Statistical Issues in the Safety Measurement and Inspection of Motor Carriers

James Gimpel

University of Maryland

jgimpel@umd.edu

May 2012

The U.S. Department of Transportation's Federal Motor Carrier Safety Administration (FMCSA) has developed a Safety Measurement System (SMS) for gauging the safety of individual motor carriers traveling U.S. highways. The methodology of the SMS is detailed in a January 2012 report prepared by the John A. Volpe National Transportation Systems Center in Cambridge, MA (Volpe Center 2012). The key aspect of this new measurement system is the inspection of motor carriers by government regulators using established criteria for determining the safety of vehicles and the fitness of drivers.

Specifically, seven safety areas are identified by FMCSA as of critical: Unsafe Driving, Fatigued Driving, Driver Fitness, Controlled Substances and Alcohol, Vehicle Maintenance, Cargo-Related security, and Crash Indication assessment. The point of scoring carriers with this system is to target firms operating unsafe vehicles for fines and penalties in an effort to prevent accidents, injuries and fatalities on the national roadway network. Violations of sufficient severity can result in a motor vehicle or driver being placed out of service. Once a vehicle or driver has been declared "out-of-service," operators must correct the unacceptable conditions before they can continue to drive or operate the vehicle.

This report documents some concerns and problems with the methodology of the SMS, and the data on which it is founded.

Data Generation Process

The data on which the SMS is based originate from inspection records from on-road safety inspections of Level III or higher and crash records reported by state government agencies. The inspections data are made available for study in the Motor Carrier Management Information System (MCMIS) database and are accompanied with motor carrier Census data containing information about firm location, fleet size, and number of drivers.

From a statistical standpoint, is important to note how these inspections are carried out, and how the data are generated. While an inspection can occur almost anywhere, historically inspections have most frequently occurred at roadside inspection stations throughout the 50 states. This has changed as states now carry out more mobile inspections at rest stops, truck stops and other roadside sites. The recorded data originate from where these inspections take place. The locations of inspection stations, their times and hours of operation, are neither random nor uniform across the highway system. Inspection records are not likely to be reflective of the traffic volume of the nationwide carrier fleet, or the geographic location of firms, but instead the idiosyncratic practices of state regulators. For instance, recent data are highly sensitive to the proliferation of inspections carried out in California, Arizona and Texas, and the relative dearth of inspections in much of the Northeast.

What local regulators choose to focus on in terms of enforcement emphasis is also highly variable. Current data (Spring 2012) on BASIC percentile scores show that firms operating out of Montana and North Dakota exhibit far lower scores on the Unsafe Driving BASIC than firms physically located in Kentucky, West Virginia, New Hampshire and Massachusetts. This is an enforcement pattern that cannot be explained away by traffic density or road conditions. The Fatigued Driver BASIC scores are highest for carriers operating out of Florida, Georgia and Idaho, and lower in Washington state –

patterns that reflect local enforcement emphases not attributes of carriers operating in these regions. Vehicle maintenance BASIC violations are highest in Florida, Texas, South Carolina and Connecticut, but lower on carriers based in Hawaii, Pennsylvania, Delaware and Maryland – variation that cannot be explained by traffic or population density measures. From a statistical standpoint, the problem is the extraordinary level of heterogeneity in measurements resulting not from the characteristics of firms, drivers, and road conditions, but due to the application of the measuring instruments by data gatherers.

Because the data generation process is a highly imperfect reflection of the nature and quality of operator activity, the data are not a reflection of a representative cross-section of the carrier operators who are directly responsible for fleet safety – the responsible parties. Based on straightforward comparisons with trucking censuses, the data vastly overrepresent the firms with very large fleets, while vastly underrepresenting the impressive number of small carriers operating two, three, or perhaps only a single vehicle. While the larger carriers probably haul the vast majority of cargo, it cannot be determined based on current data how safety practices vary and how violations are distributed across the entire population of operators nationwide. Because it is operators who are subject to penalty, they must be represented in any competent study, not cargo.

Since the data are an inferior representation of the nationwide population of motor carriers, it is fundamentally unsound to generalize from any of the information contained in the data on inspected vehicles to the broader population of all carriers. Any data analysis carried out by any entity based on the inspections data, including data contained in the remainder of this report, should be accompanied with the caveat that it represents only the particular cases contained in the data. Nothing can be extrapolated from it, and its external validity is in doubt. In summary, using data generated only by happenstance of where inspections occur, based on idiosyncratic local enforcement practices, introduces selection bias, providing a misleading picture of important statistical relationships that inform essentials of the regulatory regime. Any results based on the data are suspect due to the atypical or unusual nature of the sample.

The problem of sample selection bias cannot be dismissed by FMCSA on the grounds that it is only interested in the carriers who are sampled in the inspection process. After all, it is not merely external validity, or the generalization to non-sampled carriers, that is called into question by the bias in data. Key statistical relationships thought to be causal are misconstrued as well (Heckman 1976; 1979; Goldberger 1981). For instance, regression analysis based on the partial data will exhibit bias in the coefficients in much the same way as excluding important explanatory variables produces bias. Relationships between independent and dependent variables are not properly represented even for those carriers that have been subject to inspection and are included in the MCMIS system.

Unsafe Driving Scores and Crashes

One example of where the present data can mislead regulators is in relationships found between specific inspection violations and crash risk. What is true of that relationship among the highly overrepresented large carriers in the data may not be true of the poorly represented midsized and small carriers, or of the population of carriers writ large. This variation in safety practices across

the population of firms could result from a number of causes, including the important fact that the small carriers are frequently self-employed owner-operators, and confront different incentives for safety as well as costs associated with regulatory penalties than drivers who are employed by someone else.

Even using the data provided by FMCSA the variability in the relationship between the BASIC score for unsafe driving and the score for crash rates can be made evident if we divide it into thirds by fleet size. Such a division creates three groups of trucking firms: the bottom third consisting of those with three or fewer trucks; the second third operating between 4 and 10 trucks; and the top third comprised of those with 11 or more trucks. By industry standards, a fleet with only 11 trucks is usually not considered large, but in the MCMIS data, these firms are in the top third in size.

On the following pages, we present three scatterplots (Figures 1, 2 and 3) showing the nature of the relationship between the BASIC percentile scores for unsafe driving and the crash rate drawing upon data from Spring 2012. The first plot exhibits the bivariate relationship for carriers in the lowest one-third of size (PU<4), the second plot is for the middle third PU=4-10), and the third plot captures the relationship for the very largest carriers (11 or more). Note that these fleet size cutpoints for thirds of the distribution are not where they would be set for the entire population of carriers, just for the MCMIS sample.

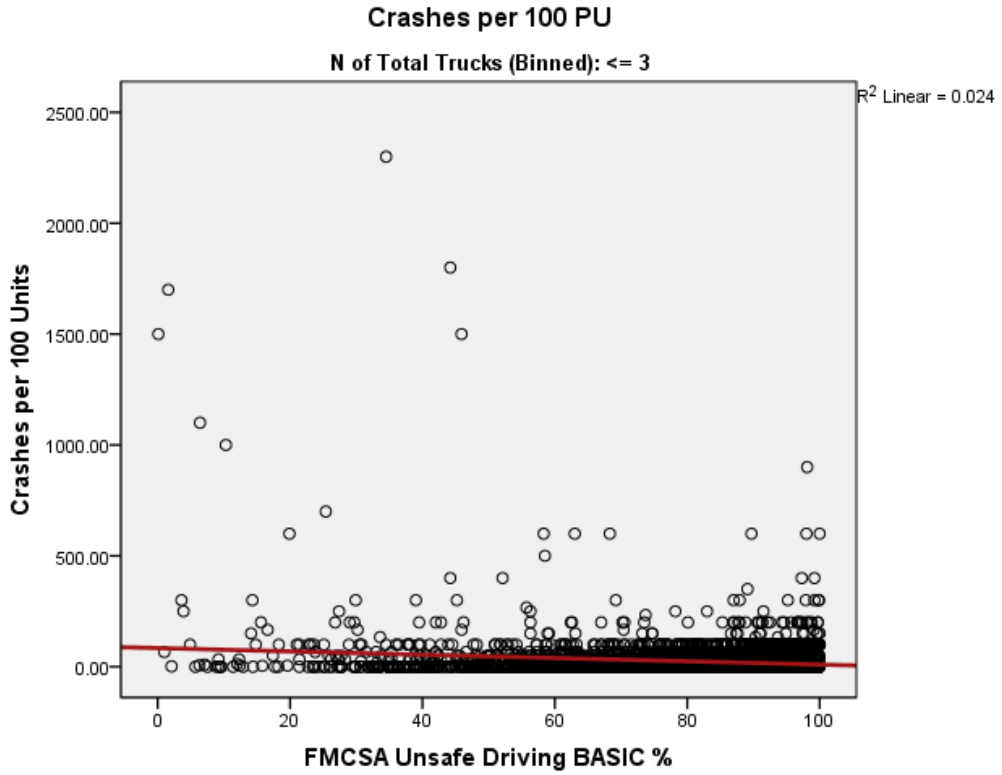

Figure 1. Bivariate Relationship between Unsafe Driving Score and Crashes Per Power Unit, Smallest Carriers, N=4,265

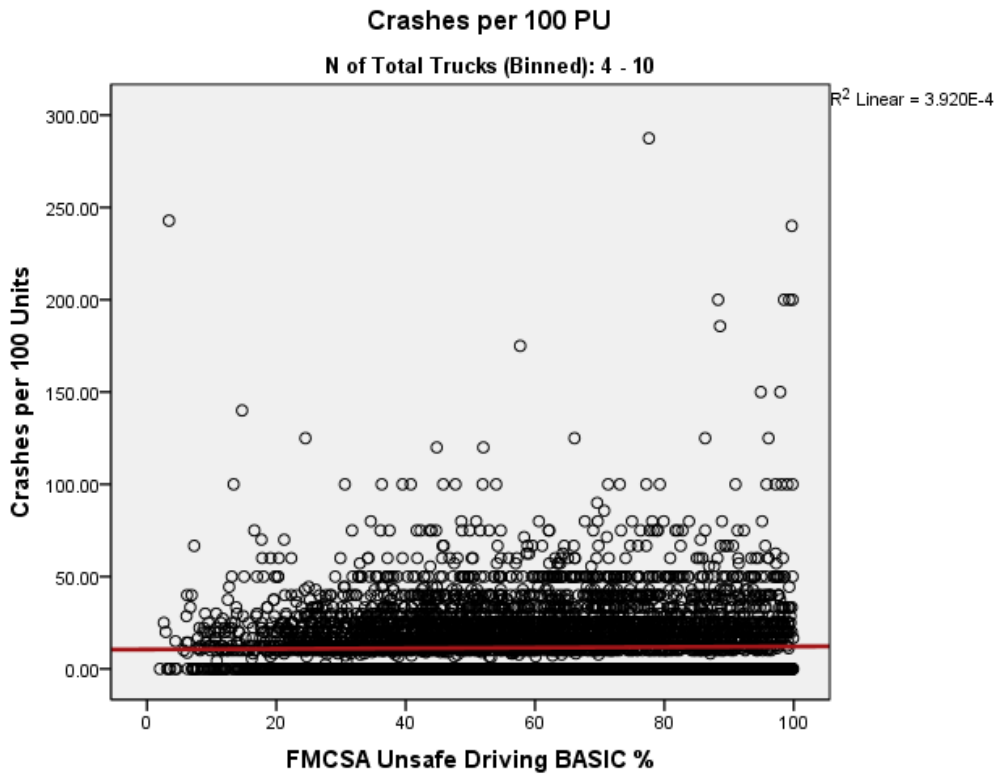

Figure 2. Bivariate Relationship between Unsafe Driving Scores and Crashes per Power Unit, Midsized Carriers, N=7,815

Due to variability in the crash ratings from some extreme outlying observations in the right tail of the distribution of those values, the differences in the regression lines are not easy to detect via visual inspection alone. But the regression coefficients reveal that for the subsample of 4,265 small carriers, the bivariate relationship is actually negative: a ten point increase in the Unsafe Driving BASIC score is associated with a modest .07 ($p \leq .001$) *drop* in the number of accidents per power unit. Taking the log transformation of the dependent variable to adjust for the skewness in the distribution did not change the relationship appreciably – it is less negative, but not statistically significant.

For the mid-sized and larger trucking operations (see Figure 3), however, the linear relationship is modestly positive. Specifically, for firms operating between 4 and 10 trucks (N=7,815), a ten point increase in the Unsafe Driving score is associated with a .001 increase in the accident rate, though this relationship does not reach conventional levels of statistical significance ($p \leq .08$). For the largest firms, operating more than 10 trucks (N=16,707), the relationship is positive, showing an increase in the accident rate of .006 ($p \leq .001$) for every 10 point increase in the Unsafe Driving score.

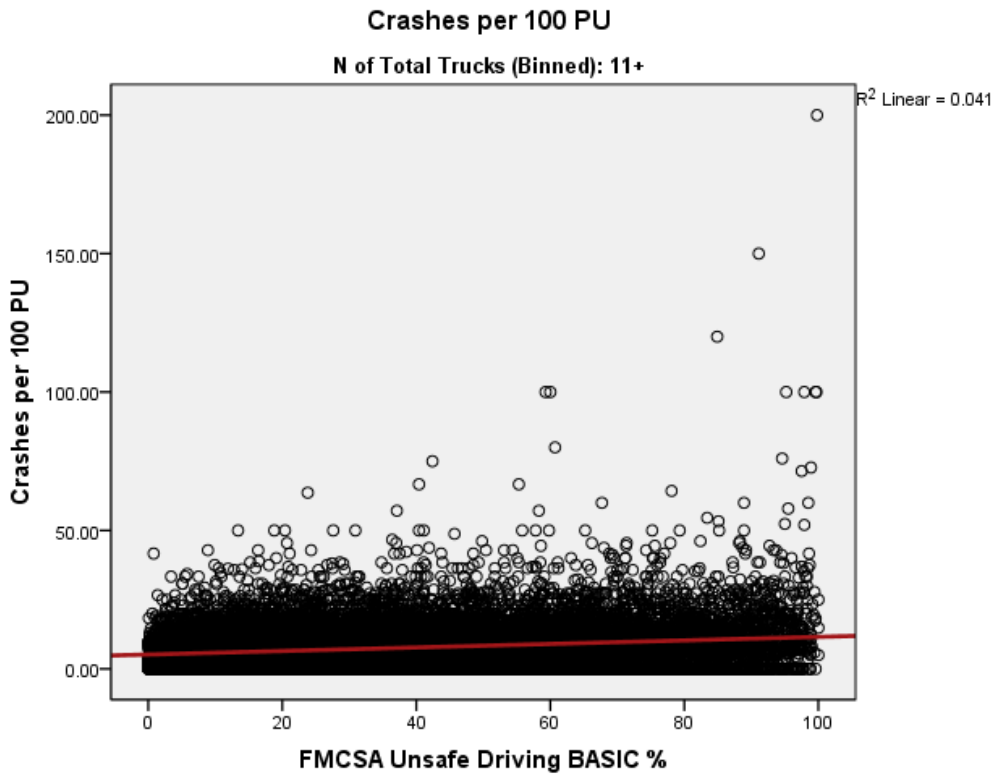

Figure 3. Bivariate Relationship between Unsafe Driving Scores and Crashes per Power Unit, Larger Carriers, N=16,707

In summary, then, based on the information provided in the MCMIS system, and stratifying the data by firm size, the estimated statistical relationships are not the same across this limited sample of inspected vehicles. The relationship between unsafe driving scores and crash rate is weakly positive and not statistically significant for mid-sized carriers (Figure 2), and *inverse* but statistically significant for the very smallest carriers (Figure 1).

Similar data analysis for other BASIC indicators shows that their relationships with crash ratings are highly variable across cohorts of firm size. For instance, for the smallest one-third of carriers, the relationship between the BASIC Vehicle Maintenance score and crash rate is weakly positive but not statistically significant, but it is positive and significant for midsized and larger carriers (Figures A1-A3). This auxiliary analysis is shown in Appendix A.

Why the Differences with the Small Carriers?

There are a number of possible explanations for why the small carriers differ from the midsized and larger carriers: some are substantive and others are technical. First, the above patterns could reflect the reality that small owner-operators are less risky drivers and better at complying with safe driving regulations than drivers in larger enterprises. After all, penalties hurt the owner-operator directly, but may not fall on contract drivers or employees. Deciding whether this explanation accounts for differences in BASIC inspection scores and crash rates is beyond the scope of this study, but the

possibility surely merits further investigation. Ownership structure has certainly been found to matter for a wide variety of other business practices.

A second reason why the relationships may be different for smaller carriers, again, has to do with the peculiarities of this sample of small trucking firms and inspection scores. Certainly it could be that since smaller firms are badly underrepresented in the inspection database that the sample is unreflective of the true relationships that exist in the population of small carriers. This cannot be determined from the data at hand, of course, but must remain speculative until representative data on small carriers is gathered and studied. In sum, whether the relationships revealed in Figures 1, 2 or 3, are real, or merely a product of sample selection is difficult to determine, which is why statisticians point out that selection bias risks confounding the substantive relationships of interest with the selection process itself (Berk 1983, 391).

Certainly another explanation of a technical nature is that the relationship for smaller carriers differs from that of larger ones as an artifact of the methodology of the scoring system itself. This has to do with the small number of inspections that smaller carriers typically experience, and the extraordinary level of variance in BASIC scores that results. This matter is explained in greater detail in the final section below.

Weaknesses of BASIC Scoring

FMCSA's BASIC methodology essentially boils down to a measurement equal to the ratio of inspections in which violations are found to the total number of inspections (Volpe Center 2012). To the agency's credit, there are reasonable adjustments made for the recency of violations and inspections and the severity of the violation, but the basic formula is a ratio of violations to total inspections, as the FMCSA puts it:

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

FMCSA then takes the additional step of "binning" or segmenting the data by categories of inspection frequency as it calculates percentile ranks for carriers within each bin. Firms with fewer than five inspections are removed from the data entirely. Carriers with no inspections in which violations are detected are also removed from the data (Volpe Center 2012, 3-15). With this second step, firms with a series of nothing but clean (no violations found) inspections do not have these inspections credited to them – *they are not even included in the data*. The exclusion of no violation inspections from the data has been a repeated complaint of reputable firms for many months now. Apparently the omission of credit for clean inspections occurs not simply as an oversight by regulators, but is part of the official policy, introducing selection bias by deliberate design.

Parenthetically, for calculation of the Unsafe Driving BASIC score, the number of total inspections is ignored, as the denominator is fleet size (average power units) adjusted for miles traveled (utilization factor):

$$\textit{BASIC Measure} = \frac{Total\ of\ time\ and\ severity\ weighted\ applicable\ violations}{Average\ Power\ Units\ x\ Utilization\ Factor}$$

The official calculation for this particular measure does not take into account clean inspections at all. This is also true of the crash indication BASIC measure.

Returning to the calculation of the BASIC scores (both those that purport to include all inspections and those for which the number of inspections is irrelevant); within each bin or "safety event group," carriers are ranked from high to low on a scale of 0 to 100 within each category. Higher scores indicate a greater number of violations.

For the Unsafe Driving indicator, XYZ Freight Company operating one truck, for example, with 9 total inspections, is ranked alongside only carriers that have experienced similar numbers of inspections. If XYZ Freight has 125 points of violations, and 9 inspections, its raw score = 125/9 = 14. OP Corporation, operating 3 trucks, has 210 points of violations over 10 inspections yielding a raw score of 210/10=21. Based on how many violations are possible and could be counted, on the face of it, these seem like pretty low scores. But the actual violation point values do not count directly, because carriers are ranked in percentile terms relative to other carriers in their safety event group.

Calculated this way, the scores do not offer information on how safe or unsafe a trucking firm is. The score only describes how safe a firm is relative to other firms experiencing similar numbers of inspections during a given period. The latter is useful information, perhaps, but it is very different from measuring safety violations in the sense in which such quantification is ordinarily understood. The distinction is important as the relative scores make it very difficult to understand what the measurement means from one year to the next, much less over a five year or ten year period of operations. A carrier's violations could rise sharply over time, becoming more of a crash risk, and yet this carrier could wind up with the same percentile ranking vis-à-vis others in its cohort. Alternatively, a carrier could improve its safety compliance, and still wind up with the same *or higher* score.

Given that the number of carriers in each event safety group is not likely to be a random subset of carriers, but will exhibit distinctive characteristics (e.g., firm size, ownership structure, geographic location of operations), this means that the entire group could be moving in an appreciably less/more safe direction over time, yet this would never be registered by the scoring methodology.

Finally, it should also be noted that the potential volatility in cross-time percentile rankings for smaller carriers may have another source: the frequent entry and exit of smaller enterprises that are regularly reconstituting the peer group. Larger firms are better established in the marketplace, and their survival across time is more assured. Small firms come and go with regularity, and they operate under a wider variety of conditions than larger firms. Given the churn in the number of small carriers, we could expect extraordinary volatility in the year-to-year variation in percentile rankings even when the number and type of violations remains the same.

Small Carriers and the Law of Large Numbers

Nowhere do we see the limitations of the BASIC scoring methodology more clearly than in that segment or group of carriers that have the fewest operators and are subject to fewer inspections. Under standard enforcement practices, the vast majority of smaller trucking firms go uninspected and therefore unmeasured. We have already noted that this is a major source of selection bias in the data, as the small number of very large carriers winds up being highly influential in regression specifications. The omission of small carriers is the quite natural result of basing data inclusion only on inspections and violations – on average smaller carriers have less exposure due to fewer travelled miles, and may also have fewer violations for reasons highlighted earlier. Consequently, the records that are included for the small carriers wind up having very few inspections counted in the denominator of BASIC formulae at any given point in time.

The problems of ratio and rate measures when denominators are small are well-known to statisticians. When ratio measures are based on fewer than 20 observations in the denominator, they are often considered unreliable, and frequently they are not even published. Twenty is a common cut-off point since beyond that changes in total variation contributed by successive measurements diminishes. Data with fewer than 20 observations in the denominator are not considered to meet a sufficient level of accuracy based on calculated standard errors. A denominator with 5 inspections is far less reliable than one with 40 when both have the same numerator. In a single 12 or 18 month period, however, many firms may have only five, six or eight inspections.

Small changes in the number of violations per inspection have a substantially larger effect when the number of total inspections is small than they do when the number of total inspections is larger. Suppose XYZ freight moves from 200 points in violations to 260 points between inspection 5 and inspection 6. That moves the raw score on which the BASIC percentile is constructed from 40 to 43. But an identical change in violation points from 600 to 660 for OP Corporation between inspection 39 and 40 moves the raw score from 15 to 16.5, having *half* the impact.

Rates based on a small number of inspections are highly variable and for that reason unreliable as measures. When rates are unstable it is virtually impossible to distinguish random fluctuation from true changes in the underlying risk of crashes or accidents. Comparisons of firms based on unstable rates can lead to spurious conclusions about safety risks.

By way of statistical background, the notion that high variability is associated with small numerators can be understood through reference to *the law of large numbers*. In statistical terms, as the number of samples increases, the average of these samples is likely to reach the mean of the whole population. Or, as the number of trials increase, the difference between the expected and actual value moves toward 0.

This explains why typically values obtained based on large numbers of observations provide stable estimates of the true, underlying quantity. Conversely, values based on small numbers of observations may fluctuate dramatically from year to year, or differ considerably from one case to another, even when there is no meaningful difference between them.

Binning the data by inspection frequency does not mitigate the high variation in scores for less frequently inspected carriers.

In summary, then, small trucking firms are subject to few inspections, meaning that whatever BASIC scores they generate, high or low, are not reliable indicators of these firms' propensity to operate safely and in compliance with regulatory standards. For firms with more trucks and greater exposure, the higher number of inspections yields an average that will be more reflective of their actual rate of safety and compliance.

Conclusions

A small share of the nationwide fleet of motor carriers is selected for inspection each year. Due to local peculiarities and pronounced biases in the selection process, the resulting data collection is an imperfect representation of the population of carriers, and especially small carriers. In addition, the measurements specified by federal regulators as part of the SMS inspections regimen are subject to wide variation in emphasis and application by geographic location. Consequently, statistical relationships detected in these data are not only a cloudy reflection of the true population, but may well be flat wrong.

The relationship between the Unsafe Driving BASIC measure and crash rates for the smallest one-third of carriers is actually the inverse of commonsense expectations, showing fewer crashes per power unit occurring with rising percentile scores. This could point to a substantively significant attribute of small vs large carriers, but could also be an artifact of the small number of inspections among this group of carriers, and the resulting high variance that calls the reliability of the BASIC scores into serious question.

Vehicle inspections may prevent accidents, but only if the appropriate aspects of driver behavior and vehicle maintenance are being monitored and inspected. Why the BASIC scores for unsafe driving are so weakly associated with crash risk across the entire MCMIS sample is likely the consequence of including safety-irrelevant aspects of operator behavior in the measure. The measures require thorough reconsideration. Trucking industry sources suggest that the vast majority of violations falling within the fatigued driver BASIC category involve minor infractions associated with recordkeeping, and therefore do not precisely capture aspects of driver disposition or vehicle roadworthiness that serve the interest of accident prevention, such as driving longer hours than safety standards allow. If the scoring for fatigued and unsafe driving were focused on those violations actually germane to common understandings of those concepts, the statistical relationships between measures and outcomes would surely be stronger.

Increasing the number of biased observations only amplifies the magnitude of the bias. Simply increasing the total number of inspections carried out is not likely to help much if current tendencies in inspection and measurement remain in place. Large operators will continue to rack up numerous inspections that do little to alter their overall measure of compliance and safety while smaller operators will be subject to wild fluctuations in their BASIC scores. Binning the data by frequency of inspection

does nothing to protect smaller carriers from the threat of being placed out of service for violations that larger carriers can largely ignore.

Sources

Berk, Richard A. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Sociological Review* 48: 3: 386-398.

Goldberger, Arthur S. 1981. "Linear Regression After Selection." *Journal of Econometrics*. 15: 3: 357-366.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5: 4: 475-492.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 45: 1: 153-161.

The Volpe Center. 2012. "Safety Measurement System (SMS) Methodology." Report prepared for the Federal Motor Carrier Safety Administration, Version 2.2. Cambridge, MA: John A. Volpe National Transportation Systems Center.

Appendix A.

Relationship of Vehicle Maintenance and Fatigued Driver BASIC % Score to Crash Ratings per 100 PU by Fleet Size

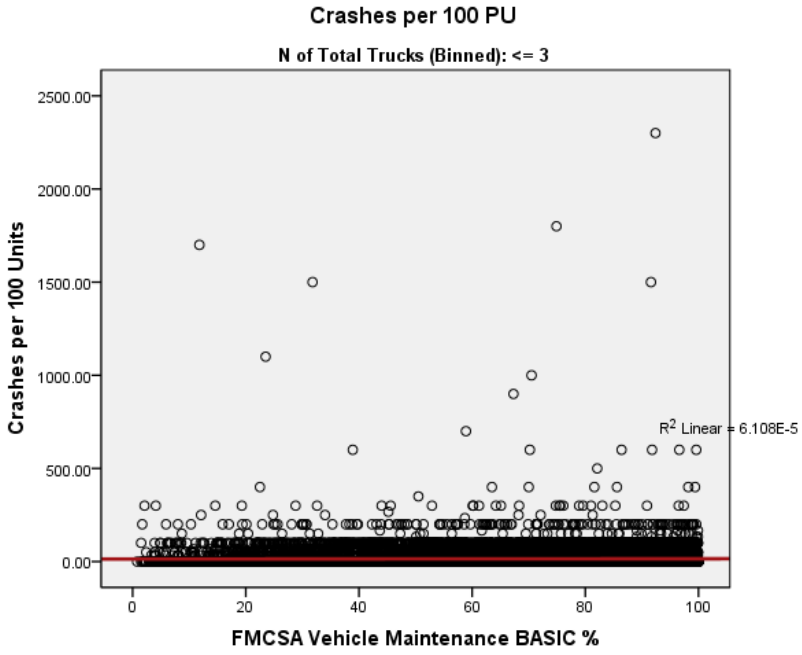

Figure A1. Bivariate Relationship between Maintenance BASIC Score and Crashes per Power Unit, Small Carriers (N=20,210)

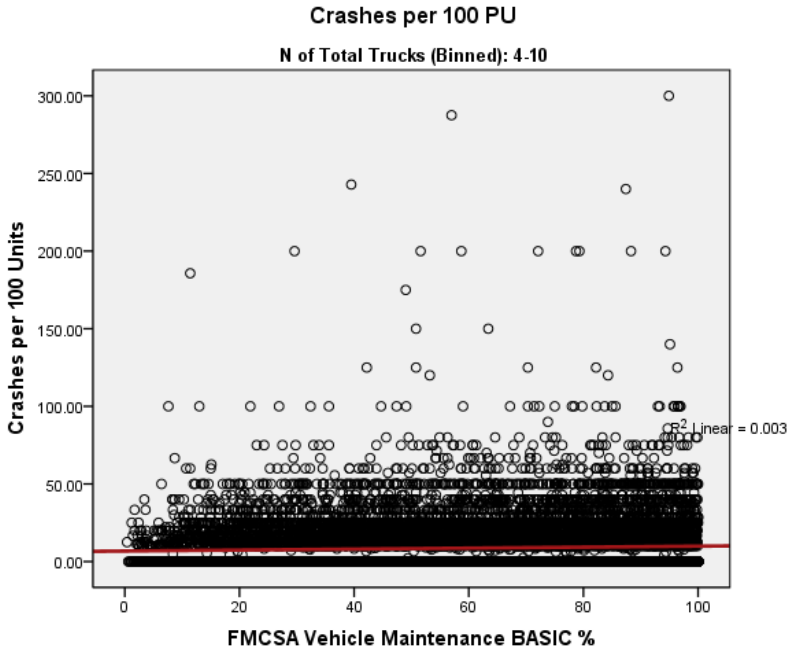

Figure A2. Bivariate Relationship between Maintenance BASIC Score and Crashes per Power Unit, Midsized Carriers (N=18,127)

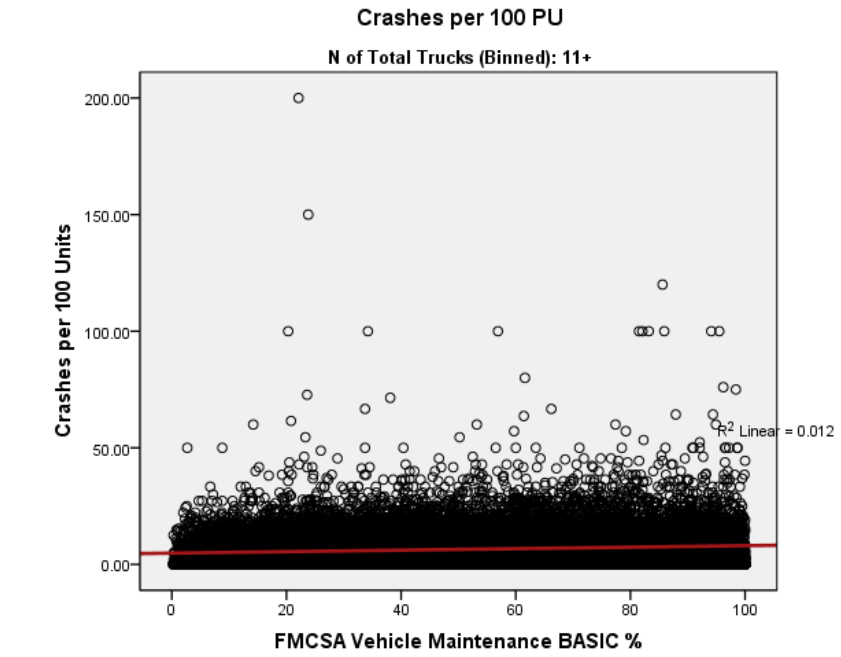

Figure A3. Bivariate Relationship between Maintenance BASIC Score and Crashes per Power Unit, Largest Carriers (N=23,548)

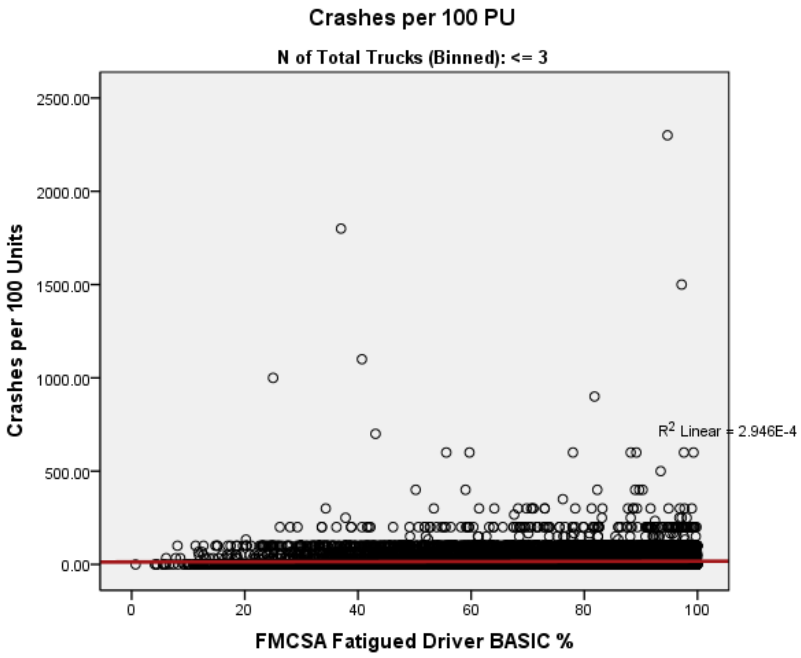

Figure A4. Bivariate Relationship between Fatigued Driver BASIC Score and Crashes per Power Unit, Small Carriers (N=16,482)

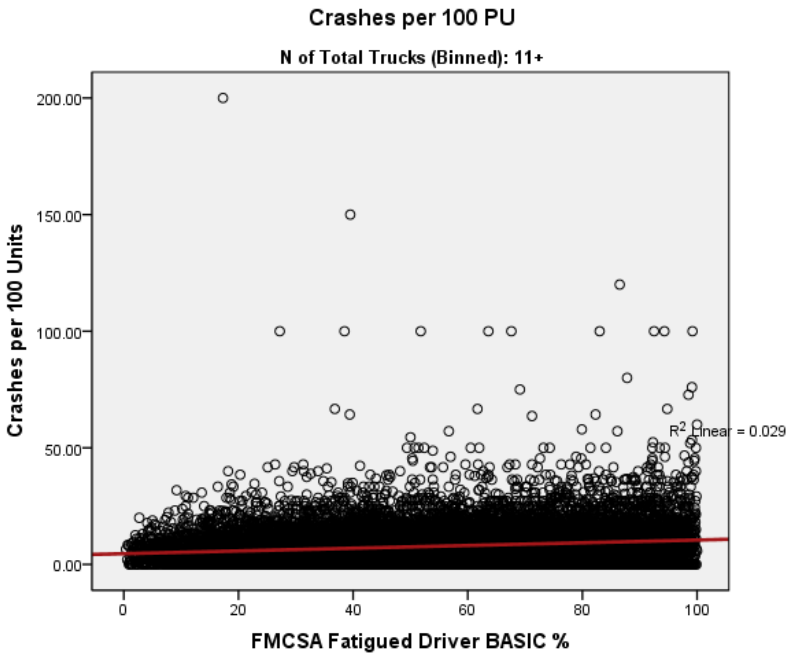

Figure A5. Bivariate Relationship between Fatigued Driver BASIC Score and Crashes per Power Unit, Midsized Carriers (N=11,718)

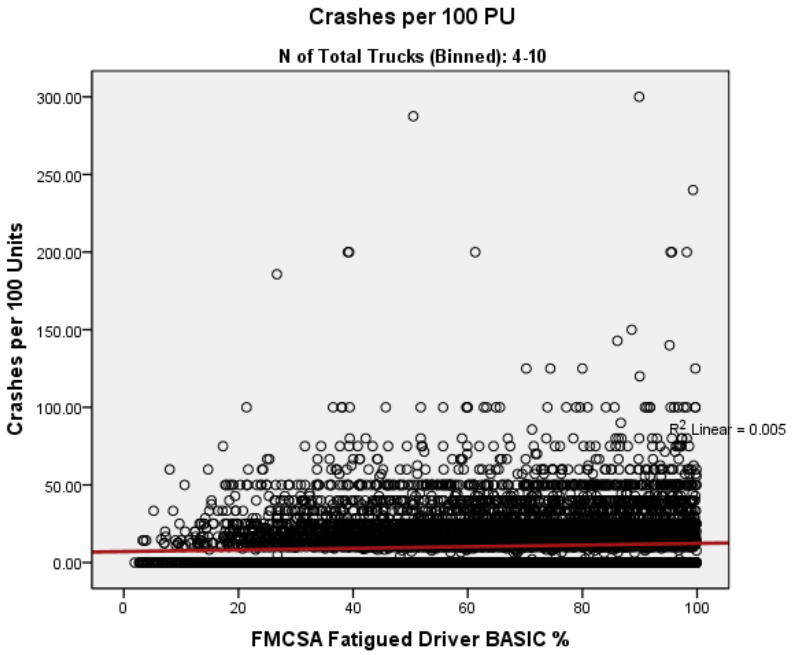

Figure A6. Bivariate Relationship between Fatigued Driver BASIC Score and Crashes per Power Unit, Large Carriers (N=15,031)