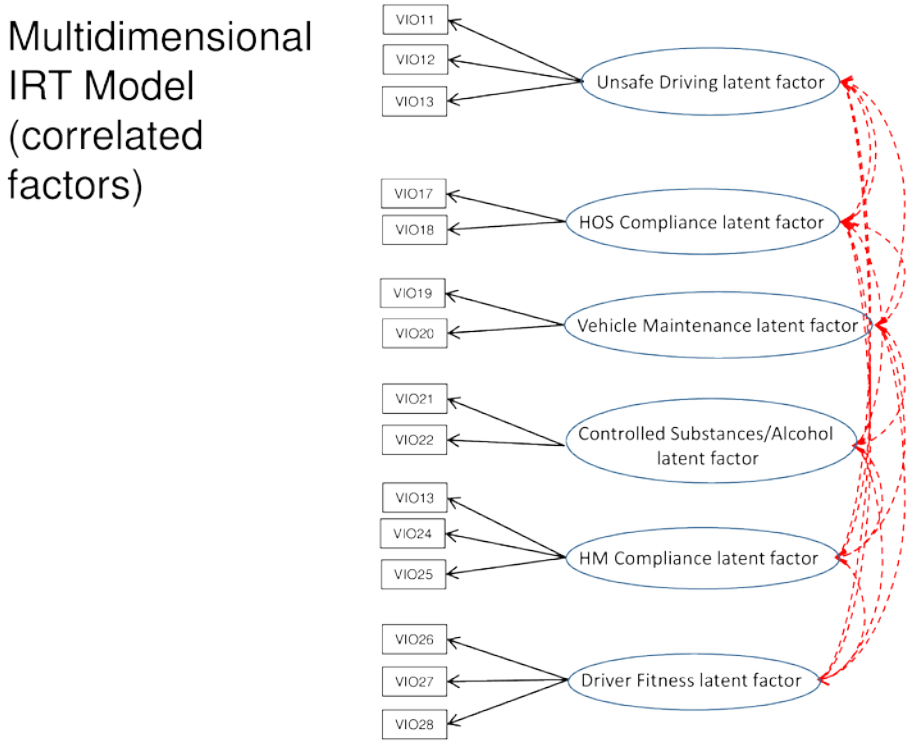


FIGURE 4-2 (c)
See note after Figure 4-2 (d)

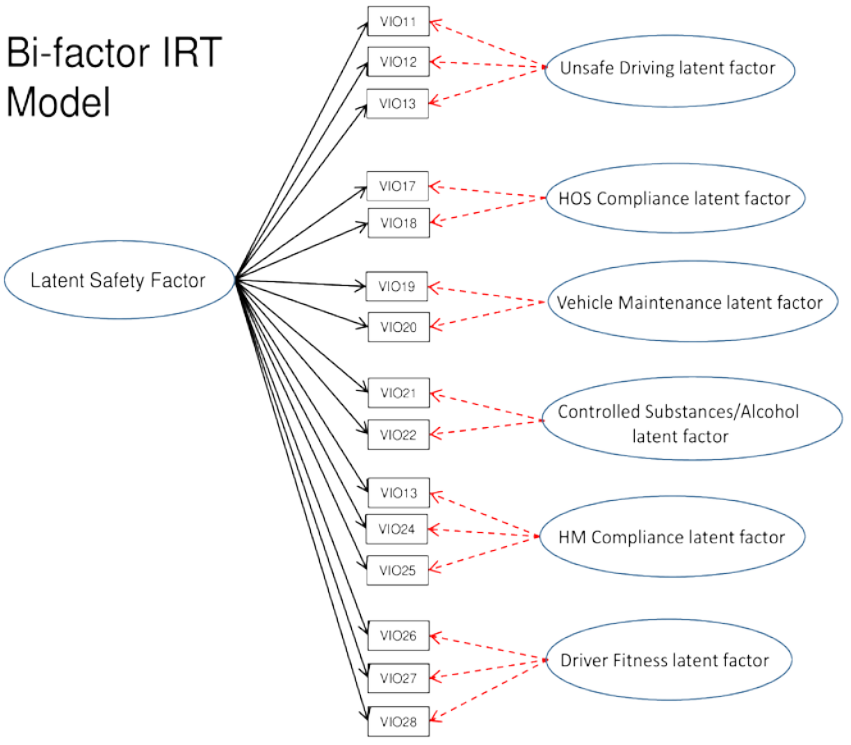


FIGURE 4-2 (d)

FIGURE 4-2 (a-d) Variations of IRT models.
SOURCE: Panel-generated.
NOTE: Panel (a): Standard IRT model where there is one latent (safety) score/factor. This one safety score impacts the probability of each violation and multiple violations are more (or less) likely to occur for the same carrier due to this relationship; Panel (b): Multidimensional IRT model with uncorrelated scores/factors. These multiple safety scores impact the probability of certain violations. For example, Unsafe Driving impacts violations 11-13 (not the actual numbers of the violations) and makes these violations more (or less) likely to occur for a given carrier. But, for example, the Unsafe Driving factor does not impact violation 21 sicne it does not diretly impact that violation and the Unsafe Driving factor is uncorrelated with the Uncontrolled Substance factor; Panel (c): Multidimensional IRT model with correlated scores/factors. This is similar to the model in Panel (b), but now the different factors are correlated and have some impact on all violations either directly (like with the Unsafe Driving factor with violation 11) or indirectly (like the Unsafe Driving factor on violation 21 through the correlation of that factor with the Uncontrolled Substance factor); Panel (d): Bifactor IRT model. Similar to the model in Panel (a), it indicates that safety score (primary factor) impacts the probability of each violation. However, there is also some additional association between distinct sets of violations that are impacted by the secondary factors. For example, unsafe driving (secondary factor) impacts violations 11-13 over and above the impact of safety score (primary factor) on these violations.

An important consideration for a multidimensional model is the assignment of violations to specific safety dimensions. Currently, the assignment of violations to the BASICs is done in a confirmatory manner. However, an exploratory approach could be used to determine how many latent traits are needed to adequately represent the violations, and to determine which violations are associated with which latent traits. These multiple latent traits would be analogous to the multiple BASICs, in that they represent different groupings of the violations. However, the way in which the items of a particular latent trait “hang together” could be determined empirically.

Expert opinion could be used to elicit informative priors on the number of dimensions of safety, and the particular items that inform on each of the dimensions.

Data on Prior Crashes

Currently, data on carriers' prior crashes is used to construct the Crash Indicator BASIC. We agree with this conceptualization of prior crashes as its own indicator, separate from other safety indicators, for several reasons. The causes behind individual crashes are multifactorial and depend on many unobserved factors such as weather and road conditions, and are thus not always under the direct control of the carrier. Thus, crashes represent a safety measure that is conceptually different from, for instance, maintenance violations. Further, crashes are relatively rare and, thus, would be more difficult to incorporate in the IRT modeling approach.

Model Assessment

It will be important to assess how well the proposed model fits the observed data. Given the proposal of using Bayesian inference for this model, a simple way to assess the absolute fit would be through the use of posterior predictive checks (Gelman et al., 2013). These checks measure and quantify how well the model captures features of the observed data as characterized by a test statistic or a discrepancy measure (Gelman, Meng, and Stern, 1996). Specific recommendations on such checks for IRT models can be found in Sinharay, Johnson, and Stern (2006).

In addition, it will be useful to carry out similar validation steps to those recommended in Chapter 3. This includes validation that the carriers selected by the IRT model as having high θ's are those that are operating unsafely. Further, it would also be necessary to validate that the carriers that are operating unsafely have frequent crashes.

SUMMARY ABOUT THE IRT MODEL FOR CARRIER SAFETY

In this chapter we outlined a probabilistic approach with which to obtain estimates of a carrier's safety. A key advantage of using a probabilistic approach over a deterministic model such as the current BASIC methodology is that probabilistic approaches yield measures of uncertainty in addition to estimated safety measures or rankings. The proposed model may need further refinements once it is fitted to the MCMIS data. For instance, the model as described assumes a violation in item *k* is independent from *k*' given the latent safety measures θ_i, an assumption that may have to be relaxed. The next chapter discusses in more detail the use of the safety measures for ranking carriers both relatively and absolutely, as well as a discussion of the use of various forms of stratification as opposed to a model-based adjustment.

Recommendation: FMCSA should develop the suggested IRT model over the next 2 years. If it is then demonstrated to perform well in identifying motor carriers for alerts, FMCSA should use it to replace SMS in a manner akin to the way SMS replaced SafeStat. Specifically, IRT models would have the following specific advantages over SMS:

1) Instead of severity weights being based on expert opinion or dated empirical information, the item discrimination parameters are estimated based on a combination of current observed data and expert opinion, and ultimately on data alone;

2) They can enhance the transparency of the evaluation system;
3) IRT models support the direct estimation of variability of scores and ranks;
4) They can account for the probability of being selected for inspection;
5) They can provide a basis with which to evaluate how data insufficiency could impact safety ratings of carriers;
6) They can provide a basis to more rigorously evaluate the structure of the current BASICs including which violations go into which BASIC;
7) They can provide for a natural way to examine the issue of further stratification;
8) They can provide for the possibility that safety is inherently multi-dimensional, which could inform as to how many BASICs are needed in the SMS model;
9) They can take account of time and thereby inform about the proper time weights in SMS;
10) They can allow for the addition of new safety measures as they become available, without having to start from scratch.
11) They can produce ranking ranges (by sampling from the posterior distribution of theta) to better understand overlap in the rankings (i.e., uncertainty);
12) They can adapt to changes in safety over time.

5
Extensions of and Implementation of the Item Response Theory Model for the Safety Management System

This chapter discusses issues related to the proposed model in Chapter 4, including how to "rank" carriers using an approach that incorporates both the safety score and its uncertainty. We also discuss issues brought up earlier in the report regarding absolute versus relative ranking of carriers and the relative merits of stratifying versus model-based adjustment for carrier characteristics (e.g., size and whether the carrier is a passenger carrier).

A NEW APPROACH FOR RELATIVE RANKING OF CARRIERS

A key drawback of the current approach to assign safety scores is that there is no quantification of uncertainty in the scores. That is, the deterministic algorithm used to produce Behavior Analysis and Safety Improvement Category (BASIC) measures does not take into account the sources of variation that are inherent to the data collection process (e.g., whether or not a carrier gets inspected on a particular trip), including the fact that some items may be found in violation even among the safest carriers (e.g., brake lights may have gone out by chance just prior to the inspection). Given the sources of variation, the ranking of a carrier's true safety measure, relative to other carriers, is not known for certain. The modeling strategy acknowledges variability by modeling the number of inspections and whether violations are present as random variables.

In developing a ranking system, it is important to acknowledge where a score lies in the distribution of scores across carriers (e.g., does the score lie in the upper 20 percent of bad scores) *and* how likely it is to truly fall in this tail. The model-based approach introduced in Chapter 4 estimates a score for each carrier and enables the calculation of uncertainty estimates about the numeric value of the score, and where the score lies relative to other carriers. The level of uncertainty depends in part on the number of inspections available for a carrier. Another factor contributing to uncertainty is the frequency with which items are found in violation. The current Safety Management System (SMS) partially recognizes uncertainty by establishing "data sufficiency" standards, with the consequence of not being able to calculate scores for a substantial portion of carriers with few inspections or violations present in the Motor Carrier Management Information System (MCMIS) data. Instead, we propose using model output to quantify the uncertainty, and include the uncertainty as part of the ranking approach.

Consider an example with 10 carriers and an assumption that the Federal Motor Carrier Safety Administration (FMCSA) takes action on the worst 20 percent of carriers. In this example, two carriers with the worst safety scores will be identified but a third will not. Now assume the score of the third-ranked carrier is very precise but the score of the second-ranked carrier is very imprecise. Figure 5-1 illustrates the estimated scores and uncertainty about the scores for these three hypothetical carriers. Uncertainty is important in ranking carriers; specifically, what the score is and how precisely is it estimated. We suggest computing an adjusted "score" that quantifies, via a probability, how likely a carrier is to truly be among the worst carriers. Such an adjusted score incorporates (implicitly) both the "estimated" score and its uncertainty.

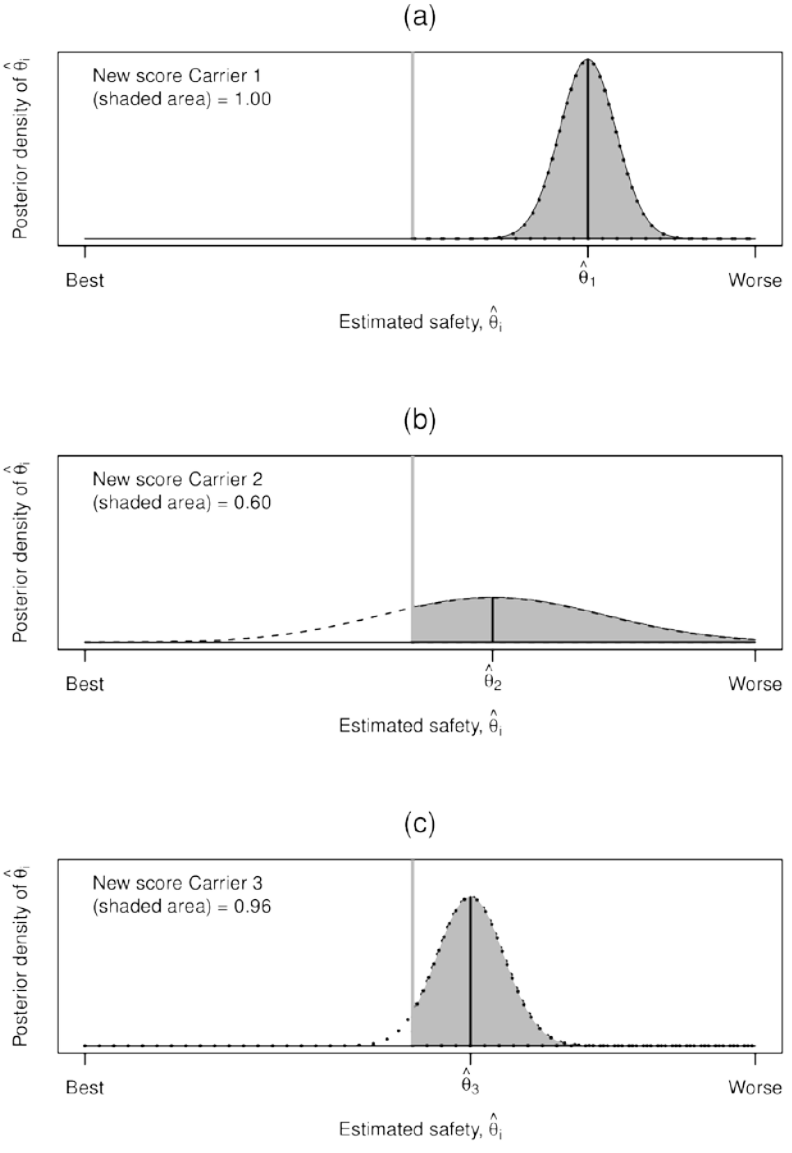


FIGURE 5-1 Posterior distributions of safety scores θ_i for three hypothetical carriers.

NOTE: Panel (a) corresponds to a hypothetical carrier (Carrier 1) with a bad safety score, θ_1 and not a lot of uncertainty. As such, the probability of being above the cut-off (denoted by the vertical line to the left of $\hat{\theta}_1$) is quite high (almost 1.0). So we are quite sure this is a bad carrier. Panel (b) corresponds to a hypothetical carrier (Carrier 2) with a slightly better score $\left(\hat{\theta}_2\right)$ than Carrier 1, but with considerably more uncertainty, which results in the probability of being above the cutoff of only 60 percent. Panel (c) corresponds to a hypothetical carrier with a safety score slightly below Carrier 2, but with more certainty about this score, resulting in a high probability of exceeding the cutoff (.96). Rather than rank carriers based on their posterior mean $\hat{\theta}_i$, the panel's proposed ranking takes into account both the mean safety score *and* the uncertainty around the estimated score. Carriers are ranked based on their probability of exceeding a cut-off, calculated as the area under the curve to the right of the gray vertical line. The top panel plots the distribution of the safety score for hypothetical Carrier 1, which has both a bad safety score

and a high probability of exceeding the cut-off. While Carrier 2 (middle panel) has a slightly worse $\hat{\theta}_i$ score than Carrier 3 (bottom panel), the precision for Carrier 3's score is greater such that the newly proposed ranking S_i is higher (worse) for Carrier 3 compared to 2.

SOURCE: Panel-generated.

In the context of our modeling approach in Chapter 4, the newly proposed adjusted score would be the following posterior probability:

$$S_i = P(\theta_i \in bottom\ q\ \%\ of\ scores | data) \qquad (5\text{-}1)$$

where as before, θ_i is a latent variable that measures the "safeness" of a carrier. The probabilities in Equation (5-1) would be ranked so that if a point estimate of θ_i is in the top q percent of worst scores, but there is a lot of uncertainty, it may not be in the top q percent of posterior probabilities as illustrated in the example above. These probabilities are also illustrated in Table 5-1. For simplicity, we focus on five carriers to determine the two worst (i.e., the 40th percentile). Table 5-1 shows that based on the estimated scores, $\hat{\theta}_i$, the two worst carriers are Carriers 1 and 2. However, there is a lot of uncertainty about the score of Carrier 2. Its probability of being in the bottom two is 60 percent, but the carrier with the slightly better score (Carrier 3) has a 96 percent probability of being one of the two worst, as shown in Figures 5-1(b) and 5-1(c). This happens because there is more uncertainty about the true safety score of Carrier 2 and less uncertainty about Carrier 3.

TABLE 5-1 Relative Scores based on Point Estimates versus a Combination of Point Estimate and Uncertainty for Five Hypothetical Carriers

Carrier (i)	$\hat{\theta}_i$ (rank)	$P(\theta_i \in bottom\ 40\%\ of\ scores)$ (rank)
1	1.5 (1)	0.99 (1)
2	.65 (2)	0.60 (3)
3	.45 (3)	0.96 (2)
4	.20 (4)	.23 (4)
5	-.10 (5)	.10 (5)

SOURCE: Panel-generated.

We recommend the adjusted score be used as it incorporates both the point estimate and its uncertainty into a single value. (Instead of computing the tail area of the posterior distribution of a carrier's measure, one could also sample from the posterior distribution many times, for each posterior sample, rank all the carriers, and then report, for example, a 95% posterior interval for each carrier's rank).

ABSOLUTE RANKING

As discussed in Chapter 2, several stakeholders expressed concern to the panel about relative rankings. For example, if the industry has large improvements and the same percentiles are used for "action," it will result in acting upon safer carriers. If the goal is to continuously improve carrier safety, given fixed resources, then the relative approach can work as intended.

On the other hand, if overall carrier safety in the industry improves, there might be interest in allocating resources to other safety initiatives. In this case, an absolute approach that would send fewer letters and initiate fewer investigations might be desired.

An absolute approach would require "connecting" the θ_i's (which are by construction a relative score centered around zero) to a future outcome measure. We suggest some possible ways to do this next.

The safety scores, θ_i are computed over 2-year windows. We can regress the number of carrier crashes in the subsequent year on θ_i using a Poisson regression, similar to that used for the number of inspections in Chapter 4, but with θ_i entered as a regressor. So we assume the rate parameter of the Poisson distribution, λ_i^c, varies smoothly with the safety score θ_i. In particular:

$$C_i \sim Poisson\left(E_i \lambda_i^c\right)$$
$$\log \lambda_i^c = f\left(\theta_i; \xi\right)$$

where C_i is the number of crashes in the subsequent year, E_i is the corresponding vehicle miles traveled (VMT), and λ_i^c is the crash rate, with $f(\cdot)$ a smooth (monotone) function of θ_i parameterized by ξ. An absolute cutoff for θ could be chosen based on the risk of a future crash per 1,000,000 VMTs, being above a certain rate determined by FMCSA. Other possible future outcome measures might include subsequent year inspections. Computation of the absolute rate for a carrier, R_i would be as follows:

$$R_i = \int\int \exp\left(f\left(\theta_i; \xi\right)\right) dF\left(\xi \mid sdata\right) dF\left(\theta_i \mid data\right) \qquad (5\text{-}2)$$

where *sdata* is the subsequent year data and *data* is the 2-year window data. This expression takes the smoothed rate, $\exp\left(f\left(\theta_i, \xi\right)\right)$, which is parametrized by ξ, averages over the uncertainty associated with the safety score (via the integration over $F\left(\theta_i \mid data\right)$ and over the parameters of the smoothing function, ξ, via the integration over $F\left(\xi \mid sdata\right)$. Equation (5-2) produces an updated crash rate for a carrier given that carrier's estimated latent risk of a crash and its vehicle miles traveled, which can then serve as an absolute metric on which decisions for interventions can be based.

While this has the look of a model of future crash risk, we are not proposing this as a model of crash risk, but as a construct that is dependent on theta, the latent safety parameter, that is an absolute measure of interest to use for decisions on interventions. In fact, we can regress any outcome of interest, such as number of crashes in the subsequent year, on θ_i using Poisson regression. The point here is that the outcome of interest is regressed on θ_i.

We could also compute posterior probabilities, P_i, similar to those in Equation (5-1), as shown in Equation (5-3). Here the probability is in terms of the rate for carrier *i* being above a cut-off and the carrier is "flagged" only if its posterior probability itself exceeds some value, for example, 0.80.

$$P_i = \int\int P\left(\exp\left(f\left(\theta_i, \xi\right)\right) > cutoff\right) dF\left(\xi \mid sdata\right) dF\left(\theta_i \mid data\right) \qquad (5\text{-}3)$$

Equation (5-3) averages over uncertainty in the same way as Equation (5-2). By using an absolute approach, the number of carriers requiring action could be reduced, carriers would be "rewarded" for making safety improvements, and resources could potentially be allocated to improve safety in other ways.

It would be worthwhile to also explore a hybrid measure that combines the proposed relative and absolute measures.

STRATIFICATION VERSUS MODEL-BASED ADJUSTMENT

Another important issue is being sure to compare sufficiently similar carriers and not to inappropriately compare different types of carriers.

One way to do this is to stratify (e.g., on the number of power units); however, stratifying on the outcome (e.g., the number of violations) should be avoided. Clearly, stratifying based on the size of the carrier or on the type of carrier (e.g., passenger carrier or not) has merit. However, there are problems with stratification, as outlined below. An alternative would be to adjust for carrier characteristics that influence the inspection rate (such as size or whether it is a passenger carrier) within the model. Referring to Chapter 4, these characteristics could be entered into the log-linear models estimating inspection rate (λ_{ij}) described in Equations (4-2) and (4-7) and their analogs.

Below we provide a list of potential issues and which approach is preferred, we discuss stratification (S) or model-based adjustment (M) for each issue. In particular, $M = S$ corresponds to the two approaches being equivalent, $M >$ S corresponds to the model-based approach preferred, and $M < S$ corresponds to stratification preferred:

- (M=S): only comparing similar carriers
- (M<S): without explicit stratification, it will be very difficult for small carriers to be classified as "bad" given their large uncertainty relative to large carriers (given the approach advocated above to "rank" carriers)
- (M>S): "discontinuities" on the strata boundaries. For example, adding one power unit can move a carrier into a different strata
- (M>S): the "cutpoints" for the strata are often, to some extent, somewhat arbitrary
- (M>S): given that the model-based approach uses all the carrier data simultaneously, safety scores will be estimated more precisely (i.e., less uncertainty)

The above issues that differentiate stratification from model-based adjustment need to be considered in deciding whether or not to stratify. They are summarized in Table 5-2. We also note that the issues in stratification of arbitrary cutpoints and score discontinuities might be lessened to some extent by using dynamic strata (see Federal Motor Carrier Safety Administration, 2014).

TABLE 5-2 Considerations in Deciding between Stratification and Regression Adjustment

Issue	Preferred Approach
Only comparing similar carriers	Equivalent
Identifying small carriers for intervention	Stratification
Avoiding discontinuities in scores	Model-based
Avoiding arbitrary cutpoints	Model-based
Increasing precision	Model-based

SOURCE: Panel generated.

IMPLEMENTATION ISSUES

Clearly, the proposed Bayesian IRT model, which involves use of 20–30 million observations and hundreds of variables to estimate hundreds of model parameters is something that requires very specific expertise, usually found in academic statisticians who carry out research on these specific models. The sparsity of data and other aspects of the problem are likely to raise some computational complexities that would require software development. The model development costs will therefore involve contracting with a small group expert in these models. However, once developed, FMCSA staff would be very capable of maintaining the model, including refitting parameter estimates, conducting model validation exercises, and incorporating improved inputs. Finally, given that the IRT model, like SMS, is sensitive to outlying values in MCMIS, FMCSA should consider institution of edit routines to identify discrepant submissions, and imputation procedures to fill in for input values that fail the edits

6
Data Availability and Quality

DESCRIPTION OF DATA IN THE MCMIS

The Behavioral Analysis and Safety Improvement Category (BASIC) measures and percentile ranks are calculated using data on crashes, inspections, and violations collected and reported by the states and federal inspectors to the Federal Motor Carrier Safety Administration (FMCSA). They are combined with other information provided directly to FCMSA through the Unified Registration System (URS) and a form called the MCS-150, which carriers complete to register with the FMCSA. All of this information goes into the Motor Carrier Management Information System (MCMIS).

The MCMIS is made up of four component files, with information to enable linking each of the components. Extracts from MCMIS files can be used to create records containing each motor carrier's demographic and safety characteristics. These four component files are:

1. *The census file:* Based on information collected from the URS and form MCS-150, this file includes, for each carrier: its estimated vehicle miles traveled; the number of trucks and buses owned, leased, and trip leased; power unit type (combination/motorcoach or straight trucks/other vehicle type) and some information on trailers; and the number of drivers employed and leased. The carriers are required to provide this information when a carrier is first registered as part of the URS, and must update the information every 2 years. Each motor carrier has a unique Department of Transportation (DOT) number assigned to it, which is used to link the carrier file to other MCMIS files. The Census file also includes noncarriers, such as brokers, shippers, and freight forwarders.
2. *The inspection file:* The inspection file contains information on vehicle inspections conducted by state and federal inspectors. The information available includes the inspection date and location, and level of each inspection. (See Glossary for the six inspections levels.) Indicator variables also signify whether the inspection was the result of a crash, and whether a vehicle is transporting hazardous material. Each separate inspection is assigned a unique identification number to facilitate linkage.
3. *The violation file (a component of the inspection file):* The violation file contains details of violations identified during inspections. Each violation is assigned a unique identification number. This file provides the following information for each violation: the DOT number for the vehicle, the violation code, and whether the violation was an out-of-service violation.
4. *The crash file:* The crash file derives from police accident reports (PARs) that the states are responsible for collecting and reporting to FMCSA. The crash file contains data on commercial motor vehicles with a gross vehicle weight rating (GVSR) or gross combination weight rating (GCWR) greater than 10,000 pounds designed to transport nine or more people, including the driver. This includes a subset of information from PARS for all trucks and buses involved in a crash where there was either: (i) a fatality, (ii) a bodily injury to a person who immediately received medical treatment away from the scene, or (iii) at least one vehicle was towed away due to disabling damage. This file

contains information on the time and date of the crash, the types of vehicles involved in the incident, counts of injuries and fatalities, whether vehicles were towed away, and whether a hazardous material was released during the crash, as well as a limited amount of other descriptive information.

It must be stressed that MCMIS is a database that draws on data supplied by the states, federal investigators, inspectors, and carriers to monitor motor carrier safety in the United States. MCMIS supports analysis of time, geography, and other trends and patterns involving large truck and bus crashes; roadside inspections disaggregated by motor carrier types; and violations. It is important to keep in mind the great deal of churn in the commercial motor vehicle (CMV) industry, which leads to a substantial percentage of newly registered carriers each year and existing carriers going dormant, commonly without notice. This means that the population of active carriers varies from one time period to the next, and therefore the current snapshot of data from MCMIS only pertains to the set of active motor carriers for this particular time period. Assessments of change over time almost necessarily involve somewhat different groups of carriers. We are basing our analysis of the MCMIS database on the November 27, 2015, version of MCMIS, which was provided to the panel by FMCSA and the Volpe National Transportation Systems Center.

Table 6-1 provides the distribution of active[1] motor carriers by number of power units (vehicles with engines) and equipment type.[2] As can be seen, the majority of motor carriers are single- truck enterprises. More than 64 percent of motor carriers operate three power units or fewer, and greater than 90 percent have 10 or fewer power units.

TABLE 6-1 Distribution of Active Motor Carriers by Power Units and Equipment Type[3]

Power Units	Number	Percentage Distribution	**Equipment Type**	Number	Percentage Distribution
1	251,456	46.3	**Straight Trucks**	279,800	51.4
2 to 3	147,641	27.2	**Tractor-Trailers**	316,768	58.2
4 to 5	50,382	9.3	**Motor coaches**	5,724	1.1
6 to 10	43,705	8.0	**Hazmat**	15,747	2.9
11 to 15	16,422	3.0	**School Buses**	3,320	0.6
16 to 20	9,037	1.7	**Vans**	10,674	2.0
Greater than 20	24,715	4.5	**Limousines**	2,135	0.4

NOTE: Table based on November 28, 2013, to November 27, 2015, version of MCMIS.

[1]Active motor carriers are those that have had an inspection, or a crash, or made insurance payments, registered as a carrier, or updated their information on form MCS-150, within the past 3 years.

[2]Equipment type is not a mutually exclusive categorization as a motor carrier can have more than one type of equipment.

[3]This table is based on the November 27, 2015, version of MCMIS.

PREPARING MCMIS FOR USE IN SMS

Prior to calculating SMS measures, various filters are used on MCMIS to identify the subset of carriers to which SMS is applied, and to identify the subsets of inspections, violations, and crashes used. Table 6-2 describes the steps used to filter these four files into the subset on which the BASIC measures are based.

First, as mentioned, the initial MCMIS files have a substantial number of carriers that have had no indication of active operations for a long while. These carriers have not had any inspections or crashes, made insurance filings, or updated their information on the MCS-150 for 3 years. MCMIS is not purged of such information because it is an administrative database, which is intended to contain information on all carriers that have registered with FMCSA. Second, BASIC measures are only calculated for motor carriers that have had enough activity of inspections, violations, or crashes during the last 2 years, which, as seen in Table 6-2, is a more restricted filtering. After intrastate nonhazardous materials carriers are removed, the remaining carriers for that particular data snapshot of MCMIS make up the set of active carriers to which SMS is applied. In addition, different data sufficiency standards have to be met before a carrier has an SMS score for a given BASIC, and the set of carriers that have sufficient data for one BASIC is not necessarily the same set of carriers that has sufficient data for another BASIC. Inspections are filtered as follows: FMCSA includes them if they occurred during the last 2 years, if there is a DOT number for the carrier involved, and if the DOT number matches to an active carrier. This reduces the number of inspections by about two-thirds. Roughly speaking, the data sufficiency standards reduce the number of carriers for which BASICs are computed to about 190,000.

Filters are also applied to violation data. FMCSA excludes violations caused due to involvement in a crash (such as broken headlights from a crash) or assigned to another entity such as a shipper or an intermodal equipment provider (such as moving a freight container to a ship). Further, FMCSA excludes violations that did not occur during the past 2 years, were not associated with a BASIC, were not linked to an active carrier, or were duplicate, which reduces the number of violations used in SMS by about 80 percent (mainly due to time duration).

TABLE 6-2 Filters Applied to MCMIS Dataset BASIC Percentiles for December 2015

	Number of Observations Removed	Number of Observations Remaining
All carriers in census (MCMIS)		**2,708,148**
Remove carriers: 1. Which are inactive or 2. Don't have an address or 3. Are not classified as a "carrier"	1,067,935	1,640,213
Remove blank rows in the file	19	1,640,194
Remove carriers without recent activity	760,859	879,335
Remove intrastate carriers	332,994	546,341
Carriers included in SMS		**546,341**
All inspections (MCMIS)		**20,708,328**
Include only inspections which occurred in the past 24 months	13,969,094	6,739,234
Include only inspections that have a DOT number listed	541,298	6,197,936
Include only inspections matched to a carrier in the final carrier file	141,144	6,056,792
Inspections included in SMS		**6,056,792**
All Violations (MCMIS)		**38,503,050**
Violations wiped out by violation review process	2,277	38,500,773
Violations that were caused by a crash	206,753	38,294,020
Violations that were assigned to an Intermodal Equipment Provider (IEP)	54,488	38,239,532
Violations that were not associated with an inspection which 1. occurred in the past 24 months and 2. was associated with an active carrier.	28,116,904	10,122,628
Violations not associated with any BASIC	1,546,949	8,575,679
Violations that were assigned to a shipper	302	8,575,377
Violations in an inspection which was not relevant to the BASIC of that violation	103,431	8,471,946
Violations from a postcrash inspection, and "Unsure" if violation was postcrash	4,173	8,467,773
Duplicate violations	1,112,558	7,355,215
Violations included in SMS		**7,355,215**
All Crashes (MCMIS)		**3,022,849**
Include only crashes that occurred in the past 24 months	2,710,259	312,590
Include only crashes that have a DOT Number listed	75,399	237,191
Include only crashes matched to an active carrier	10,432	226,759
Remove duplicate crashes for the	896	225,863

same carrier	
Crashes included in SMS	**225,863**
Carriers assigned at least one BASIC measure	**487,885**

NOTE: Table based on November 28, 2013, to November 27, 2015, version of MCMIS.

MCMIS DATA QUALITY

Information on Exposure

The number of crashes motor carriers have is highly correlated with the amount of the carriers' exposure, where exposure is a measure of all trips taken by a carrier's vehicles. Exposure is certainly a function of how far a carrier's vehicles have traveled, since a carrier whose trucks or buses travel twice as far as the trucks and buses for another carrier would be expected to have twice as many crashes, everything else being equal. Exposure, which is essential to calculating crash risk, should also be sensitive to the measurable contributing factors in crashes such as the road type and surface conditions, visibility, day or night, congestion, and others, which if available would also be taken into consideration. However, these factors are not collected in MCMIS, in part because they would be very difficult to collect, and in some cases, are not feasible to collect reliably. Were they collected, especially on a trip-by-trip basis, it would be possible to calculate exposure that was specific to the routes, times, and seasons driven by vehicles under the control of each motor carrier, and crash risk could be more accurately estimated for each carrier. To do so would require much greater detail in the data reported by motor carriers than what is available in MCMIS, and therefore, making use of this expanded view of exposure is currently not feasible.

Limiting exposure to an assessment of the length of all trips, in MCMIS, the two most commonly used variables for exposure by FMCSA in SMS are vehicle miles traveled (VMT) and average power units (APU) times a utilization factor. VMT is the total number of miles traveled by all vehicles managed by a motor carrier, while the number of power units is the number of non-trailer vehicles managed by the carrier. APU is an average of the number of power units a carrier has currently, 6 months ago, and 18 months ago (if updated by the carrier during the 2-year window).

The Utilization factor is an adjustment for carriers with high (above average) use of their trucks, measured by VMT per PU, which attempts to prevent such operations being overly identified in the Unsafe Driver and Crash Indicator BASICs. Utilization factor is defined as follows. For combination trucks, for VMT per PU less than 80,000 (e.g., carriers at or below average utilization in segment) or greater than 200,000, no correction is applied and the utilization factor is set to 1. For VMT per PU between 80,000 and 160,000, the utilization factor increases linearly from 1 to 1.6. Between 160,000 and 200,000 VMT, the utilization factor, instead of remaining linear, is truncated at 1.6, the value for 160,000. In a sense, VMTs greater than 200,000 are not considered trustworthy and so the utilization factor is set to 1. VMT per PU values between 160,000 and 200,000 are treated as if they were equal to 160,000, and values between 80,000 and 160,000 are treated as useful indicators of the VMT for a carrier, and so utilization factor is a linear function of VMT when it is between those values. Similarly, for straight trucks, the utilization factor for carriers with VMT less than 20,000 (e.g., carriers at or

below average utilization in segment) and greater than 200,000 is 1; between 20,000 and 60,000 the utilization factor is a linear function from 1 to 3; and between 60,000 and 200,000, utilization factor, instead of remaining linear, is truncated at 3. So, again in a sense, values greater than 200,000 are not trusted.

Total VMT and number of power units are reported by motor carriers when they fill out the Form MCS-150 (or the URS), and this information is required to be updated at least biennially. Figure 6-1 shows that the distribution of VMT for recently active carriers is skewed with a long right tail. Also, 7.8 percent of carriers fail to report VMT during the 2-year period and 5.1 percent report zero. While zero can be a valid score if the carrier is a leasing firm and does not operate trucks itself, it would be useful to provide such carriers with a way of indicating this information so they can be separately treated by SMS. The distribution of APU is also highly skewed, but only 0.5 percent of carriers report zero, and only 0.05 percent fail to report (see Figure 6-2).

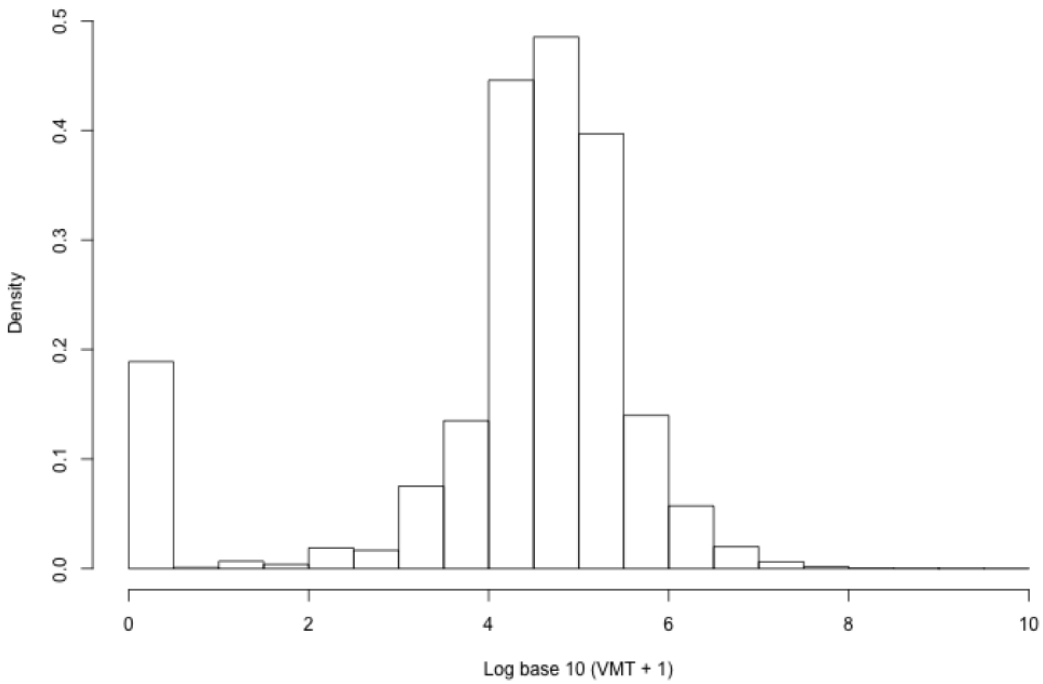


FIGURE 6-1 Logarithmic distribution of vehicle miles traveled.
NOTE: Figure based on November 28, 2013, to November 27, 2015, version of MCMIS.

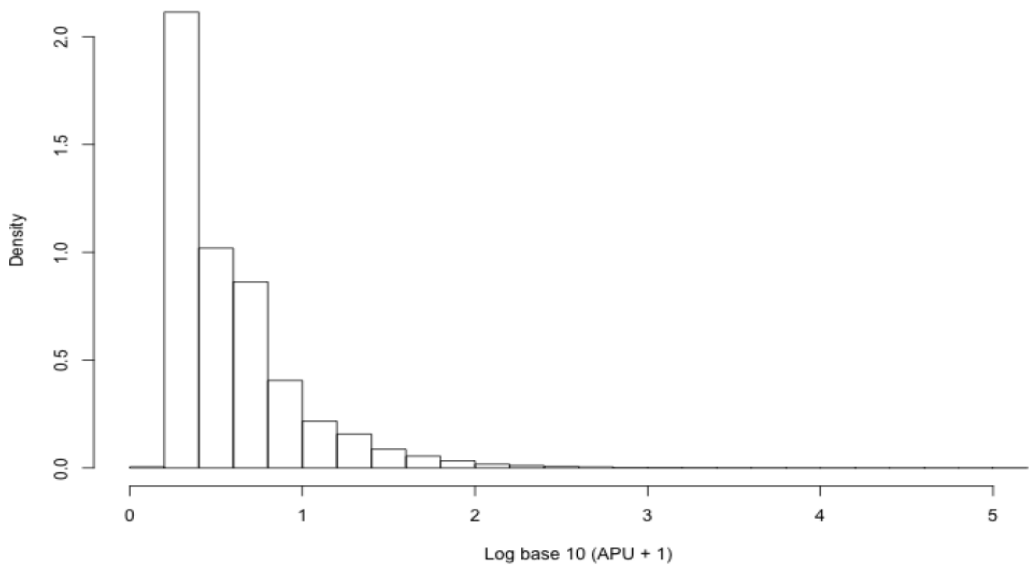


FIGURE 6-2 Logarithmic distribution of average power units.
NOTE: Figure based on November 28, 2013, to November 27, 2015, version of MCMIS

The timeliness of VMT data is a concern, as seen in Figure 6-3, which shows the frequency distribution of VMT by its reported year. While a majority of motor carriers do update their exposure information, a large number of carriers fail to do so.

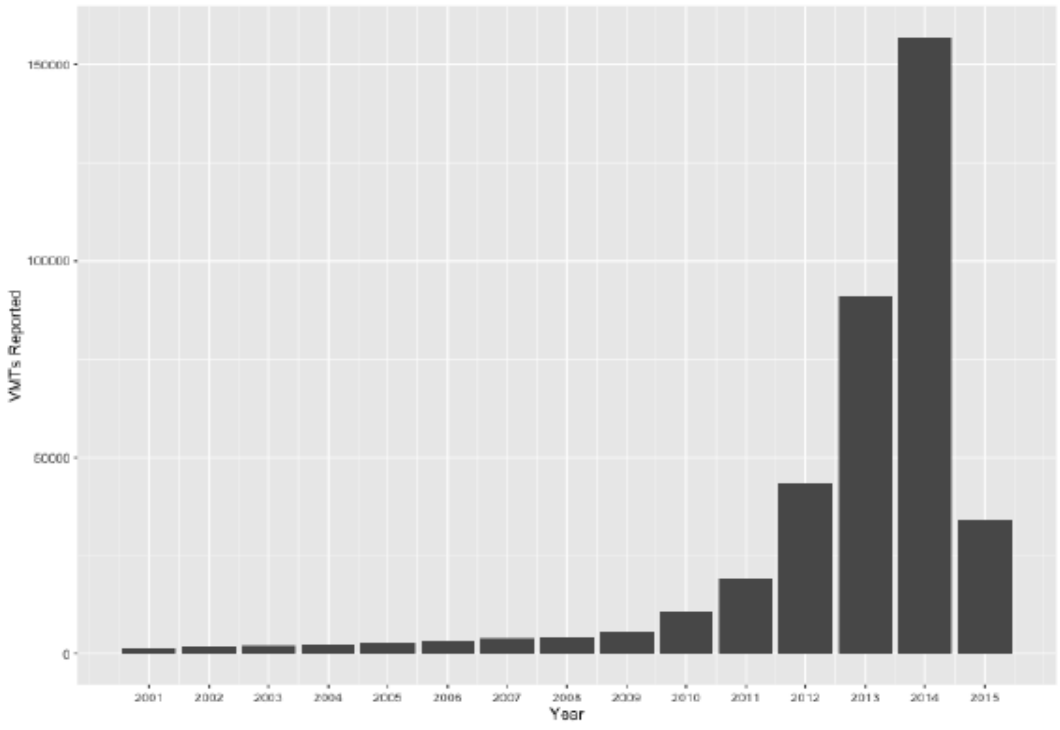


FIGURE 6-3 Most recent VMT report year.
NOTE: Figure based on November 28, 2013, to November 27, 2015, version of MCMIS.

For two of the BASIC measures, Crash Indicator and Unsafe Driving, the denominator is a measure of exposure of a motor carrier. If the denominator is not of high quality, those two BASIC measures will be a poor measure of a carrier's safety performance. The lack of updated and high-quality exposure information can clearly bias the results.

We carried out an analysis of the differences between VMT from the MCS-150 and the VMT from investigations (MCS-151), which would likely be of much higher quality. There were substantial differences between the two, but this analysis was deficient since we could not tell the precise time period for each entry. Therefore, we mention it is suggestive, but not definitive, that the MCS-150 information is not of high quality.

Estimates of Prevalence of Safety Inspections

Between November 27, 2013, and November 26, 2015, roadside inspections were conducted on 332,271 motor carriers, out of the total of 546,341 active carriers during this time period. So, only about 61 percent of active carriers were inspected during the 2-year period. Acknowledging the infrequency of crashes, this means that MCMIS has no information on which to base an assessment of the safety performance for about 40 percent of active carriers.

We point out that 51 percent of inspections conducted were clean, meaning that there were no violations recorded (which given the earlier mention of failure to report clean inspections, is likely an underestimate). Table 6-3 demonstrates considerable variability in the frequency of clean inspections across inspection levels. (See the Glossary for a description of the inspection levels.)

TABLE 6-3 Distribution of Inspections and Clean Inspection by Inspection Level

Inspection Level	Volume	Frequency	Percentage of Clean Inspections (%)
Full Inspection (=1)	1,680,683	.31	38
Walk Around Inspection (=2)	1,829,943	.33	43
Driver Only Inspection (=3)	1,845,840	.34	70
Special Inspection (=4)	NA	NA	NA
Terminal Inspection (=5)	102,629	.02	73
Radioactive Inspection (=6)	754	.00	97

NOTE: Figure is based on November 27, 2013, to November 26, 2015, version of MCMIS. Also, level IV inspections are not used in SMS.

Distribution of Violations

Out of the total of 899 possible violations, 790 of them occurred during the period November 27, 2013, to November 26, 2015. Some violations were recorded more frequently than others. Table 6-4 shows the top 20 most frequently recorded violations.

TABLE 6-4 Top 20 Violations from Level 1 (Full) Inspections

	Violation	BASIC	Count	Percent
1	Inoperative required lamps	Vehicle Maintenance	285456	8.4
2	Log violation general/form and manner	Fatigued	117301	3.5
3	Inspection/repair and maintenance parts and accessories	Vehicle Maintenance	108173	3.2
4	Non-English-speaking driver	Fitness	107215	3.2
5	Oil and/or grease leak	Vehicle Maintenance	106331	3.1
6	Tire other tread depth less than 2/32 of inch	Vehicle Maintenance	102121	3
7	Clamp/RotoChamber type brakes out of adjustment	Vehicle Maintenance	103567	3
8	No/discharged/unsecured fire extinguisher	Vehicle Maintenance	91579	2.7
9	Driver's record of duty status not current	Fatigued	88871	2.6
10	Failing to secure brake hose/tubing against mechanical damag ...	Vehicle Maintenance	83265	2.5
11	Automatic brake adjuster CMV manufactured on or after 10/20/ ...	Vehicle Maintenance	74089	2.2
12	Inoperative turn signal	Vehicle Maintenance	71072	2.1
13	Driving beyond 8 hour limit since the end of the last off du ...	Fatigued	68566	2
14	No/defective lighting devices/reflective devices/projected	Vehicle Maintenance	68169	2
15	Operating a CMV without periodic inspection	Vehicle Maintenance	64477	1.9
16	State/Local Laws Speeding 6-10 miles per hour over the speed ...	Unsafe Driving	62146	1.8
17	Windshield wipers inoperative/defective	Vehicle Maintenance	58913	1.7
18	ABS malfunctioning lamps towed CMV manufactured on or after ...	Vehicle Maintenance	50710	1.5
19	Inoperative head lamps	Vehicle Maintenance	44977	1.3
20	Brakes general	Vehicle Maintenance	44512	1.3

NOTE: Percent is percent of all violations from level 1 inspections (a total of 1,680,683 inspections).
SOURCE: Panel computations.

As noted, SMS groups the 899 violations into six BASIC categories. Looking at the distribution of violations by BASIC category shows that Vehicle Maintenance violations are the ones most frequently cited. Figure 6-4 provides the frequency distribution of violations by BASIC categories.

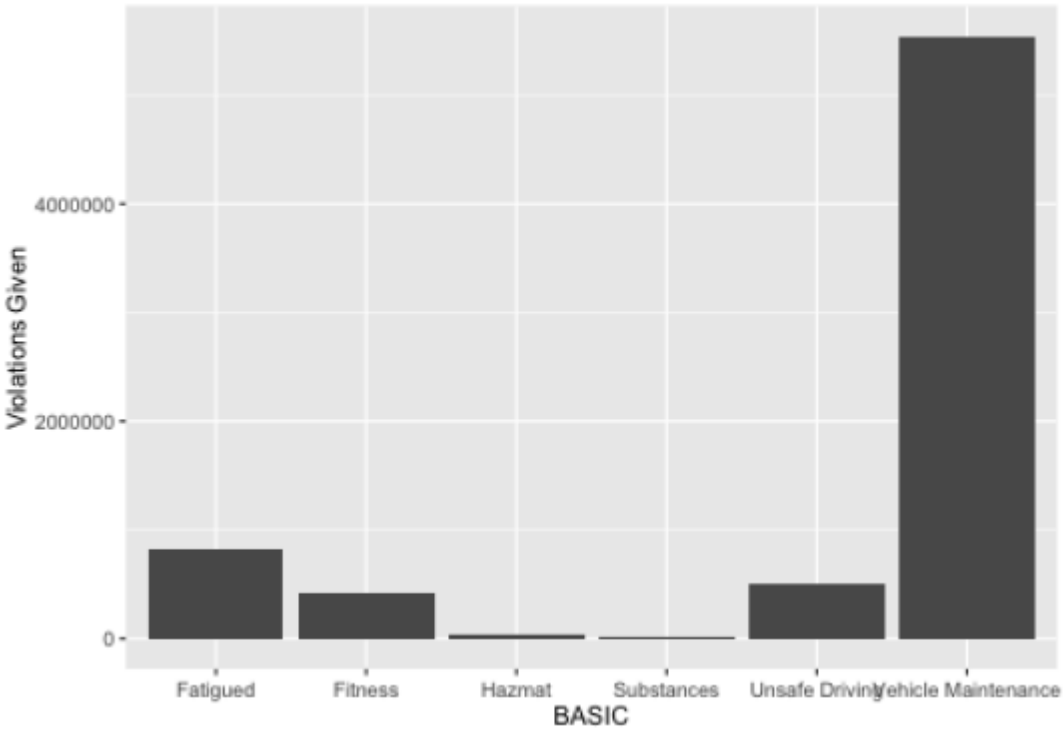


FIGURE 6-4 Frequency distribution of violations by BASIC categories.
NOTE: Figure based on November 28, 2013, to November 27, 2015, version of MCMIS.

NEEDED IMPROVEMENTS TO MCMIS TO SUPPORT SMS

It is important to reiterate that the collection and hence the quality of MCMIS data is not unilaterally controlled by FMCSA. However, FMCSA does have some influence over this data collection, and in fact, its efforts have resulted in important improvements to the quality and timeliness of the data collected over the past several years. For example, FMCSA has a continuing program to measure certain aspects of data quality and publish the results (e.g., through a ranking of states by reporting timeliness, completeness, and other factors). This includes the State Safety Data Quality (SSDQ) program; the Data Qs program, which allows carriers to request and track requested corrections; and the Safety Data Improvement Program (SaDIP).

The goal of this section of our report is to encourage FMCSA to continue in its efforts, but also to provide additional focus as to which elements are in most need of improved quality, knowing any effort to improve data quality will necessarily involve collaboration with the states and other organizations, such as CVSA. The following are, in the panel's view, the MCMIS variables that need to be collected with higher quality in order for MCMIS to better support SMS.

Better Measures of Exposure

The numerator of SMS's Crash Indicator BASIC score is the time and severity weighted number of crashes a carrier has experienced in the previous 2-year period. This numerator must be normalized through use of a denominator that is an estimate of the exposure to crashes to provide a statistic that can be compared across carriers. It is typical to use the number of VMT as an exposure estimate, which would result in a statistic of time and severity weighted crashes per mile. As its measure of exposure, FMCSA uses APU multiplied times a utilization factor, which is a function of VMT that tries to reduce the impact of discrepant values (see Appendix B for details). The Unsafe Driving BASIC uses the same denominator, since VMT is a reasonable way to normalize the numerator of the number of severity and time weighted violations for that BASIC.

Unfortunately, because the current estimates of VMT are not considered to be of high quality, the Crash Indicator and the Unsafe Driving BASIC measures could be substantially flawed depending on the specific quality of VMT and of APU reported by individual carriers. A 10 percent error in the denominator would translate to a 10 percent error for the score, which in turn could have a large impact on a carrier's percentile rank. Thus there may even be incentives for self-reporting incorrect data.

Therefore, it is important that FMCSA has access to an estimate of higher-quality VMT estimates than the existing estimates. This might mean exerting greater pressure on carriers to report APU and VMT more often, maybe annually, and possibly by fining carriers that either fail to respond or are found to intentionally misrespond (possibly as a result of an investigation). With the advent of electronic log books, it is at least worth exploring the possibility of automated reporting of mileage without relying on motor carrier self-reports, although this idea would face considerable pushback from the industry despite its potential for improving the accuracy of safety ratings.

Further, if possible, data on two other important factors that are known to impact crash risk should be collected and incorporated into the exposure variable (which could be done using statistical models). We believe the great majority of carriers can disaggregate their VMT by month and state, and this could be collected through an expansion of the data request on the MCS-150. By acquiring this information, the result could be a better measure of exposure that would reflect the fact that different states in different months are more or less safe to drive in. Again, this data should be obtainable electronically; however, there would be considerable pushback from the industry against sharing detailed trip-level data with the government.

Better Data on Crashes

Research by Blower and Matteson (2003a,b; 2004) and Matteson and Blower (2005) has shown the MCMIS crash file, in the early 2000s, missed between 15 to 30 percent of crashes, depending on the state. Especially if there is a strong differential aspect to these missing crashes, for instance if the missing crashes are disproportionately those from particular states, the carriers that drive through those states more or less often will have measures that are lower or higher than they ought to be. The research also showed that missing crashes were related to truck size and type and crash severity. This could result in a bias against carriers with specific types of operations. To combat this, FMCSA should be more aggressive in encouraging all states to find out which crashes are omitted from MCMIS and to determine ways to reduce the degree of

missingness. The SaDIP program is aimed specifically at improving the completeness of crash reporting at the state level through grant funding tied to specific data improvement projects proposed by the states.

Further, police officers provide additional data. They record details on the state crash report form, such as dozens of detailed data elements, crash sketches, and crash narratives, which could add to the current information on MCMIS as to what happened and contributory factors. A major complication is that what police officers provide differs by state. The Model Minimum Uniform Crash Criteria (MMUCC: http://www.mmucc.us/) is a voluntary guideline for crash data element definitions suitable for states to use in designing police crash reports (PCRs) and statewide crash databases. The current version is MMUCC's 4th edition, but a 5th edition is in preparation and will be published in 2017. MMUCC is a joint product of the National Highway Traffic Safety Administration (NHTSA) and the Governors' Highway Safety Association (GHSA). Relevant standards and data element definitions are found in the American National Standards Institute (ANSI) D16.1 Manual on Classification of Accidents (also due for an updated release in 2017), and the Fatality Analysis Reporting System (FARS). MMUCC, ANSI D16.1, and FARS have been coordinated in recent editions so that they share definitions for common data elements as far as is practical. The MCMIS crash data element definitions and attributes have been incorporated into the MMUCC data element definitions related to commercial motor vehicles so that the two are closely aligned.

MMUCC is divided into sections corresponding to the data elements that describe the entire crash (date, time, location, environmental contributing circumstances, weather, etc.), the vehicles, and the people (drivers, other occupants, and nonoccupants) involved in crashes. As of the 5th edition, the data elements are grouped by: crash, vehicle, driver, occupants, nonoccupants, fatality-specific, and commercial-motor-vehicle-specific. MMUCC is intended as a minimum; it is the recommended set of data elements that need to be collected on every crash (as relevant to the circumstances, involved vehicle types, and severity of the crash). States are encouraged to adopt the MMUCC data element definitions and to add more data elements that they need for their own purposes.

There is no national summary of the level of MMUCC compatibility showing the results for each state. NHTSA offers a free-to-states mapping comparison of the state's PCR and crash database contents to the MMUCC guideline. States are not required to adopt the MMUCC guideline, nor are they required to assess the agreement between their crash data and the MMUCC guideline. However, NHTSA manages traffic records improvement grants (referred to as 405c grants), eligibility for which requires states to address MMUCC compatibility. Traffic Records Assessments, conducted by NHTSA on a 5-year cycle in all states, address MMUCC compatibility based on data quality measurement of uniformity. The assessments also review each state's strategic plan for traffic records improvement, which is an opportunity for the state to report plans to update its PAR and crash database, potentially adopting additional MMUCC data element definitions in the process.

Over time, more states are adopting more MMUCC data element definitions. In recent years, several states have moved to near-complete implementation of the MMUCC guideline. Part of the benefit of such adoption for states is that they can save time and money when they are revising their PAR and crash system database by adopting the already-defined MMUCC elements. There is a much larger benefit at the national level. When states adopt MMUCC data element definitions, NHTSA and others can combine data across states and use the larger dataset to examine the circumstances contributing to crash frequency and severity. The shared data

element definitions make it far easier to develop a merged dataset including multiple states' data. This facilitates comparisons among states and helps increase the effectiveness of data analyses.

For SMS, if states were to adopt the MMUCC data element definitions, it would be much easier for FMCSA to expand its database to include the information not already in MCMIS. This would help FMCSA and others characterize CMV-involved crashes based on the expanded information about noncommercial vehicles and their drivers, expanded lists of contributing circumstances, detailed information about the location, and more. In particular, if FMCSA is interested in downweighting crashes for nonpreventability, this information could help to make such assessments. Further, a MCMIS crash file that included all variables in the MMUCC would permit a much richer evaluation of truck and bus safety than the very limited current number of variables in MCMIS, so that MCMIS could become a more useful CMV safety research dataset. FMCSA has the option to try to add more data elements to MCMIS. In the absence of widespread adoption of MMUCC by states, the job of adding data elements from state PCRs and crash databases is complicated by the fact that each state has a unique set of data definitions. To bring the appropriate data into MCMIS would require translation between the state data definitions and those implemented in MMUCC for those extra data elements. (To capture police sketches, more states are going to electronic PCRs, and in theory, sketches and officers' narratives could be obtained from those that gather and store them electronically.) With widespread MMUCC adoption, this process becomes much easier, assuming MCMIS adopts the MMUCC data element definitions for the to-be-added elements. Information would not require special data collection from motor carriers, but would require translation from multiple states' unique crash database definitions into a standard dataset applicable to all.

The panel believes that FMCSA (and its partners in NHTSA and FHWA) should strongly encourage the states to increase their level of compatibility with MMUCC through various means including grants, partnerships, and guidance.

Better Data on Violations and Severity Weights

As mentioned earlier, there remains some degree of ambiguity about how to represent the specific violations identified during an inspection. Two examples are given here, but many more could be provided.[4] First, assume during the winter, a CMV slides off the road because the driver was driving too fast for conditions. This action can be cited in multiple ways. In many states, driving too fast for conditions is a subset of the speeding statute. It could be listed as violation 392.2, which is a local law violation with a severity weight of zero; as violation 392.25, which is speeding/speeding related with a severity weight of 1; or as violation 392.14, which is failing to use caution for hazardous conditions with a severity weight of 5.

Second, consider a driver who does not have his or her medical card. This action can be cited as a 383.23(a)(2) for operating a CMV without a CDL, which has 8 severity points. It can also be cited three different ways using 383.51 for driving while disqualified or suspended, which has a severity weight of between 1 and 8, or it can be cited as a violation number 383.71H for failing to submit medical documentation as required, which has a severity weight of 1.

This ambiguity as to how to score violations is a source of noise in the system, because these different ways of representing the same situation are associated with substantially different severity weights. The variability has nothing to do with the safety performance of a carrier; it

[4]We thank Collin Mooney for his help in formulating these examples.

results from inspector variance. Therefore, the resulting measures, especially for small carriers, can depend on which inspector was involved and how they decided to code violations.

This situation could be alleviated in two ways. First, inspectors use various software tools to translate their notes into objective codes. These software tools are quite different and as yet there is no standardization. Therefore, some way of requiring these tools to behave alike when faced with various circumstances would be extremely helpful. In addition, FMCSA and CVSA should revisit the entire coding system to look for ways to make clearer how to code various violations.

Finally, in addition to situations where severity weights can differ for the same action, the attribution of severity weights to separate violations might be capable of improvement. Looking at which violations receive higher and lower severity weights, some violations seem more severe but receive what seem to us to be smaller severity weights, and vice versa. Therefore, FMCSA would benefit from subject-matter experts reviewing the severity weights to see if they are internally consistent in this sense. (By internal consistency, we mean that severity weights, within a BASIC, across violations must be ranked in a way that strongly correlates with the subject-matter expertise as to the degree to which a violation rate would be expected to be related to future crashes**.)** More importantly, FMCSA should use MCMIS data to see whether severity weights can be quantified on a more empirical basis. (We note that the new approach to SMS described in Chapter 4 will automatically provide a mechanism for revising severity weights over time.)

Better Data on Clean Inspections

While it is not clear the degree to which the problem persists, there was evidence from the American Transportation Research Institute (2014) that clean inspections are often not reported. Obviously, since the BASICs are weighted frequencies of violations, an unreported clean inspection results in the associated carrier having a higher score than is warranted under SMS. Therefore, inspectors need to be strongly encouraged to report all inspections, regardless of whether they discover violations.

Better Data on Type of Business

We have argued that type of business is a possible reason for additional stratification of SMS and that FMCSA should examine the trade-offs associated with increased stratification. For instance, there is an argument that CMVs that operate locally and those that engage in long-distance trips might be worth separate treatment because they operate in different risk environments. However, any assessment of such a possibility depends on the quality of the responses on the URS and the MCS-150 about a carrier's type of business, and such responses are believed to be inadequate for this purpose. This may be due to the wording used to request the information, the fact information is at times not regularly updated, or because carriers believe that it is to their advantage to include any type of business so that they do preclude themselves from those activities in the future. For this reason, we believe that FMCSA should examine the language on the URS and the MCS-150 to see whether improved information on type of business can be collected, and also to motivate higher quality and frequently updated responses. Again, trip-specific data may be available, but they are not shared with FMCSA, and there would likely

be industry pushback against sharing the level of detail that would truly help improve the utility and validity of information on utilization.

In summary, we make the following recommendation:

Recommendation: FMCSA should continue to collaborate with states and other agencies to improve the quality of MCMIS data in support of SMS. Two specific data elements require immediate attention: carrier exposure and crash data. The current exposure data are missing with high frequency, and what are collected is likely of unsatisfactory quality. Further, to improve the exposure data collected involves not only collecting higher-quality VMT data, but also collecting this information by state and by month. This will enable SMS to (partially) accommodate existing heterogeneity in the environments where carriers travel. Crash data are also missing too often. Also, there is information available from police reports that is currently not represented on MCMIS that could be helpful in understanding the contributing factors in a crash. Such information could help to validate the assumptions linking violations to crash frequency. To address these issues, FMCSA should support the states in collecting more complete crash data, and in universal adoption of the Model Minimum Uniform Crash Criteria, as well as developing and supplying the code needed to automatically extract the data needed for the MCMIS crash file.

NEW VARIABLES IN NEW DATA SOURCES USEFUL IN A NEW SMS

The data deficiencies discussed above in support of SMS would be remedied through use of the same sources currently used in MCMIS, though modified by asking for additional details or ensuring higher quality. Such modifications represent relatively minor changes in the current system. The question remains whether other auxiliary sources of data could augment the current sources of data for MCMIS to benefit the current SMS or the new approach described in Chapter 4.

To examine this question, the panel considered previous findings by a National Academy of Science, Engineering, and Medicine panel (2016), which argued that it is reasonable to think of the causes of CMV crashes due to the following four categories: (1) characteristics of the driver, including those that change from trip to trip, (2) characteristics of the vehicle, (3) the driving environment, and (4) practices and procedures of the carrier. These categories include the following contributory factors (a list not intended to be complete):

1. Driver factors: Demographics, human capital, health conditions, medications used, degree of fatigue, recent sleep history, circadian effects, driving experience, safety record, decision-making ability, work demands;
2. Vehicle factors: Type and age of truck/bus, quality of brakes, quality of tires, other mechanical conditions, maintenance frequency and history, crash history, technology on board for distraction avoidance, collision avoidance;
3. Driving environment factors: Weather, degree of precipitation, time of day, traffic density, road type, degree of road lighting, hazards, safety features, availability of rest stops, impact of other drivers (if not collected as individual driver factors), light condition;

4. Carrier factors: Operation type, type of freight, fleet size, scheduling, logistics, driver turnover, fatigue management, safety culture, compensation level and method.

In any study of the causes of a crash, failure to include some contributory factors (confounding influences) can lead to improper inferences. In our case, omission of such factors can result in SMS misidentifying carriers for intervention. It is clear that the current variables in the MCMIS are inadequate to represent many important factors known to contribute to crashes. In particular, since SMS is concerned with carrier behavior, the fact that MCMIS does not have information on scheduling, logistics, driver turnover, fatigue management policies, and compensation type and level suggests that if such information could be collected by FMCSA, it would benefit the functioning of SMS or the alternative proposed here.

Since we believe a substantial fraction of crashes are due at least in part to carrier operations, and it is those that SMS is attempting to modify, it is clearly important to consider how to gain knowledge of those factors when evaluating SMS. It would be extremely useful to know about the carrier's operations, how it deals with scheduling issues and other logistics, what the driver turnover rate is, how it deals with fatigue management, what its form of compensation is and what level is typical, and what the approach to safety is. Should FMCSA issue an intervention to a carrier, it would be informative to see what aspects of carrier operations were modified to try to address the intervention. Towards this end, we suggest to FMCSA to look into how the following carrier characteristics might be collected. Some of them could be collected through the current MCMIS process, but some of them might need other types of data collections.

- **Information on turnover rate.** We believe most carriers could provide their driver turnover rates if requested. More specifically, they could provide how many drivers they hired (as employees or contractors) and how many they terminated during the previous year. This variable could be very predictive of a company's treatment of its employees, which could be related to safety operations. In addition, a low turnover rate is likely associated with employment of drivers with longer tenures and hence greater experience.
- **Information on type of cargo carried.** Since current questions on type of business are producing lower-quality information, it might be preferable to ask a carrier about its *typical* cargos. (It might be better to ask what cargoes were carried "last year," that is, over a definite period of time.) The response is nearly the same as type of business and might be easier to answer.
- **Information on compensation level/ method.** Similar to turnover rate, it is known that drivers who are better compensated, and those not compensated as a function of miles traveled, have fewer crashes (Rodriguez, Targa, and Belzer, 2006; Belzer, Rodriguez, and Sedo, 2002; Belzer and Sedo, 2017). The mode of compensation can incentivize behavior known to be less safe; pay-by-the-mile compensation will reward drivers for possibly excessive miles. Therefore, information on level and method of compensation, including whether carriers pay for loading and unloading, could be collected, would likely be useful in identifying carriers that are behaving differently with respect to safety operations.
- **Better information on exposure:** We are not certain that the only source of VMT by state and by month is the URS and the MCS-150. We believe that tax information is a possible source of high-quality VMT data, and therefore, we suggest that FMCSA interact with the state taxing authorities, such as the International Registration Plan (IRP)

and the International Fuel Tax Agreement (IFTA), to see whether interagency agreements can be reached to share this information in support of SMS.[5] In addition, at the end of this year, electronic on-board recorders (EOBRs) will be required for most carriers. Results for all carriers could be reported to FMCSA. Having the number of VMT at the end of a year would be extremely easy to produce and would be definitive. Finally, urban/rural and type of road have been shown to be predictive of crash risk. If EOBRs could be used in conjunction with detailed data on the road infrastructure, it might be possible to provide this information.

In conclusion, we have the following recommendation:

Recommendation: FMCSA should investigate ways of collecting data that will likely benefit the recommended methodology for safety assessment. This includes data on carrier characteristics—including information on driver turnover rate, type of cargo, method and level of compensation, and better information on exposure. This additional data collection will likely require additional funds for research and development of the data collection instrument, and greater collaboration between FMCSA and the states as to how to undertake this new data collection effort so that it is standardized across the states.

[5]Also, MCS-150 does not specify which mileage is to be reported, and it should request fuel tax miles for the last full calendar year. If FMCSA can gain access to fuel tax miles, this will provide validation of the miles self-reported by the carrier. Further, fuel tax reports contain miles operated by the carrier in each state, which will allow for analysis of differences among states.

References

Almeida, C., Braveman, P., Gold, M. R., et al. (2001). Methodological concerns and recommendations on policy consequences of the World Health Report 2000. *Lancet,* 357, 1692 1697.

American Transportation Research Institute. (2012). Analyzing the Relationship of Scores to Crash Risk. Micah D. Lueck, October 2012; Minneapolis, MN.

American Transportation Research Institute. (2014). Evaluating the Impact of Commercial Motor Vehicle Enforcement Disparities on Carrier Safety Performance. Amanda Weber and Dan Murray, July 2014; Minneapolis, MN.

American Transportation Research Institute. (2015). Assessing the Impact of Non-Preventable Crashes on CSA Scores. Caroline Boris and Dan Murray, November, 2015; Minneapolis, MN.

Belzer, M. H., Rodriguez, D. A., and Sedo, S. A. (2002). Paying for Safety: An Economic Analysis of the Effect of Compensation on Truck Driver Safety. Washington, DC: U.S. Department of Transportation, Federal Motor Carrier Safety Administration.

Birkmeyer, J. B., Dimick, J. D., and Birkmeyer, N. J. (2004). Measuring the quality of surgical care: Structure, process or outcomes? *Journal of the American College of Surgeons, 198,* 626–631.

Blower, D. F. (1999). The relative contribution of truck drivers and passenger-vehicle drivers to truck/passenger vehicle traffic crashes. *UMTRI Research Review, 30*(2), 1–15.

Blower, D. F. and Campbell, K. (2002). The Large Truck Crash Causation Study. Prepared for the U.S. Department of Transportation, Federal Highway Administration, DTFH61-96-C-0038

Blower, D. and Matteson, A. (2003a). Evaluation of the Motor Carrier Management Information System Crash File, Phase One. University of Michigan Transportation Research Institute, sponsored by the Federal Motor Carrier Safety Administration.

Blower, D. and Matteson, A. (2003b). Patterns of MCMMIS Crash File Underreporting in Ohio. University of Michigan Transportation Research Institute.

Blower, D. and Matteson, A. (2004). Evaluation of Missouri Crash Data Reported to MCMIS Crash File. Prepared for Federal Motor Carrier Safety Administration Task D MCMIS Crash File Evaluation. University of Michigan Transportation Research Institute.

Blower, D. and Matteson, A. (2013). Evaluation of 2010 New Jersey Crash Data Reported to the MCMIS Crash File. UMTRI-2013-48. University of Michigan Transportation Research Institute.

Camilli, G., and Fox, J. P. (2015). An aggregate IRT procedure for exploratory factor analysis. *Journal of Educational and Behavioral Statistics, 40*(4), 377-401.

Chen, L. M., Staiger, D. O., Birkmeyer, J. D., Ryan, A. M., Zhange, W, and Dimick, J. B. (2013). Composite quality measures for common inpatient conditions. *Medical Care, 51*(9), 832–837.

Cohen, R. J. and Swerlik, M. (2001). *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. New York: McGraw-Hill.

Courtney, J., Krumholz, H., Wang, Y., and Turnbull, B. (2002). Using composite measures for the public reporting hospital performance data. *Connecticut Medicine 66*(10),633–634.

Coyne, J. S. and Hilsenrath, P. (2002). The World Health Report 2000. *American Journal of Public Health, 92*(1), 30–34.

Craft, R. (2012). Coding Scheme for Motor Carrier Crash Accountability: A Test of Using a Modified Critical Reason Methodology. Washington, DC: U.S. Department of Transportation, Federal Motor Carrier Safety Administration.

Donabedian A. (1980). *Explorations in Quality Assessment and Monitoring: The Definition of Quality and Approaches to Its Assessment.* Ann Arbor, MI: Health Administration Press.

Federal Motor Carrier Safety Administration. (2015). Large Truck and Bus Crash Facts. Available: https://www.fmcsa.dot.gov/safety/data-and-statistics/large-truck-and-bus-crash-facts.

Federal Motor Carrier Safety Administration. (2016a). Safety Measurement System (SMS) Methodology: Behavior Analysis and Safety Improvement Category (BASIC) Prioritization Status, Version 3.0.5; Methodology revised September 2015, Document Revised February 2016.

Federal Motor Carrier Safety Administration. (2016b). CSA Effectiveness Measures. U.S. Department of Transportation, June 30, 2016.

Federal Motor Carrier Safety Administration. (2017). Mission Statement. Available: https://www.fmcsa.dot.gov/mission.

Federal Register. (2010). Withdrawal of Proposed Improvements to the Motor Carrier Safety Status Measurement System (SafeStat) and Implementation of a New Carrier Safety Measurement System (CSMS). Available: https://www.fmcsa.dot.gov/regulations/notices/2010-8183

Gelman, A, Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, Third Edition. Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science.

Gelman, A., Meng, X.-L., and Stern, H. S. (1996), Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807.

General Accounting Office. (1999). Truck Safety: Effectiveness of Motor Carriers Office Hampered by Data Problems and Slow Progress on Implementing Safety Initiatives. GAO/T-RCED-99-122.

Gibbons, R. D. and Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423. doi:10.1007/BF02295430.

Government Accountability Office. (2014). Federal Motor Carrier Safety: Modifying the Compliance, Safety, Accountability Program Would Improve the Ability to Identify High Risk Carriers. GAO-14-114.

Green, P. E. and Blower, D. (2011). Evaluation of the CSA 2010 Operational Model Test. University of Michigan Transportation Research Institute, UMTRI-2011-08.

Gregory, R. J. (2003). *Psychological Testing: History, Principles and Applications*. 4th ed. Boston: Allyn & Bacon.

Hatfield, L. A., Hodges, J. S., and Carlin, B. P. (2014). Joint models: when are treatment estimates improved? *Statistics and Its Interface, 7*(4), 439–453.

Independent Review Team. (2014). Blueprint for Safety Leadership: Aligning Enforcement and Risk. Appointed by Secretary of Transportation Anthony R. Foxx to Review the Federal Motor Carrier Safety Administration's Safety Oversight of the Motor Carrier Industry. Available:

https://www.fmcsa.dot.gov/sites/fmcsa.dot.gov/files/docs/FINAL%20REPORT%20-%20IRT_July%2015.pdf

Institute of Medicine. (1999). Measuring the Quality of Health Care: A Statement by The National Roundtable on Health Care Quality. Division of Health Care Services, Institute of Medicine, M. S. Donaldson (Ed.). Washington, DC: National Academy Press.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Landon, B. E., Normand, S-L. T., Blumenthal, D. et al. (2003). Physician clinical performance assessment: prospects and barriers. *Journal of the American Medical Association, 290,* 1183–1189.

Matteson, A. and Blower, D. (2005). Evaluation of North Carolina Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. Sponsor: U.S. Department of Transportation, Federal Motor Carrier Safety Administration.

National Academy of Sciences, Engineering, and Medicine. (2016). *Commercial Motor Vehicle Driver Fatigue, Long-Term Health, and Highway Safety: Research Needs.* Washington, DC: The National Academies Press. doi: 10.17226/21921.

National Highway Traffic Safety Administration. (2013). Traffic Safety Facts: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. DOT HS 812 139.

Nunnally, J. C. (1978). *Psychometric Theory*. 2nd ed. New York: McGraw-Hill.

Reise, S., Moore, T. M., and Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personal Assessment*, *92*(6), 544–559. doi: 10.1080/00223891.2010.496477.

Rodriguez, D. A., Targa, F., and Belzer, M. H. (2006). Pay incentives and truck driver safety: a case study. *ILR Review*, *59*(2), 205–225.

Rothstein, R. (2000). Toward a composite index of school performance. *The Elementary School Journal*, *100*(5), 409–441.

Shahian, D. M., Wolf, R. E., Iezzoni, L. I., Kirle, L., and Normand, S-L. T. (2010). Variability in the measurement of hospital-wide mortality rates. *New England Journal of Medicine, 363*, 2530–2539.

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298.

Volpe National Transportation Systems Center. (2010). Carrier Safety Measurement System (CSMS) Violation Severity Weights.

Volpe National Transportation Systems Center. (2014). Table 2 in The Carrier Safety Measurement System (CSMS) Effectiveness Test by Behavior Analysis and Safety Improvement Categories (BASICs).

Appendix A
Agendas from Open Portions of Panel Meetings

First Meeting of the Panel on the Review of the Compliance, Safety, Accountability Program of the Federal Motor Carrier Safety Administration
June 29–30, 2016
The Keck Center of the National Academies, 500 5th St. NW, Washington, DC
Room 201

Open Session, 10:30am–5:30pm

10:30 AM **Welcome and Introductions**
- **Connie Citro,** *Director, Committee on National Statistics*
- **Jack Van Steenburg,** *Chief Safety Officer and Assistant Administrator, Federal Motor Carrier Safety Administration (FMCSA)*

10:45 AM **Commercial Motor Vehicle Safety Alliance**
- **Collin Mooney,** *Executive Director, CVSA*

11:30 AM **Fixing America's Surface Transportation Act**
- **Joe DeLorenzo,** *Director, Office of Enforcement and Compliance, FMCSA*

12:15 PM ***Working Lunch (3rd Floor Cafeteria)***

1:00 PM **Continuation of Fixing America's Surface Transportation Act**
- **Joe DeLorenzo,** *Director, Office of Enforcement and Compliance, FMCSA*

1:45 PM **Review/Critique of Safety Management System by ATRI**
- **Daniel Murray,** *Vice President, American Transportation Research Institute*

2:45 PM **Review/Critique of Safety Management System by GAO**
- **H. Brandon Haller,** *Assistant Director, Physical Infrastructure Team*
- **Jeff M. Tessin,** *Senior Statistician, Applied Research and Methods*

3:45 PM **Response from FMCSA**
- **Joe DeLorenzo,** *Director, Office of Enforcement and Compliance, FMCSA*

4:00 PM ***Break*** (refreshments available inside conference room)
4:15 PM **Additional Questions from Panel Members**
4:30 PM **Hearing from Stakeholders**
4:45 PM **Public Comment**

5:30 PM **Adjournment**

****

Second Meeting of the Panel on Review of the Compliance, Safety, Accountability Program of the Federal Motor Carrier Safety Administration

August 30–31, 2016

The Keck Center of the National Academies, 500 5th St. NW, Washington, DC

Room 100

Open Session, 10:00am–5:30pm

10:00 AM **View from Australia**
Ann Williamson, *University of New South Wales*

10:45 AM **Break**

11:00 AM **View from Schneider National**
Tom DiSalvi, *Schneider National*
Don Osterberg, *Schneider National (retired)*

11:45 AM **View from Panther**
Irwin Shires, *Panther Premium*

12:00 PM ***Working Lunch to continue morning discussion (3rd Floor Cafeteria)***

1:00 PM **Bus Perspective**
Rudolph Supina, *DATTCO*

1:30 PM **How CSA treats buses**
Joe DeLorenzo, *Director, Office of Enforcement and Compliance, FMCSA*

2:30 PM **View from Truck and Bus Interest Groups**
Julie Heckman, *American Pyrotechnic Association*
Cary Catapano, Sr., *National School Transportation Association*
Ken Presley, *United Motorcoach Association*
Tom Weakley, *Owner-Operator Independent Drivers Association*

3:30 PM ***Break*** (refreshments available inside conference room)

3:45 PM **Quality of the MCMIS Data**
Scott Valentine, *FMCSA*

4:30 PM **Public comment**

5:30 PM **Public adjournment, panel in recess**

6:00 PM **Working Dinner**

****

Third Meeting of the Panel on Review of the Compliance, Safety, Accountability Program of the Federal Motor Carrier Safety Administration

December 15–16, 2016

The Keck Center of the National Academies, 500 5th St. NW, Washington, DC

Room 208

8:45 AM View of Safety Experts

- 8:50 AM: William Voss (ICAO), Jacqueline Duley (Engility Corp.)
- 9:20 AM: Rob Molloy (NTSB)
- 9:50 AM: John Lannen (Truck Safety Coalition)

- 10:20 AM: Shuie Yankelewitz (Central Analysis Bureau), Jean Gardner (Central Analysis Bureau)

10:50 AM ***Break***

11:05 AM Relative vs. Absolute Measures — Joe DeLorenzo, FMCSA

11:30 AM Report of the Insurance Institute for Highway Safety — Eric Teoh, IIHS

12:05 PM ***Working Lunch to continue morning discussion (3rd Floor Cafeteria)***

Appendix B
Details of the SMS Algorithm

DATA INPUT

The data used for input to the Safety Measurement System (SMS) comes in four separate segments of the Motor Carrier Management Information System (MCMIS): (1) carrier census data, (2) inspection data, (3) violation data, and (4) crash data. Carrier census data include: Department of Transportation (DOT) number, vehicle miles traveled, address of headquarters, active status, interstate/intrastate, operational status code, equipment type and count, cargo classification, number of power units, and whether the carrier transports hazardous materials. Inspection data include: date/time of the inspection, truck or bus carrier, vehicle configuration, cargo body type, gross vehicle weight range, date, state and county, inspection level, number of violations, and whether the vehicle had enough hazardous material to require a placard. Violation data—which are able to be linked to the relevant inspection—include violation code, whether a citation was issued, whether the violation has been adjudicated and the result, whether the violation is a post-crash violation, whether the violation resulted in an out-of-service order, the associated Behavior Analysis and Safety Improvement Categories (BASIC), the associated severity weight, and the out-of-service weight (see below). Crash data include DOT number, location, time/date, sequence of events, number of vehicles involved, type of vehicle, weather, light condition, injuries, fatalities, whether a vehicle was towed-away, whether hazardous materials were released, and the severity crash weight (see below).

Input Data Filters. There are several filters or checks that have to be passed before data are used in the calculation of the BASIC measures. (These are in addition to the data sufficiency standards, discussed below.) They are: (1) for carrier data: carriers must be active; (2) for inspection data: the inspections must be in the last 24 months, and the type of inspection must match the BASIC they are being applied to; (3) for violation data: in cases of multiple counts of the same violations, SMS only counts a violation once, and if any of the duplicates have an out-of-service attribute, that is made an attribute of the single counted violation; (4) for crash data: the crash must have occurred in the past 24 months, and it must have been a tow-away crash, or one with injuries or fatalities.

NUMERATORS OF MEASURES: WEIGHTINGS OF VIOLATIONS

The numerators of the seven BASIC scores are sums over violations of the product of weights. The following are the weights that make up these products:

Severity Weights. Severity weights are assigned to violations to reflect the degree of their correlation with crash occurrence and/or crash consequences, ranging from 1 for the least important to 10 for the most. Severity weights are only comparable within and not across BASICs. In other words, a severity of, for example, 5 for one BASIC is not necessarily comparable to a severity of 5 for a different BASIC. A severity weight is increased by 2 when

the violation in question results in an out-of-service order. Also, the sum of all violation severity weights for any single inspection in any one BASIC is limited to 30.

Crash Severity. Crashes are assigned different weights according to the seriousness of the crash. (This is only to compute the Crash BASIC.) The idea is to assign greater weight to crashes involving injuries, fatalities, release of hazardous materials (HM), and a lower weight is assigned otherwise. The crash severity weights are as follows:

Crash Type	Crash Severity Weight
Involves tow-away but no injury	1
Injury or fatality	2
Involves an HM release	Crash Severity Weight + 1

Time Weights. Inspections, violations, and crashes are weighted according to how recently they occurred. Violations and crashes that took place within the last 6 months receive a weight of 3, those that took place between 6 months ago and 1 year ago are given a weight of 2, and those that occurred between 1 year and 2 years ago receive a weight of 1.

Therefore, the numerators of the BASICS are either sums of the product of severity weights and time weights for the noncrash BASICs, or sums of the product of crash severity weights and time weights for the Crash BASIC.

CALCULATION OF DENOMINATORS: MEASURES OF EXPOSURE

Since the number of crashes experienced and violations obtained during inspections are a function of time on the road, it is important to normalize the number of violations, or the number of crashes, by a measure of time driving, or exposure, in order to make comparisons meaningful. SMS makes use of three measures of exposure to construct its denominators.

For BASICs related to driver inspections, the denominators used are the total time weight associated with the relevant inspections. This normalization is used for HOS Compliance, Controlled Substances/Alcohol, and Driver Fitness. The second type of denominator relates to vehicle inspections; the normalization is similar to the first denominator, except that it pertains only to vehicle-based inspections and the resulting violations, and is used for BASICS for Vehicle Maintenance and HM Compliance (with the latter only for inspections for HM carriers). Third, for the Unsafe Driving BASIC, for behaviors that usually prompt an inspection, such as speeding or swerving, the normalization is the Average Power Units (APU) multiplied times the utilization factor. Similarly, the Crash Indicator BASIC is also normalized by the same denominator.

TABLE B-1 Definition of Utilization Factors for Combination Segment

Combination Segment	
Vehicle Miles Traveled (VMT) per Average Power Unit (PU)	Utilization Factor
< 80,000	1
80,000 – 160,000	$1 + \frac{VMT\ per\ Average\ PU - 80{,}000}{133{,}333}$
160,000 – 200,000	1.6
➢ 200,000	1
No Recent VMT Information	1

SOURCE: Table 3.1 in SMS Methodology, FMCSA (2016).

TABLE B-2 VMT per Average PU for Straight Segment

Straight Segment	
VMT per Average PU	Utilization
< 20,000	1
20,000 – 60,000	$\frac{VMT\ per\ Average\ PU}{20{,}000}$
60,000 – 200,000	3
➢ 200,000	1
No Recent VMT Information	1

SOURCE: Table 3.2 in SMS Methodology, FMCSA (2016).

DEFINITION OF THE SEVEN DIFFERENT BASICS

SMS groups violations into categories, which are referred to as Behavior Analysis and Safety Improvement Categories (BASICs). They are as follows:[1]

- Unsafe Driving BASIC – This BASIC measures "Operation of commercial motor vehicles (CMVs) in a dangerous or careless manner. *Example violations include: speeding, reckless driving, improper lane change, texting while operating a CMV, not wearing safety belts* [2,3] The measure is defined as follows:

$$Unsafe\ Driving\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ violations}{Average\ PUs\ x\ Utilization\ Factor}$$

[1]Taken from FMCSA (2016a).
[2]In cases of multiple counts of the same violation, each violation counts only once per inspection.
[3]Violations can also result from inspections motivated by traffic enforcement stops for moving violations.

- Crash Indicator BASIC – This BASIC measures "Historical pattern of crash involvement, including frequency and severity. The BASIC is based on information from state-reported crashes that meet reportable crash standards. All reportable crashes are used regardless of the carrier's or driver's role in the crash. This BASIC uses crash history that is not specifically a behavior but instead the consequence of a behavior or a set of behaviors." This BASIC uses state-reported crash data in MCMIS to form the following measure:

$$Crash\ Indicator\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ crashes}{Average\ PUs\ x\ Utilization\ Factor}$$

  A reportable crash is one that results in at least one fatality, one injury requiring transportation to a medical facility, or one vehicle needing to be towed from the scene.

- HOS [Hours of service] Compliance BASIC – This BASIC measures "Operation of CMVs by drivers who are ill, fatigued, or in noncompliance with the HOS regulations. This BASIC includes violations of regulations pertaining to records of duty status (RODS) as they relate to HOS requirements and the management of CMV driver fatigue. *Example violations include: operating a CMV while ill or fatigued, requiring or permitting a property-carrying CMV driver more than 11 hours, failing to preserve RODS for 6 months/failing to preserve supporting documents"*
  The measure is defined as follows:

$$HOS\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ violations}{Total\ time\ weight\ of\ relevant\ inspections}$$

- Vehicle Maintenance BASIC – This BASIC measures "Failure to properly maintain a CMV and prevent shifting loads, spilled or dropped cargo, and overloading of a CMV. *Example violations include: inoperative brakes, lights, and other mechanical defects, improper load securement, failure to make required repairs"* The equation used to calculate the BASIC is as follows:

$$Vehicle\ Maintenance\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ violations}{Total\ time\ weight\ of\ relevant\ inspections}$$

- Controlled Substances/Alcohol BASIC – This BASIC measures "Operation of CMVs by drivers who are impaired due to alcohol, illegal drugs, and misuse of prescription or over-the-counter medications. *Example violations include: use or possession of controlled substances or alcohol, failing to implement an alcohol and/or controlled substance testing program"*
  The measure is defined as follows:

$$Controlled\ Substances\ /Alcohol\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ violations}{Total\ time\ weight\ of\ relevant\ inspections}$$

(Since all such violations are considered to be out-of-service violations, it is not the case that two additional severity points are added to any violations for this BASIC.)

- HM [hazardous materials] Compliance BASIC – This BASIC measures "Unsafe handling of HM on a CMV. *Example violations include: failing to mark, label, or placard in accordance with the regulations, not properly securing a package containing HM, leaking containers, failing to conduct a test or inspection on a cargo tank when required by the United States Department of Transportation (U.S. DOT)"*
  The measure is defined as follows:

$$HM\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ violations}{Total\ time\ weight\ of\ relevant\ inspections}$$

- Driver Fitness BASIC – This BASIC measures "Operation of CMVs by drivers who are unfit to operate a CMV due to lack of training, experience, or medical qualifications. *Example violations include: failing to have a valid and appropriate commercial driver's license (CDL), being medically unqualified to operate a CMV, failing to maintain driver qualification files.*"
  The measure is computed as follows:

$$Driver\ Fitness\ Measure = \frac{Total\ of\ time\ and\ severity\ weighted\ violations}{Total\ time\ weight\ of\ relevant\ inspections}$$

DATA SUFFICIENCY STANDARDS

BASIC scores are only computed for carriers that have had sufficient activity to justify that the computation will not have an unreasonably large variance. The sufficiency standards that have been instituted by FMCSA are different for the different BASICs, reflect information from the previous 24 months, and are as follows:

- Unsafe Driving: Carriers with at least three inspections with at least one violation
- Crashes: Carriers with at least two applicable crashes
- Hours of Service: Carriers with at least three relevant inspections and at least one inspection with at least one violation
- Vehicle Maintenance: Carriers with at least five relevant inspections and at least one inspection resulting in one violation.
- Controlled Substance/Alcohol: Carriers with at least one violation.
- HM Compliance: Carriers with at least five relevant inspections and where at least one inspection resulted in a violation.
- Driver Fitness: Carriers with at least five relevant inspections and one inspection resulting in a violation.

During 2014–2015, 38 percent of active carriers had sufficient data to be scored for at least one BASIC, and these carriers had 92 percent of the crashes during that time period. (There is

also a Critical Mass Threshold Test that is similar to the data sufficiency standard, but it sets a higher minimum standard before an intervention can be taken.)

STRATIFICATION BY TYPE AND SIZE OF CARRIER

SMS makes use of two types of stratifications in the sense that its computations and rankings are carried out separately within these strata:

Type of Carrier. Unsafe Driving and Crash Indicator BASICs are stratified (referred to as being segmented) by combination trucks/motorcoaches and straight trucks/other vehicles. For carriers with both types of vehicles, if 70 percent or more of the power unit types are combination trucks/motor coaches, the carrier is assigned to the combination stratum, and otherwise it is assigned to the straight stratum.

Safety Event Groups. Carriers are meant to be compared to other carriers that are of a comparable size. For that reason, for each BASIC, SMS places carriers in what are referred to as safety event groups based on the number of events over the previous 24 months (e.g., inspections, violations, and crashes depending on the BASIC in question) in which they have been involved during the previous 2-year period. This allows comparison of carriers that have had roughly comparable levels of exposure. The safety event groups are defined as follows for the Combination Segment (Table B-3) and Straight (Table B-4) Safety Event Groups.

TABLE B-3 Definitions of Combination Safety Event Groups

Safety Event Group	Unsafe Driving BASIC	Crash Indicator BASIC	Hours of Service Compliance BASIC	Vehicle Maintenance BASIC	Controlled Substance / Alcohol BASIC	HM Compliance BASIC	Driver Fitness BASIC
	Number of Inspections with Unsafe Driving Violations	Number of Crashes	Number of Relevant Inspections	Number of Relevant Inspections	Number of Relevant Inspections	Number of Relevant Inspections	Number of Relevant Inspections
1	3-8	2-3	3-10	5-10	1	5-10	5-10
2	9-21	4-6	11-20	11-20	2	11-15	11-20
3	22-57	7-16	21-100	21-100	3	16-40	21-100
4	58-149	17-45	101-500	101-500	4+	41-100	101-500
5	150+	46+	501+	501+		101+	501+

TABLE B-4 Definitions of Straight Safety Event Groups

Safety Event Group	Unsafe Driving BASIC	Crash Indicator BASIC	Hours of Service Compliance BASIC	Vehicle Maintenance BASIC	Controlled Substance/Alcohol BASIC	HM Compliance BASIC	Driver Fitness BASIC
	Number of Inspections with Unsafe Driving Violations	Number of Crashes					
1	3-4	2	See above for combination segment	See above for combination segment	See above for combination segment	See above for combination segment	See above for combination segment
2	5-8	3-4					
3	9-18	5-8					
4	19-49	9-26					
5	50+	27+					

TRANSFORMING SCORES TO PERCENTILES

FMCSA believes that SMS should, at its core, be a relative measure of carriers' safety performance, rather than an absolute measure. This is because FMCSA's resources in support of interventions are fixed, and therefore SMS has to function as a priority ranking system, used to identify a specific number of carriers for intervention. Towards this end, the scores for carriers within a safety event group are ranked and the percentile associated with each score is computed for each carrier.

There are a few additional rules regarding computation of percentile ranks. Percentiles are only generated from scores of U.S. interstate and hazardous material (HM) carriers. For the remaining carriers, percentiles are assigned based on interpolations of the percentiles for interstate and HM carriers.

Further, there is a 12-month rule, in which the percentiles for carriers that have had all of their violations (or crashes) occur more than 12 months ago are dropped (sometimes there is the added condition that no violation was recorded in the last inspection whenever it occurred). The percentiles for the remaining carriers remain unchanged after those are eliminated.

Carriers with percentiles above a designated threshold are flagged to receive an Alert, which typically indicates that they will receive an intervention. (Alerts can also be triggered as a result of investigations. Alerts were not discussed in this report.) The thresholds are summarized in Table B-5.

TABLE B-5 Intervention Thresholds for BASICs

BASIC	Intervention Thresholds		
	Passenger Carrier	Hazardous Materials (HM)	General
Unsafe Driving, Crash Indicator, HOS Compliance	50%	60%	65%
Vehicle Maintenance, Controlled Substances/Alcohol, Driver Fitness	65%	75%	80%
HM Compliance	80%	80%	80%

NOTE: There are specific definitions of passenger carriers and HM carriers on pp. 2–9 of FMCSA (2016a).

FMCSA's analysis has demonstrated that the Unsafe Driving, Crash Indicator, and HOS Compliance BASICs have the strongest linkage to crash risk, and therefore their thresholds, seen above, are lower. In addition, because the consequences of any resulting crashes are greater, the thresholds are lower for passenger carriers and for HM carriers.

Appendix C
IRT Example Using MCMIS Data

Jacob Spertus

The following provides a brief worked example of an IRT model as discussed in Chapters 3, 4, and 5 applied to actual SMS data. It uses a very small proportion of the full data, purposefully selected for quality and carrier size. It is not intended to be a comprehensive, unbiased analysis but merely a proof of concept. A full implementation would look roughly similar, but with many more carriers and violations included.

This appendix is included to show that the proposed model is relatively easy to implement in MCMIS. However, this model is in many ways a toy model. It is only applied to 865 carriers (not 200,000) and only unsafe driving violations that appeared for 5 or more carriers are included; there are none of the features, including multidimensionality (in fact, only one BASIC is used); and the priors have not been updated over time. Therefore, the quality of the fit should not be seen as an indication of the quality of the fit of the full model as represented in Chapters 4 and 5.

Suppose for each carrier $i \in \{1, \ldots, C\}$ we have a measure of exposure E_i, a count of eligible inspections n_i^k, and a count of violations y_i^k, where $k \in \{1, \ldots K\}$ indicates the type of violation. Note that inspections are superscripted with k because the type of inspection determines which violations can be recorded. We propose the following model:

$$\mathbb{P}(N_i = n_i | E_i) = \text{Poisson}(\lambda E_i) \tag{C-1}$$

$$\mathbb{P}\left(Y_i^k = y_i^k \middle| n_i^k, p_{ik}\right) = \text{Binomial}(n_i^k, p_{ik}) \tag{C-2}$$

$$\text{logit}(p_{ik} | \theta_i) = \beta_k + \alpha_k \theta_i \tag{C-3}$$

$$\theta_i \sim \mathcal{N}(0,1) \tag{C-4}$$

$$\beta_k \sim \mathcal{N}(0, 3^2) \tag{C-5}$$

$$\alpha_k \sim \text{log-}\mathcal{N}(1,1) \tag{C-6}$$

The parameter λ in equation (C-1) estimates the rate of inspections per VMT across the population. p_{ik} in equation (C-2) represents the probability of carrier i receiving violation k at a given inspection. It is modeled as a logistic function of the prevalence (or difficulty) parameter β_k, which reflects the marginal prevalence of violation k in the data, and discrimination parameter α_k, which reflects the association of violation k and latent safety. Thus for a given

violation k a higher value of β_k indicates that it is observed more frequently and a higher value of α_k indicates it is more associated with safety or danger.

To implement this model on a small scale, we use SMS data from 2014 to 2015, selecting a subset of the carrier population and including only unsafe driving violations. The carrier subpopulation we select are those carriers with 100 or more average power units, more than 80,000 vehicle miles traveled (VMT) per average power units (APU) reported, and less than 200,000 VMT per APU reported (medium to high utilization carriers according to SMS methodology). This leaves only 865 carriers. Finally in order to set a lower bound on sparsity, we drop violations that occur in less than 5 carriers. After this processing, 35 different types of unsafe driving violations (features) remain for 865 carriers (observations).

We used the 'rstan' interface to Stan for Hamiltonian Monte Carlo to fit the fully Bayesian models described above (Stan Development Team 2016). In the first step, a simple Poisson model (equation (1)) is fit. For exposure E_i, we take 100,000 VMT. λ is specified with a non-informative prior and thus gives, essentially, the average inspections per 100,000 VMT. The observed mean in the population is 2.014, while the posterior mean (95% CI) of λ is 1.806 (1.802, 1.811)

The second step uses equations (C-2) and (C-3), and priors (C-4), (C-5), and (C-6), to fit a Binomial item response theory (IRT) model. The prior on θ_i is selected to enforce identifiability of the model. The log-normal prior on α_k, which has support only over the positive real numbers, encodes the assumption that α_k is strictly positive. We provide default choices for the priors on $\beta_k$and α_k by matching the specification in the 'edstan' R package for IRT models (Furr, 2017). Note that substantive knowledge about the effect of certain violations on safety could be included via the prior (C-6). The 'trials' for this IRT model are inspections eligible for unsafe driving violations, i.e. driver inspections (level 1, 2, 3, or 6). A brief caveat: unsafe driving violations (e.g., speeding) often precipitate an inspection, so inspections might not be the best denominator for these violations. VMT might be a better exposure variable, requiring a Poisson model with VMT as offset.

Figure C-1 shows the characteristic curves for θ. The probability of receiving a given violation at an inspection is plotted against θ. Thus more dangerous carriers (with higher θ) receive more violations over all, and certain violations with especially high frequency.

The top 20 most discriminating violations (highest α) are shown in Table C-1.

Figure C-2 plots unsafe driving violations per inspection against θ. There is a very close mapping between these two metrics, such that θ more or less returns the number of unsafe driving violations per inspection.

Figure C-3 plots crashes per 100,000 VMT against θ. There is a positive relationship but also very high variance about the least squares line (low R^2). The Spearman correlation between crashes per 100000 VMT and θs from the Binomial IRT is 0.23.

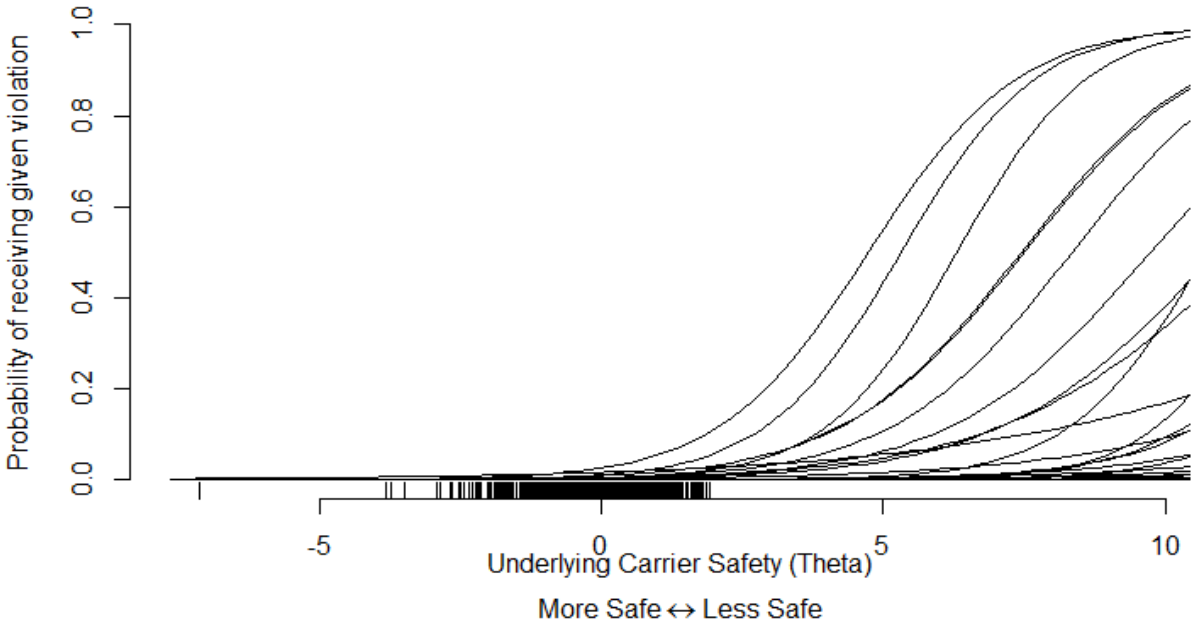


FIGURE C-1: Characteristic curves showing on the y-axis the probability of receiving a given violation (at a single inspection), and given safety level theta on the x-axis. Parameters are generated from Binomial IRT. Rug shows actual estimated thetas. Larger values of theta correspond to less safe carriers.

TABLE C-1: Alpha, Beta, and Prevalence for the 20 Violations with the Highest Alpha Value Generated from Binomial IRT.

Description	Alpha	Beta	Prevalence
Speeding	0.931	-11.2	10
Following too close	0.871	-5.5	3010
Lane restriction violation	0.839	-4.5	8328
State-local laws operating a CMV while texting	0.833	-8.9	95
Inattentive driving	0.771	-11.0	12
State-local laws speeding 6-10 miles per hour over limit	0.760	-3.6	18946
Operating a CMV while texting	0.734	-9.6	47
Driving a commercial motor vehicle while texting	0.635	-8.8	111
Improper lane change	0.631	-5.3	3614
Failing to use seat belt while operating CMV	0.627	-4.7	6532
State-local laws speeding 11-14 miles per hour over limit	0.620	-4.7	6629
Using a handheld mobile telephone while operating a CMV	0.568	-5.5	2752
Allowing or requiring driver to use a handheld mobile telephone	0.565	-10.4	20
Failure to slow down approaching a railroad crossing	0.540	-11.4	7
State-local laws speeding in work/construction zone	0.530	-5.8	2134
Failing to stop at railroad crossing bus	0.489	-11.3	8
State-local laws speeding 15+ miles per hour over limit	0.459	-5.3	3549
Railroad Grade Crossing violation	0.441	-10.4	21
Improper passing	0.426	-7.3	463
Scheduling run to necessitate speeding	0.400	-10.7	15

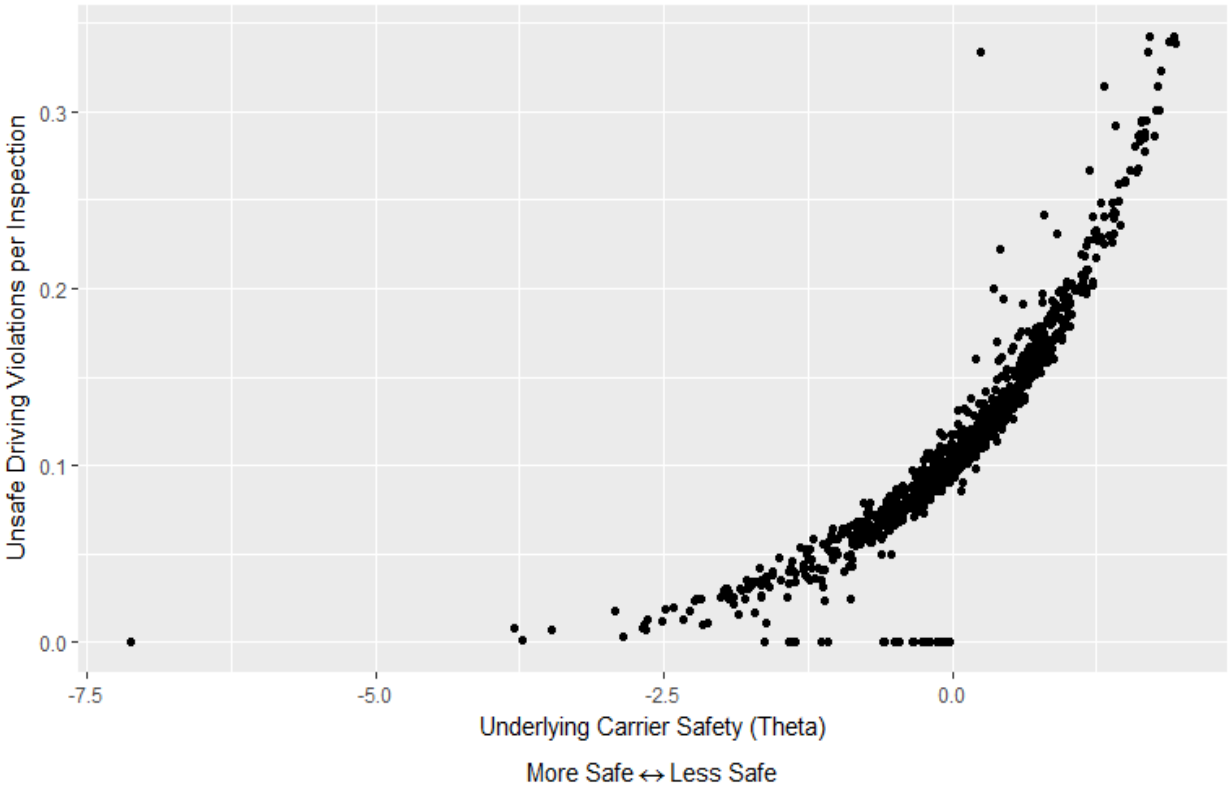


FIGURE C-2 Unsafe driving violations per inspection plotted against theta generated from Binomial IRT. Larger values of theta correspond to less safe carriers.

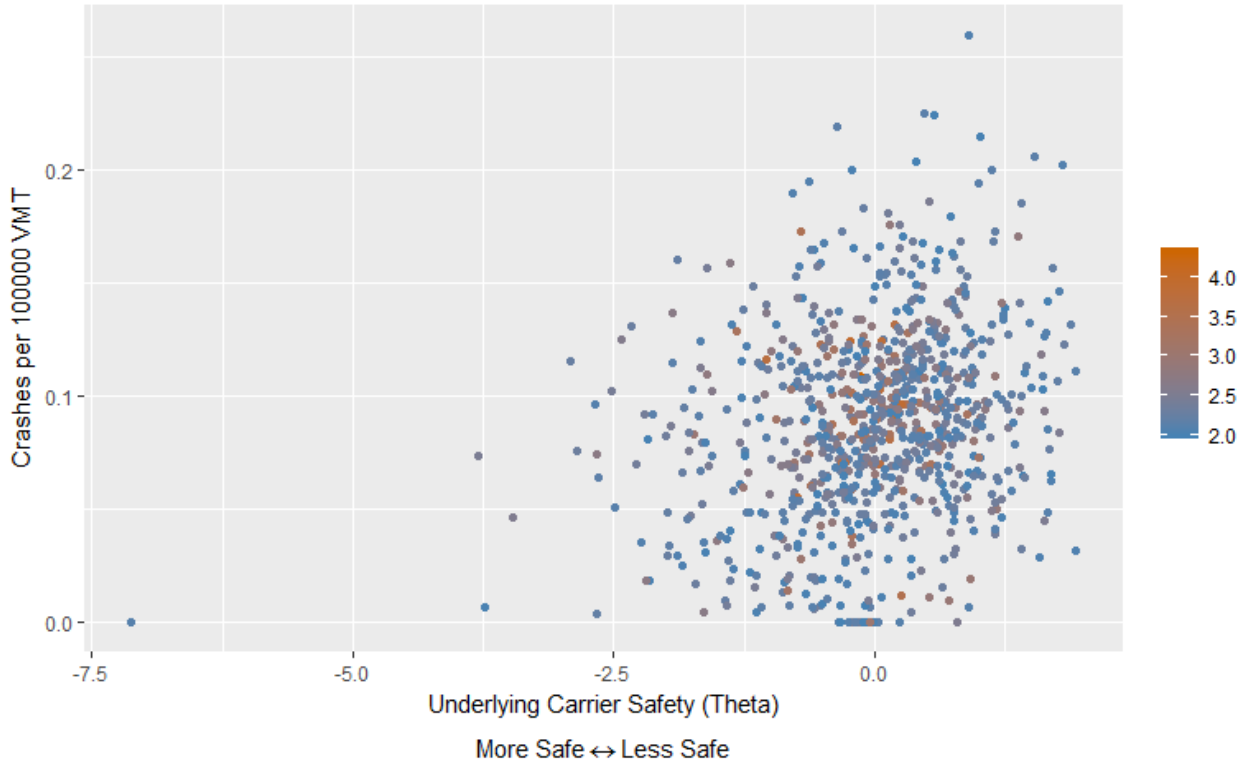


FIGURE C-3 Crash per 100000 VMT by safety level theta for each carrier in subpopulation estimated from Binomial IRT. X-axis is theta, y-axis is crashes per 100000 VMT, color is log base 10 APU. Larger values of theta correspond to less safe carriers.

References

Furr, Daniel C. 2017. edstan: Stan models for item response theory, version 1.0.6. https://CRAN.R-project.org/package=edstan

Stan Development Team. 2016. RStan: the R interface to Stan, version 2.14.1. http://mc-stan.org

Appendix D
Biographical Sketches of Panel Members and Staff

JOEL GREENHOUSE (*Cochair*) is a professor in the Department of Statistics at Carnegie Mellon University. Prior to his current appointment, he was associate and assistant professor in the Department of Statistics at Carnegie Mellon and has served as an associate dean for academic affairs. He is also an adjunct professor of psychiatry, and adjunct professor of epidemiology at the University of Pittsburgh. His areas of research include methods for research synthesis and the use of Bayesian methods in practice. He is a fellow of the American Statistical Association, a fellow of the American Association for the Advancement of Science, and an elected member of the International Statistical Institute. He has served as editor-in-chief for *Statistics in Medicine*, associate editor for *Statistics, Politics, and Policy*, associate editor for the *Journal of the American Statistical Association*, and editor of the *IMS Lecture Notes – Monograph Series*. He received a B.S. in mathematics from the University of Maryland at College Park, and M.P.H. in biostatistics, A.M. in statistics, and Ph.D. in biostatistics all from the University of Michigan.

SHARON-LISE NORMAND (*Cochair***)** is a professor of health care policy (biostatistics) in the Department of Health Care Policy and in the Department of Biostatistics at the Harvard School of Public Health. She has made important contributions to the use of hierarchical models, and to a better understanding of causal inferential techniques. Further, she has served as director of MASS-DAC, the data-coordinating center responsible for collecting, analyzing, and reporting on the quality of care for adults discharged following a cardiac procedure from all non-federal hospitals in Massachusetts. She is a fellow of the American Statistical Association. She has served as associate editor of *Circulation, Cardiovascular Quality and Outcomes, Statistics in Medicine, Biometrics,* and *Health Services and Outcome Research Methodology*, methods editor for *Psychiatric Services*, and guest editor for a special issue of *Health Services and Outcomes Research Methodology*. She earned B.Sc. and M.Sc. degrees in statistics from the University of Western Ontario and a Ph.D. in biostatistics from the University of Toronto.

MICHAEL BELZER is associate professor of economics in the College of Liberal Arts and Sciences at Wayne State University. He also is a research scientist at the University of Michigan's Institute of Labor and Industrial Relations, and is associate director of the Alfred P. Sloan Foundation's Trucking Industry Program, which focuses on trucking industry operations and industrial relations and where he directs its Trucking Industry Benchmarking Program. Current benchmarking efforts include an Owner Operator Cost of Operations Survey, in partnership with the Owner Operator Independent Drivers Association, and a Safety Best Practices Program. His research interests include trucking industry organization and operations, labor management relations, employment policy, and safety. He currently is studying the relationship between truck driver pay (both method and level) and safety, as well as issues related to truck driver hours of work. Earlier, he spent 10 years driving trucks. He received a Ph.D. from Cornell University.

DANIEL BLOWER is associate research scientist emeritus with the Center for the Management of Information for Safe and Sustainable Transportation Group at the University of Michigan

Transportation Research Institute. He has extensive experience with all primary national crash data files, and many state crash data files. His primary area of research is traffic crash causation. Past projects include investigating the crash experience of younger truck drivers, developing an event tree for heavy truck accidents, and developing statistical models relating vehicle configuration and operating environment to the probability of accident involvement. He is chair of the Michigan Truck Safety Commission. He received a B.A. and Ph.D. in history from the University of Michigan.

LINDA NG BOYLE is professor and chair of Industrial & Systems Engineering at the University of Washington, and also has a joint appointment with Civil & Environmental Engineering. Previously, she was an associate professor at the University of Iowa and a senior researcher at the U.S. Department of Transportation's Volpe Center. Her research centers on driving behavior, crash countermeasures, crash and safety analysis, and statistical modeling. She is an associate editor for the journal *Accident Analysis and Prevention*. She received a B.S. in industrial engineering from SUNY Buffalo and M.S. in inter-engineering/human factors and Ph.D. in civil and environmental engineering from the University of Washington.

MICHAEL DANIELS is professor and chair of the Department of Statistics and Data Sciences, and professor in the Department of Integrative Biology, at the University of Texas at Austin. His research interests lie in Bayesian methodology, biostatistics, hierarchical modeling, incomplete data models, and causal inference. He is coauthor, with Joe Hogan, of *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analyses*. He currently serves as coeditor of *Biometrics*. In the past he served as associate editor for *Biometrics, Journal of the American Statistical Association, Statistics and Probability Letters,* and *Biostatistics.* He is a fellow of the American Statistical Association and an elected fellow of the International Statistics Institute. He received an A.B. in applied mathematics from Brown University and Sc.D. in biostatistics from Harvard University.

DON HEDEKER is a professor of biostatistics in the Department of Public Health Sciences at the University of Chicago. He is an expert in the analysis of longitudinal data, missing data analyses, and hierarchical modeling. He has developed several freeware computer programs for statistical analysis of such data (MIXREG for normal-theory models, MIXOR for dichotomous and ordinal outcomes, MIXNO for nominal outcomes, and MIXPREG for counts). He has served as associate editor for *Statistics in Medicine* and *Journal of Statistical Software*. Along with Robert Gibbons, he is coauthor of *Longitudinal Data Analysis*. He is a fellow of the American Statistical Association and elected member of the International Statistical Institute. He received a Ph.D. in quantitative psychology from the University of Chicago.

BRENDA LANTZ is the associate director of North Dakota State University's Upper Great Plains Transportation Institute (NDSU/UGPTI) as well as the program director for the Transportation Safety Systems Center branch of UGPTI. As program director she leads a team responsible for the development of inspection and investigative commercial vehicle safety systems for use by FMCSA. Her primary research interests include commercial vehicle safety systems and analysis. She has worked on projects involving the CSA/SMS system, including Commercial Vehicle Inspection and Investigative Systems Software Development. She also participated in the Commercial Motor Vehicle Driver Risk Factors Study and *An Evaluation of*

Commercial Vehicle Drivers' and Safety Inspector's Opinions Regarding the MCSAP, the Roadside Inspection Process, and Motor Carrier Safety. She received a B.S. in sociology and M.S. in applied statistics from North Dakota State University and Ph.D. in business administration, supply chain management and information systems from Pennsylvania State University.

DAN McCAFFREY is principal research scientist at Educational Testing Service in Princeton, New Jersey. Previously he was head of the statistics group at RAND. He has focused his recent research on applications of statistics to education policy issues, especially the effectiveness of value-added models for use in assessing the performance of elementary and secondary teachers, which involves use of hierarchical models. He is a fellow of the American Statistical Association. He has served as editor of the *Journal of Educational and Behavioral Statistics* and *Statistics and Public Policy* and as associate editor of the *Journal of Educational and Behavioral Statistics*, *Elementary School Journal*, *Journal of the American Statistical Association*, and *Journal of Computational and Graphical Statistics*. He received a Ph.D. in statistics from North Carolina State University.

BRISA SANCHEZ is associate professor of biostatistics in the Department of Biostatistics at the University of Michigan. Her areas of research include structural equation and latent variable models, with applications to environmental epidemiology. She has served as associate editor for *Statistics in Medicine* and the *Journal of the Royal Statistical Society, Series C.* She received a B.S. in mathematics and M.S. in statistics from the University of Texas at El Paso, M.Sc. in biostatistics from the Harvard School of Public Health, and Ph.D. in biostatistics from Harvard University.

ROBERT SCOPATZ is a senior transportation analyst with the safety practice of VHB, Inc. His work has focused on improving safety behavior and advising states and the federal government on data quality improvement. His work in commercial motor vehicle safety includes projects for FMCSA evaluating hours of service regulations, developing predictive analyses of safety inspections and crashes, and supporting the first round of Safety Data Improvement Program (SaDIP) grants. For the AAA Foundation for Traffic Safety, he conducted an evaluation of large combination vehicle safety and crash data quality. He is also secretary, past president, and long-time executive board member of the Association of Transportation Safety Information Professionals. He received a Ph.D. in experimental psychology from Columbia University.

JUNED SIDDIQUE is associate professor in the Feinberg School of Medicine at Northwestern University. His research interests, in addition to medical applications, lie in the treatment of missing data, the use of hierarchical models, and causal inference. He has served as associate editor of *Statistics in Medicine*. He is a fellow of the American Statistical Association. He received a B.S. in economics from the University of Wisconsin, M.S. in statistics from George Washington University, and Ph.D. in biostatistics from the University of California at Los Angeles.

MICHAEL L. COHEN (*Costudy Director*) is a senior program officer for the Committee on National Statistics. He also serves as a consultant on statistical analysis for other divisions in the National Academies of Science, Engineering and Medicine. Previously, he was a mathematical

statistician at the Energy Information Administration and held positions at the School of Public Affairs at the University of Maryland and at Princeton University. His general area of interest is the use of statistics in public policy, with particular focus in census undercount, model validation, and robust estimation. He is a fellow of the American Statistical Association and an elected member of the International Statistical Institute. He has a B.S. in mathematics from the University of Michigan and an M.S. and a Ph.D. in statistics from Stanford University.

ESHA SINHA (*Costudy Director*) has worked on a variety of CNSTAT panel studies, workshops, and planning meetings. She co-edited *Improving Measurement of Productivity in Higher Education* and was co-rapporteur for several other workshops. She organized and presented at a workshop on Measuring Productivity in Higher Education in India. She has an M.A. in economics from GIPE, India, and worked as a research assistant in the Indian Institute of Management, Ahmedabad, before attending SUNY Binghamton. She has a Ph.D. in economics from SUNY Binghamton.

JACOB SPERTUS (*Consultant*) is a research assistant for Sharon-Lise Normand in the Department of Health Care Policy at Harvard Medical School. His research interests include Bayesian methods and high-dimensional causal inference, on which he has written a number of statistical articles. He has presented his research at Harvard Medical School, the Society of Biological Psychiatry, and the U.S. Food and Drug Administration. He received a B.A. in mathematics from Bowdoin College.

RICHARD PAIN (*Consultant*) recently retired from the Transportation Research Board, where he was the transportation safety coordinator in the board's Technical Activities Division. He served as staff to a wide range of committees, including studies of truck and bus safety, statistics in transportation, visualization in transportation, and future truck and bus safety research opportunities, as well as an international conference on research on the health and wellness of commercial truck and bus drivers. Prior to his work for the board, his work focused on human factors and safety research and evaluation in transportation, nuclear, civil, and military areas; training, development, conduct, and evaluation studies; and human engineering reviews. He has a B.A. in psychology from Hofstra University, and an M.A. in clinical psychology and a Ph.D. in applied experimental psychology from Michigan State University.