Crash Weighting Analysis

Peer Review Comments

Reviewers:
Sam Faucette, Safety Director of Old Dominion Freight
H. Scott Matthews, Carnegie Mellon University
Peter Savolainen, Wayne State University
Bob Scopatz, Independent Researcher
Eric Teoh, Insurance Institute for Highway Safety

January 2015

The Federal Motor Carrier Safety Administration engaged a respected trucking industry consultant, Gene Bergoffen, to help identify independent industry and academic subject matter experts to peer review the Crash Weighting Analysis Report. The reviewers were asked to evaluate the report based on eight criteria. Mr. Bergoffen organized the comments and provided the following summary of peer review findings and recommendations. The peer reviews are compiled in the document below.

Author: Gene Bergoffen, MaineWay Services

Clarity of Hypothesis *–Is the objective and hypothesis clearly stated at the outset, in a manner that enables a logical progression throughout the report?*

- All reviewers found the hypothesis clear, with one suggestion regarding organization of the report

Validity of Research Design

- A number of recommendations and observations were made, with suggestions relating justification of focus on fatal crashes, screening of data sets and attention to carrier concerns regarding weighting of non-assigned crashes
- A number of suggestions made for additional supporting details

Quality of Data Collection Activities

- General support for data collection approach, with some suggestions and a comment.

Robustness and Depth of Analysis Methods

- Reviewers accepted the analysis methods and made some specific suggestions for further justification of approaches uses

Appropriateness of Methods for Hypotheses Being Tested

- Reviewers raised some questions, made some suggestions, and asked for further in-depth justification of methods employed

Extent to Which Conclusions Follow Analysis

- Reviewers generally found the conclusions did follow the analysis
- One reviewer suggested further conclusions might have been drawn from the effort

Strengths and Limitations of the Overall Product

- Over all, the reviewers found value in the product
- A number of specific points were made as to potential limitations

Specific Recommendations for Improvement of the Product

- All reviewers had a number of recommendations for improvement

Author: Sam Faucette, Safety Director of Old Dominion Freight

Clarity of Hypothesis:

1.1 Overview of the SMS- Stakeholders (specifically carriers) are not proponents of Crash Severity Weighting in particular, but modifying the Crash Indicator to reflect causation only. The consensus is a carrier should not be included in Crash Data when exposure is the only contributing factor to involvement.

Validity of Research Design:

The complexity of the Research Design is systematic and comprehensible. The conclusions seem to have a pre-determined focus on existing methods without considering an obvious solution. While PAR accuracy is essential for inclusion, a remedy is not proposed. This being a uniform reporting standard for PAR's.

Screens for Carrier Data Set- as this seems to include the >500 peer grouping (38%) of carriers only. Larger data pools should show the most comprehensive results as is rational. This supports the general SMS issues of a high percentage of carriers/drivers who are not visible or rated.

Quality of Data Collection Activities:

Available data sources are limited but accessed and reflected.

Robustness and Depth of Analysis methods Employed:

Reliability of PAR's- Reporting discrepancies with PAR's relating to the contributing factors could be defined by example. While the weather/road conditions are recorded with accuracy, it is beyond understanding why contributing factors are not.

Appropriateness of Methods for the Hypotheses Being Tested:

Coding Crash Events without a PAR Review using MCMIS data on single vehicle preventable crashes provided a very limited view. The Pre-Crash OOS condition is not explained and gives an impression of a very limited amount of data. Unclear if the OOS condition in the Analysis Results is attributed to the Post Accident inspection as the condition is often the results of damage incurred in the crash.

Extent to Which the Conclusions Follow the Analysis:

PAR Reliability and Sufficiency- emphasis is not placed on causation as should be more focus oriented. While it is understood the purpose of the Crash Weighting Analysis is to improve future crash predictability, the emphasis on causation lacks in detail beyond available data.

Crash Weighting Benefits – weighting crashes when the causation is determined to not attributed to the CMV aligns with current methods of recording crash involvement.

Strengths and Limitations of the Overall Product:

Data is very conclusive demonstrating the difficulty with the hypothesis of assigning crash causation using PAR's. The conclusive number of successful assignments should reinforce the importance of causation vs. involvement but leaves a void that will need to be resolved before dependency with the method can be established.

Specific Recommendations for Improvement of the Product:

Stakeholder concerns should be given due consideration with the inclusion of additional study on crash causation without weighting. While this does not reflect FMCSA's purpose of the Crash Weighting Report, (future crash involvement) the conclusions will need to be solidified and documented. Inclusion of weighted crash involvement without causation does not improve the current method of the SMS Crash Indicator.

Peer group modifications to a more accountable level to the carrier may be helpful.

Vehicle Miles Traveled may have an impact on results in the current formula(s).

Author: H. Scott Matthews, Carnegie Mellon University

1. **Clarity of Hypothesis:** Is the objective and hypothesis clearly stated at the outset, in a manner that enables a logical progression throughout the report?

I find that the overall objective any hypothesis are good, and aside from some organizational issues I mention below, flow logically throughout the report.

2. **Validity of Research Design**

This is a high quality research effort. While I do not doubt the research effort per se, some aspects are not very well justified in the report, and as such might call it in to question. I have identified a few such items, which can likely be addressed fairly cosmetically without new research. Its possible though that my comments are not able to be justified, and thus would need to be corrected.

- Section 4.1.2.4.1 – the report says "There are no drivers in the "Both" group, because no driver in this analysis had more than one fatal crash in the two-year pre-period." Isn't there an important difference regarding whether the pre-period driver **was** the fatality? If so, you wouldn't expect more crashes of course. Was this accounted for? How?

- Section 4.2 – more detail needed in main report (not appendix) about specifics of model used (e.g., equations). Same thing for the crash indicator measures.

- Section 4.3.2.1 – isn't the opposite of the churn rate (those who would fall below threshold?) Which is worse? Both are bad.

- I found the cost assessment to be reasonable, but the presentation of results underwhelming and hard to follow. Table 25 should have 4 rows, not 2 rows with already summed totals separated by slashes. It is hard to follow the "life cycle cost" discussion below without seeing these breakouts.

- Table 26 - Might be more clear (I presume) that you're not saying a QC reviewer only spends 3 minutes looking at each record, but that this represents an allocated amount based on statistical sampling of all codings done? Otherwise it gives the impression that such reviewing is trivial.

- 5.2.2.4 - Why would the US code **all** of anything, as opposed to statistical samples? A benefit could be to suggest what an appropriate sample might be, and show those costs.

- Section 3.2.2.2 – Be more clear what review entailed. Is this the only double-coding done? Were the coded results randomly chosen? Text says "Reviewers agreed". Was this a team of reviewers looking at each PAR, or individuals reviewing each other and reporting back? Did the reviewers otherwise use the same coding process/guidelines?

Did they have the same expertise and training? Was the "match" reported at high level or the justifications level?

3. **Quality of Data Collection Activities:** Have the authors utilized appropriate data collection given the available Federal data sources and the nature of trucking industry information sources.

Yes. I only had one comment about data collection.

- Section 3.1.1.1.1 - Is this ALL accidents involving trucks in both UMTRI and PA, or were these already sampled? I ask because it looks like the PA data isn't comparable to the US level (just 1.5% of crashes – shouldn't it be higher?)

4. **Robustness and Depth of Analysis methods Employed**

The analysis done is of high quality, but could be more robust and could be deeper in places. I give some examples identified below:

- Table 20 – More justification is needed for the weights used in default and in the original and modified cases afterwards. Why +1, etc?

- I would say if weights changed more if you might expect different results. For example, it might be worth doing a "2D" table showing how high the weights need to change to make the percent from baseline numbers go up by bigger (relevant?) amounts.

5. **Appropriateness of Methods for the Hypotheses Being Tested**

The methods seem to be appropriate but should be justified better.

- The report references and uses the "Wilcoxon-Mann-Whitney test" but it has not bee discussed in terms of how it has been used before, how and where it has been applied, etc. This should be added to lend credibility to it. I would also suggest a brief mention in the report as to what kind of results it gives and how they are interpreted.

- Sections 4.1.2– more detail is needed to be able to explain what statistical differences were being assessed. I don't understand how a sentence like "Carriers had the highest distribution of future crash rates" says anything about an assessment of differences. Please define and describe the test done, what was compared, and how the statistical difference is assessed. If necessary, change the bullet point descriptions to match the "Test" text so its clear what differences are being compared and how.

- Section 4.2.2 – You are using a different "Test" than above, and you first need to briefly describe the test, and also convince the reader whether a certain threshold difference in AIC/BIC value is required so as to be deemed relevant. As written, you seem to have judged that 118,753 vs. 118,738 (rows 1 and 4) aren't different enough to be considered

better or worse – yet 118,000 vs. 120,000 was. What is the typical difference needed to make claims? Given this, do your results change?

6. **Extent to Which the Conclusions Follow the Analysis**

The conclusions, in the main report and in the Appendix, in general are representative of the work done.

There are various subsections that include intermediate conclusions, and others that do not. This shoud be standardized, and likely that means to add some sentences of conclusions to those currently missing them.

6.1 - My takeaway of this section in the report was that this was not likely to be sufficient. However the bullets here are a bit generic (avoiding opinion is important, but seems to paint a rosier picture than the section did)

7. **Strengths and Limitations of the Overall Product**

Overall, the written product is very good, but it is not able to be as emphatic and convincing as it could be.

- Add a clear definition of crash weighting in the report and in the executive summary. Perhaps an example from current practice as well. This report basically says "we're trying to see if crash weighting is possible" without defining it.
- End of section 3.2.2.2 - I agree with this intermediate conclusion, but would state it more strongly with the percentages, e.g., that more than half disagreed on whether there was even enough information available to do it. This is a substantial finding that leads to the challenge of doing all of this.
- Under table 12, Need intermediate conclusion sentence like in previous subsections, in this case talking about how this analysis doesn't help support weighting.

8. **Specific Recommendations for Improvement of the Product.**

While this version of the report is of high quality, I would recommend the following changes (I also submitted a marked up Word document with even more specific suggestions not listed here):

- The executive summary excludes the review versus coded comparison (which I found to be quite dramatic). It should be included.
- Section 2.1 is called PAR completeness and accuracy – but it wasn't clear what reported had to do with completeness
- Section 2.2 largely summarizes past work on drivers not carriers

- Section 3.1.1.2 – be more clear in summary field table what was coded versus copy/pasted.
- How was crash severity assessed in the coding? Where was it used after coded?
- It might be useful in Table 4 to note how many matched in all 5 fields (must be small).
- Under Table 4, I'd suggest to add a brief additional comment about traffic-way flow, which is close to 50% and has the same reason as the first two mentioned. Main message is roadway surface and weather are mostly perfect matches, the other 3, decreasingly so, and usually because it's not in the PAR.
- On page 14 when "defining" critical reason you might add examples like you did for critical events
- Table 6 - Is the last row zero, or a "other" category? Text above implies there were multiple ways of getting to an "unknown" kind of result – this implies the answer is zero. Text below implies "other". So Change category name or add rows for detail?
- Text under Table 10, This explanation is good, but given the previous section's sampling review of coding results, its surprising that the match is so high given that NMVCCS had more information and review prospects. You might add a bit more reflecting on what to me is a surprising result (that the coder summary matched NMVCCS much more often than it matched what another randomly done re-coding said).
- Table 11 formatting and legend confusing – see Word document attached
- Text under Table 11 - Add footnote/explanation on whether one should expect detailed "match"? MCMIS is implying preventable/at fault so really it's the 94% vs 6% that matters – the CMV driver/vehicle levels (600 and 30) are just for the readers' benefit and shouldn't really be "matching".
- Section 4.1.1 – for assigned crashes, saying "0 not assigned crashes" is like a double negative. Is there an easier way to say this?

Author: Peter Savolainen, Wayne State University

1. **Clarity of Hypothesis:** The research hypotheses are very clearly outlined. The authors aim to address three primary questions that are of great interest to the truck safety community:

(a) How reliable and sufficient is the information provided in police accident reports?
(b) Would a crash weighting process provide stronger prediction than overall crash involvement?
(c) Procedurally, how could such a crash weighting process be implemented by the FMCSA?

One of the principal motivating factors for this study is to address the perception that some carriers may be unfairly targeted for intervention programs based upon crashes in which the carrier and/or driver was not at fault.

2. **Validity of Research Design:** Generally speaking, the research design allows for direct examination of the three research hypotheses. However, there is less of a direct focus on the concerns of carriers with regard to selection for interventions based upon prior crash involvement. While this concern is addressed implicitly as a part of the analyses (by comparing "Assigned" versus "Not Assigned" crashes), some further analysis or discussion of this issue would help to address this issue explicitly.

3. **Quality of Data Collection Activities:** The authors have appropriately utilized a variety of data sources that provide important information that is necessary to obtain defensible answers to the research questions of interest. The integration of information from the Motor Carrier Management Information System (MCMIS), Fatality Analysis Reporting System (FARS), and the police accident reports (PARs) results in datasets that are quite robust and, as such, the results of the analyses can be generalized to a variety of analytical settings.

4. **Robustness and Depth of Analysis methods Employed:** The authors employ fundamental statistical techniques as a part of their analyses. In most cases, the resultant findings are likely to be relatively robust.

With respect to the reliability and sufficiency of the police accident reports, these techniques are more qualitative and provide an excellent snapshot of the issues inherent with the crash investigation and reporting process, as well as how such issues would impact crash indices. These findings can help to inform subsequent policy and program decisions by FMCSA and other agencies.

The methods utilized as a part of the "crash prediction" analysis are somewhat simplistic. This may be due, at least in part, to limitations as to the level of data available through the MCMIS, FARS, and the PARs. However, some fundamental issues should be clarified as noted in the following section of this review.

The implementation procedures for a crash weighting process are generally well presented. This section would be of interest to agencies beyond FMCSA, as the results may prove useful to a wide range of agencies involved in various aspects of traffic safety. Consequently, further details would be appreciated in this section of the report.

5. **Appropriateness of Methods for the Hypotheses Being Tested:** The formal statistical testing procedures are principally utilized in Section 4 of the report, so these comments apply specifically to that area of the report. The statistical tests/models utilized as a part of this study include the Wilcoxon-Mann-Whitney (MWW) test, as well as the family of negative binomial (NB) regression models.

Ultimately, the objectives of these tests are to compare the crash-involvement rates between four groups of carriers/drivers: (a) those who have been assigned fault in a crash previously; (b) those who have been involved in a crash, but not assigned fault; (c) those who have been involved in multiple crashes, including at least one where they were at fault and one where they were not at fault; and (d) those who have not been involved in a crash.

Both the MWW test and the NB models are of a univariate nature, in that they contrast the rate of crashes (per power unit) among these groups. These techniques inherently assume that these four groups (a through d above) are homogeneous, except for their prior crash experience. This analytical framework essentially allows for a comparison of whether the crash rates in the future tend to be different among the four groups. However, one of the principal concerns to such an analysis from a methodological standpoint is whether this assumption of homogeneity is appropriate. What are the impacts of this assumption? Ultimately, it would be quite useful if details of the carrier and driver populations were provided to inform the reader of the broader context of the analysis. How many total carriers were available in the population that was sampled? How many drivers? What types of carriers and/or drivers would not have been included in the sample?

To this end, there are some qualitative notes of differences between the four groups. For example (pg. 30), the "Both" group has the fewest carriers and tends to include much larger carriers. Given these facts, differences (or lack thereof) between this and other groups may reflect some of these fundamental differences between the groups. Analytically, it would be very useful and interesting to attempt to account for these factors as a part of the analysis, or at least include a discussion of aggregate-level differences between the groups. Were carrier segment (straight truck versus combination) and truck utilization (vehicle miles traveled per power unit) examined directly? If not, could they be? Computing crash rates per power unit makes the implicit assumption that the travel rates (VMT per power unit) are consistent across the four groups.

It is suggested that the crash rates are explicitly presented in the report (e.g., in Table 13). Calculation of the rates based upon the information from Table 13 shows the rates among the

first two groups are virtually indistinguishable (0.05255 future crashes per power unit vs. 0.05205) and the difference from the third group is also marginal (0.05119). The fourth group has a lower rate (0.04730), but there are some potential concerns as to differences in the characteristics of this group and the other groups. The same comments apply for the fatal crash analysis presented subsequently. It seems that this group (no crash involvement) would tend to include a disproportionate number of smaller carriers (who would be involved in less travel than the larger carriers). Are any data available to assess differences between these groups? The a priori assumption of using the MWW test is that these groups are homogenous, except for the group variable (i.e., whether they have assigned crashes or not).

Pg. 31 – Footnote 37 is difficult to follow. The Wilcoxon-Mann-Whitney test is a non-parametric alternative to the t-test. I assume the authors are just clarifying that the average of the carrier rates are being compared (as opposed to the overall average rates when combining all data within the groups). Ultimately, I would suspect that the results of the MWW test are not substantively different than those of an independent sample t-test.

How do the groups of carriers and drivers vary? Specifically, there were 108,824 future crashes among the carriers examined, but only 3,382 among the drivers. What are the differences with respect to the characteristics of these groups? Are certain types of carriers (or drivers) overrepresented?

Some revisions to the technical discussion of the NB model are recommended. Instead of referring to the "number of successes", it is advised to simply refer to the number of crashes. It should also be noted that this variable (r) would actually the number of crashes, not the rate of crashes. If the models were estimated correctly, the exposure (i.e., volume, fleet size, etc.) would be used as an offset (a covariate with a coefficient constrained to equal one). Presumably, this is how the models were estimated, but no details are provided to substantiate this.

For the negative binomial models, AIC and BIC values are presented (Tables 17-19). However, these are not particularly meaningful to the reader in a practical sense. What about the parameter estimates? Continuing, the use of the NB model assumes that the crash data are overdispersed (i.e., the variance of the crash counts is greater than the mean). Is this assumption valid? It seems that this would depend largely upon what level the data are analyzed at. For example, at the individual driver level, the data are probably more likely be underdispersed.

6. **Extent to Which the Conclusions Follow the Analysis:** The conclusions generally follow the analysis results. Specifically, the three stated hypotheses are directly addressed and discussed in the analysis and conclusions sections of the report. While there are some concerns as to the underlying assumptions and methods, addressing these issues would help to reinforce these conclusions.

7. **Strengths and Limitations of the Overall Product:** Overall, the report provides a lot of information that is very useful to the FMCSA and others in the traffic safety community. In its

current form, the strengths of the report are particularly the reliability/sufficiency assessment for the PARS, followed by the development of a procedures and processes for implementing a crash weighting system. The statistical methods employed as a part of the crash weighting process are the principal limitation as discussed above.

8. **Specific Recommendations for Improvement of the Product:** The following are general comments beyond the technical issues that have been discussed previously.

In the Comparison of Data Fields (beginning on pg. 12), how is the "Driver Contributing Factors" category assessed? There can be up to four contributing factors. Do all four need to match? It appears that the PAR is particularly deficient in this area, which is not surprising. While the authors discuss some of the reasons for this discrepancy, it is unclear exactly how a «match» was discerned.

Missing data appears to be a problem in various aspects of this study. Is imputation of values a feasible alternative in such instances? For example, the conclusions from Section 3 note that the PARs often lack sufficient information for analysis purposes.

It is not surprising that All Crash Involvement is a better predicted than Single-Vehicle Preventable or Coded Assigned Crashes. This is likely to largely a byproduct of the larger sample, which results in reduced standard errors.

In section 4.1.2.1.2, the authors note which carriers had the "highest distribution of future crash rates". This seems as if it would be more appropriately stated with respect to which carriers tended to experience higher crash rates (i.e., the term "highest distribution" is nebulous).

The crash weighting determination process is interesting, particularly as it relates to the economic estimates of implementing this process. Further details as to how the underlying cost estimates were obtained would be useful to the reader.

Was the SMS Crash Indicator Measure developed empirically? How were the weighting factors determined? It would be interesting to estimate a model, based upon carrier/fleet characteristics, that was directly tied into «future» crash data.

The authors note that on pg. 40 that "...the fatal crash model shows an improved Crash Indicator...". However, this is not found to be the case for all crashes. Given the sample size differences, it appears that much of this «improvement» in the fatal crash model is just capturing the greater uncertainty given a more limited sample size. Ultimately, some clarification is needed as to what the practical results of these "Effectiveness Tests" actually mean.

Minor Notes:

Pg. x – There is some ambiguity with respect to comparing fields “on the PAR” with fields “in the FARS”. Ultimately, FARS is based upon PARs (which is described elsewhere). This point could be clarified here.

Pg. xiv – The authors refer to the MWW test as a "statistical model", it would be more aptly described as a test than a model.

Traffic-way flow – What does this refer to? Does this designate a flow rate? A type of road facility? A type/directional designation? I assume this is consistent with FARS, which would indicate this field designates direction, whether the road is divided and/or includes a median, etc. This should be clarified throughout the document to improve readability.

Pg. 10 – I would advise defining FARS here (though I acknowledge this is done elsewhere in the report).

At several points in the manuscript, it is stated that “The study did…” or “The study used…” Technically, the study did not “do” or “use” these things. These should be simple syntax fixes.

Author: Robert A. Scopatz, Independent Researcher

This review follows the outline provided to me via email by Gene Bergoffen. The review's objective and key points to address are listed in the text box below:

Objective: To enable the Author(s) to improve the product

1. **Clarity of Hypothesis:** Is the objective and hypothesis clearly stated at the outset, in a manner that enables a logical progression throughout the report
2. **Validity of Research Design**
3. **Quality of Data Collection Activities:** Have the authors utilized appropriate data collection given the available Federal data sources and the nature of trucking industry information sources.
4. **Robustness and Depth of Analysis methods Employed**
5. **Appropriateness of Methods for the Hypotheses Being Tested**
6. **Extent to Which the Conclusions Follow the Analysis**
7. **Strengths and Limitations of the Overall Product**
8. **Specific Recommendations for Improvement of the Product**

SUMMARY

This study was conducted to evaluate the differential ability to predict future crash involvement by motor carriers and drivers when past crashes were (a) used as a predictive variable without reference to the carrier's/driver's role (responsibility) versus (b) first filtered so that only crashes in which a positive determination could be made that carrier/driver bore some assignable responsibility for the crash. The data used in the study were a combination of police accident reports (PARs) obtained from the states as part of the 2005-2007 National Motor Vehicle Crash Causation Study (NMVCCS) and the Fatality Analysis Reporting System (FARS) for 2008-2010, along with crash and inspection records from the Motor Carrier Management Information System (MCMIS). A total of 10,892 PARs were examined and, using the critical reason coding method from the Large Truck Crash Causation Study (LTCCS), assigned a critical reason in all but 304 (2.8%) of cases. A second coding method used MCMIS data only and assigned the critical reason to the commercial motor vehicle (CMV) if (a) it was a single-vehicle "preventable" crash, or (b) a post-crash inspection revealed that there was a pre-crash out-of-service condition.

Any crash in which the CMV was assigned the critical reason (through either method) is termed an "Assigned" crash for purposes of analysis. All other crashes were termed "Not Assigned". Crash "weighting" in various analyses could be all-or-none (a weight of 1 versus 0 depending on whether the crash was "assigned" or "not assigned" or indexed such that assigned crashes were given a higher severity rating than not assigned crashes, but in which all crashes had a non-zero weighting.

In a series of analyses, the researchers assessed the reliability and sufficiency of PARs as a basis for crash weighting, the sufficiency of an automated crash weighting determination based on existing data in MCMIS, the potential benefits of crash weighting for predicting future crashes, and practical methods
for implementing crash weighting in the FMCSA Safety Management System (SMS). The research team also developed an estimated cost for implementing and maintaining a crash weighting system.

This research report is thorough in that it attempts to quantify three major factors that would affect the decision of whether or not to implement crash weighting as a method of judging motor carriers' and drivers' safety. Those factors are: the ability to use available data (the PARs and/or MCMIS) to
determine if a particular crash should weigh against a carrier or driver; the ability to reliably assign a crash to the CMV; and the benefits and costs of weighting.

The review that follows addresses the requested assessment factors in the sequence shown in the text box.

REVIEW

1. Clarity of Hypothesis

The question asked in the review criteria is: *"Is the objective and hypothesis clearly stated at the outset, in a manner that enables a logical progression throughout the report?"* On page 3 of the document, Section 1.2 (Report Scope) the research aims are clearly stated, as follows:

- Do Police Accident Reports (PARs) provide sufficient, consistent, and reliable information to support crash weighting determinations?
- Would a crash weighting determination process offer an even stronger predictor of crash risk than overall crash involvement, and how would crash weighting be implemented in the SMS?
- Depending upon the analysis results for the questions above, how might FMCSA manage the process for making crash weighting determinations, including public input to the process?

Though not explicitly labeled "hypotheses", these three questions tell the reader exactly what the study is designed to accomplish. These are easily understood as hypotheses in the traditional sense.

In addition, the three research questions provide the organizational basis for the remainder of the document, making it obvious which question is being addressed in each subsequent section.

The only remaining part of this first criterion, then, is whether or not there is a *logical* progression through the three questions and thus, through the document. Again, the answer is "yes." As noted in my summary, there are three factors that should be addressed in support of the decision whether or not to move to a crash weighting scheme based on the CMV's role in and responsibility for a crash. These are:

A. Can we achieve a reliable determination, based on available data, of when a crash-involved driver, vehicle, or motor carrier should be assigned the critical role in (and thus responsibility for) the crash?

B. Does weighting crashes in this manner result in improvements in our ability to characterize the safety of motor carriers and drivers?

C. Do the benefits of weighting (in terms of improved safety analysis) outweigh the costs of implementation and maintenance?

The report is sequentially organized to address these three decision factors.

2. Validity of Research Design

This question must be answered in sections corresponding to each of the analyses conducted. In this review, I have labeled each according to the section number of the report in which the research design (analytic approach) is presented.

Section 3: PAR reliability and sufficiency

Overall, the approach is heavily influenced (10,505 of the total 10,892 records) by a comparison of original PAR data to the record ultimately created as a result of the FARS coding process. To a lesser extent, NMVCCS data (387 records) influence the analyses and conclusions throughout this section (3.1 to 3.3). The FARS coding process, and likewise the NMVCCS process, were created because PAR data alone are not sufficient to characterize crashes to support robust analysis. First and foremost, the crash records are not standardized among the States so building a national dataset requires interpretation and recoding. It is also well known that FARS coders (and one assumes the NMVCCS coders as well) frequently find errors in the PARs and must work to complete a case accurately and within applicable data quality standards. The same is true for data entered via the SafetyNet system that serves as the core crash data in MCMIS.

It also must be recognized that crash reports of fatalities are given the greatest level of attention among police reported crashes. That means that the officers and supervisors scrutinize the reports and attempt to ensure that the information is accurate and complete. In many jurisdictions an accident reconstruction team is deployed and the PAR may either be influenced by or completely derived from their efforts.

In short, there is good reason to expect differences between any PAR and the data ultimately captured in a national database, but, at least for fatal crashes, there is also reason to expect that the original PAR would be of the highest quality possible. This has implications both good and bad for the research designs throughout Section 3.

3.1: Reliability of PARS

The reliability analysis showed that there is a low match for selected key fields on the crash report in comparison to the data coded in FARS. A large part of the two sources match for items like Driver Contributing Factors, First Harmful Event, and Traffic-way Flow; however, is due to a lack of data on the PARs. It's not that there are errors on the PARs, but that, in the researchers' judgment, the required information is not present (and thus the data do not match**). The methodology for how this comparison was accomplished is not specified**. Was it automated (in which case there is no wonder that it failed—data integration across multiple states' crash data into a single standard has been tried multiple times over the years, each time ending in at best partial success. If the comparison was based on human judgment, then the method needs to employ some measures of inter-rater reliability and criterion-based training in order to assure the reader that it was conducted with measurable standardization.

In addition to the above, the prior material on this analysis mentions both NMVCCS and MCMIS, but **the comparisons between PAR data and these two datasets is not presented or mentioned in sections 3.1.1.4 (comparison of data fields) or 3.1.2 (analysis results).**

Overall, this analysis is not compelling in disputing the reliability of PARS. It does show clearly, however, that a simplistic method of matching crash data elements to the FARS records is not workable. By extension one could reasonably conclude that a crash weighting scheme that did not involve human decision makers (coders) would fail. The PARs being examined are known to be the best quality received from law enforcement. If an automated coding system would not work with reports of fatal crashes, it would fail even more spectacularly with data collected at the scene of less severe crashes.

Recommendations:

- Add in the specifics of how the match between PAR data and the FARS records was conducted. If automated, state that clearly. If it involved human judgment specify how the people doing the match were trained and tested for inter-rater agreement.
- Say more about the data collection tool. In a typical FARS case, the coder is well trained to identify items that this research says are largely "missing" from the PARs. In fact, FARS coders get a great deal of this information from the

PARs but they do so by reading the narrative and diagram, and interpreting the facts collected throughout the crash report form. They have other sources of information as well. A key question here is: Did the researchers make use of any FARS coders to advise them on how to translate from PAR to FARS? This is not to say that an automated coding scheme would work—it will not or we wouldn't need FARS analysts. The point is, however, that this analysis may have failed for uninteresting reasons such as the failure of the data collection tool designers to understand the nuances of each state's crash report form.

- Make the larger point much more explicitly. An attempt was made to automate the coding of crash reports to test for reliability of key fields that would be part of a determination of "assignability" was made. It failed to match what trained humans (the FARS analysts) can do with the same data resource. Therefore, an automated tool for assigning crashes is unlikely to work. If the best quality crash reports (PARs from fatal crashes) can't be reliably interpreted through automated means, we shouldn't even bother trying with reports of less severe crashes.

3.2: Sufficiency for making crash weighting determinations

The methodology for weighting crashes (i.e., assigning a critical reason for the crash) is well described in Section 3.2.1.2. The methodology says "All of the coders that reviewed the PARs were experienced in using the LTCCS methodology to make critical reason determinations." This is vague. Were they all coders on the LTCCS project? Were they equally experienced in coding cases in LTCCS? Or, were they merely "users" of the LTCCS information and thus experienced in that they understood the coding process?

Another missing piece of information is whether or not the researchers conducted a quality review of the experienced coders **to determine the level of inter-rater reliability prior to using the resulting assignments**. This should have been done for a randomly selected portion of crash reports and any coder who displayed a low agreement with other experienced coders should have been given remedial training or dropped from the study. This is a standard practice when human judgments are involved in created a research data set.

It appears that the research team used themselves as "reviewers" in a post-hoc attempt to assess the quality of the coding done by the experienced raters. The reasons for this choice should be explained—that is, why did they use post-hoc independent review as a QC step, and how does it substitute for measuring inter-rater reliability and criterion-based training prior to the coding?

The comparison of critical reason determinations for the cases that were matched to the NMVCCS is a good analysis. It shows that the trained coders in this study agreed about 90% of the time with the original coders in the NMVCCS. That is not a bad level of agreement and lends some credence to the assertion that this study's coders were experienced and well trained. It is somewhat concerning, however, that the false alarm rate (all but the highlighted cell of the

"Truck/Bus Driver" column of Table 10) is higher than the miss rate (all but the highlighted cell in the "Not Assigned to This Truck/Bus" column of Table 10). This is an indication of bias toward assigning reasons to the truck/bus driver in the current study…at least in comparison to the NMVCCS. That is somewhat worrisome in that it may indicate that the coders were somehow influenced by the knowledge of the aims of this current research effort.

Overall, this value of this analysis hinges directly on the skill of the coders. Taking as a given that they were all experienced coders and would have agreed substantially on assignments of the critical reasons, this analysis offers compelling proof that trained human judges can sufficiently interpret crash reports to support a crash weighting program. While the post-hoc QC analysis helps to show that that the assignments are probably reasonable, such a review can be easily compromised by researchers' expectations. The reviewers had the coders' assessments in front of them when deciding if those codes were valid, so this was not an independent QC step, but merely a confirmation check for reasonableness (face validity). The comparison to NMVCCS coding is reassuring that this study did in fact use experienced coders; however, the authors need to explain whether the bias evidenced in the relationship between false alarms and misses shown in Table 10 is important to their ability to interpret the data. The bottom line, however, is that this analysis is compelling proof that human observers can code crash weights based on the PAR alone.

Recommendations:

- Describe the coders' level of experience. How many coders were there? What is their background and experience with LTCCS? Describe the training in more detail. Was it criterion-based training?
- Describe the quality control processes, specifically any inter-rater reliability testing that preceded the assignment phase of this analysis. Additionally, were there any inter-rater reliability checks during the assignment (coding) phase? Present the values for inter- rater reliability in the text.
- Explain the choice of using post-hoc review rather than criterion-based training of coders and testing for inter-rater reliability.

- Address the apparent bias (in comparison to NMVCCS coders) in the current studies' coders as evidenced in Table 10. How does this affect the overall interpretation of the results (if at all)?

3.3: Coding of crash events without a PAR review

In this analysis, the researchers reasoned that it might be possible to assign crashes based on the information in MCMIS for each crash. Two classes of crashes were determined to be "assignable": (1) "preventable" single vehicle crashes—where preventable means that there were no obvious external causes for the crash; and (2) crashes in which a post-crash inspection revealed a pre-existing out-of-service condition for the CMV. Where a match between the PAR and MCMIS could be obtained (1,438 cases), the coded assignments to the CMV from analysis 3.2 were compared to an automated selection from MCMIS. The results show that the automated assignment matched the human assignments 94% of the time for single-vehicle crashes, but less than half the time for crashes where the CMV had a pre-existing out-of-service condition. This is explained as likely resulting from the fact that the coders in analysis 3.2 did not have access to the post-crash inspection information. This latter analysis, however, would only have been compelling if a high percentage of agreement had been achieved. A low agreement (as found here) means nothing with respect to the value/sufficiency of the PARs for coding crash assignments. There are multiple types of out-of-service violations and, just as a drunk driver may not actually be at fault in a crash (despite the pre-existing violation), an out-of-service driver or vehicle may have not caused the crash in which they were involved.

Overall, the analysis of single vehicle crashes is more compelling than the out-of-service analysis. However, the truth is that automated assignment of crashes is unlikely to succeed. MCMIS simply doesn't have all the right data fields to make this determination, and we've already seen (in 3.1) that automated interpretation of crash report data is a difficult undertaking that meets with only limited success.

Recommendations:

- No recommendation. This analysis had to be attempted, but its results were predictable and so it serves a purpose of showing that an automated method based on MCMIS data would not be sufficient.

Section 4: Crash Weighting Benefits

The analyses in this section are designed to quantify the improvement in crash prediction when past crash assignments (crash weight) is known. The weighted crashes are the same as those from Section 3. The Mann Whitney U test (aka Wilcoxin-Mann-

Whitney), negative binomial regression, and a calculated metric are used, respectively in the three analyses presented in Section 4.

4.1: Comparing future crashes using crash weighting

This pre-/post-analysis compares four groups of carriers defined based on the crash assignment processes described in Section 3. MCMIS data were filtered to identify carriers that had at least one crash in the pre-period (2009-2010) along with other selection criteria spelled out in the methodology. Carriers were grouped by whether they also had been involved in one of the assigned crashes and whether or not the assignment identified the carrier (i.e., the carrier was assigned the critical role or not). From among the group of assigned crashes, some carriers (Assigned Group) had only assigned crashes, some (Not Assigned Group) had only crashes which were ultimately not assigned to the carrier, some (Both Group) had crashes of both types, and a final group (Neither Group) had neither assigned or not-assigned crash experience. It must be recognized that the carriers in all the groups may also have had crashes that were not part of the assignment process.

The statistical test is based on average rank orderings of the carriers' crash experience in 2011 and 2012 and asks the question: Do the carrier groups defined by crash assignment differ in terms of their rank-ordered crash rates (total crashes/#power units). This analysis was conducted twice: first for all crashes regardless of severity and then for fatal crashes only.

To understand the problem in this analysis, consider that crashes are low probability events and fatal crashes are even more rare. The election of fatal crashes as the standard data set for assignment means that whenever a carrier happened to fall into the Both Group it was almost guaranteed to have a high crash rate. This result was obtained in the two carrier analyses (All crashes and fatal crashes only) but not for the driver analyses (which generally did not show any significant findings.

Overall, a better way to conduct this analysis would have been to compare only carriers with assigned and not assigned status, leaving out the Neither group altogether. Also, it would have improved validity to conduct a new crash assignment process based on a random selection of crashes from all levels of severity. Section 3 proved that the assignment process worked. Section 4 could then have used a broad selection of crashes to avoid the problem associated with only looking at fatal crashes (for the vast majority of cases). It is unlikely that the project budget would have borne the cost of another independent data gathering and coding process. So, this analysis is likely as good as one can get within a reasonable cost and time frame.

Recommendations:

- None. This analysis is technically well designed. It suffers from a potential bias due to the use of fatal crashes as the majority of the assignment data set, but that is unavoidable.

4.2 Predicting future crashes using crash weighting

This analysis used negative binomial regression to assess the relative predictive value of models of future crash rates (crashes per power unit) and driver crash counts considering (alone and in combination) all crashes; single-vehicle preventable crashes; coded assigned crashes; and coded not assigned crashes. The various possible models were compared using two standard methods (AIC and BIC) to identify the strongest models. The analysis was done once for all crashes and again for fatal crashes only. The results show that the strongest models did not need to take into consideration the coded crashes. Using MCMIS data to identify the preventable single-vehicle crashes helped improve the model for all crashes and for the driver analysis.

It should be noted that because the coded crashes comprised the fatal crashes in the pre-period, this analysis logically boils down to asking whether or not knowing that a carrier was responsible (in some tangible way) for a fatal crash in 2009-2010 adds predictive ability beyond just knowing their overall crash rate in that same period. It would be surprising to find the result that the model was sensitive to this. Fatal crashes are thought to be random events in most respects (e.g., why did the crash result in a fatality versus serious injury only) but that the likelihood of being involved in a fatal crash should track somewhat to a carrier's overall exposure to crashes. That is a function of their overall safety performance (how crash prone they are) and the number of miles they drive (for which number of power units serves as a surrogate). In short, the "assigned code" crash category doesn't really add information to the analysis because the predictive value of "carrier A was involved in a fatal crash" is already captured to some extent in their overall crash rate.

Overall this analysis does not represent a fair test of the hypothesis that weighting crashes adds predictive value. It is practically the same as an analysis that looks at the value of adding "fatal crash involvement" to an analysis of the predictive power of knowing overall crash involvement. This is a direct consequence of having used FARS cases for the majority of the coded cases for assignment. It is interesting to note that, where it was tested, the only unambiguous notation of "carrier/driver at fault" (the single vehicle preventable crashes) does add strength to the model. Recall, however, that single vehicle crashes are a small percentage of all crashes.

Recommendations:

- Restate the findings to show that there is only one fair test of the value of knowing crash assignment—that is the test using the single-vehicle

preventable crashes. The analysis of "coded" crashes is not a fair test of the value of knowing that a crash was assigned to a carrier for the reasons stated.

- Run a second analysis predicting ONLY single vehicle crashes (crash rate or involvement) in the post-period. Examine the value of knowing pre-period single- vehicle preventable crashes on the ability to predict post-period single-vehicle crashes. **That analysis should give FMCSA the *upper bound* on the value of crash weighting.** If it adds a great deal of strength to the statistical model that incorporates "all crashes" then that should be the basis for determining the value of crash weighting (assuming one could do universal crash weighting for all crash types and severity).

4.3: Implementing crash weighting in the SMS

After reading this analysis section several times, it is not entirely clear what is being presented in Tables 23 and 24. The percentages are described as "percent improvement for both crash weighting methods compared to the baseline model" but I don't actually see how the new weightings are independent of the old method that already adds increased weight for fatal crashes. This new model appears to simply make assigned fatal crashes much more important while leaving the base model's weight for fatal crashes intact. Non-assignment could reduce the penalty for fatal crash involvement (by 25% from the base model or removing those crashes altogether), whereas assignment now boosts the crash value by a large additional weight (or not at all in the removal version of the analysis). This amounts to a marginal change in the base model, and, unsurprisingly, a marginal improvement in the model's predictive value seems to have accrued. The improvement is larger for fatal crash analyses only, as one would expect because of the use of fatal crashes as the basis for the assignments.

Recommendations:

- Provide a more complete explanation of the analysis and meaning of Tables 23 and 24. The description is not sufficient to ensure that the reader fully understands what the analysis is showing.
- Run another analysis where a similar weighting scheme is used but for single-vehicle preventable crashes. This was the truly important result of Analysis 4.2 (that these assignments were potentially valuable). It isn't clear why they weren't the subject of an alternative weighting analysis in 4.3.

Section 5: Implementation of a crash weighting determination process

The cost analyses here appear valid and present reasonable estimates of what a manual process would involve. I have no additional comments on them other than that the time per PAR estimates appear to be in line with what I know to be the case for manual data management processes in several states. They may, in fact, be underestimates of the real

costs because it there is no cost shown for training coders. Assuming some reasonable level of turnover, that would be a recurring cost that could be quite high.

Recommendations:

- Address the start-up and recurring costs of training under various models such as (a) States code all the crashes; (b) a centralized staff (contractor?) codes all the crashes for the entire US.

3. Quality of Data Collection Activities

Data collection activities were well conducted. There are serious implications throughout of using (mostly) fatal crashes for the dataset of weighted crashes. In hindsight, a better approach might have been to select a small number of FARS and NMVCCS cases to use in validating a coding method, and then taking a truly random sample of all cases entered into SafetyNet for 2009-2010. This would have avoided the problems in Section 4 with models that are not logically different from a model based on Fatal crashes as predictors of future crash experience. Unfortunately, that would have raised the cost of the study significantly and would have raised the potential for other problems.

The researchers hit on a very interesting idea in examining single-vehicle preventable crash involvement as a surrogate for weighting. This was validated in Section 3 and used effectively in some of the modeling in Section 4. Unfortunately, the implications of this insight were not fully explored in the analyses. Recommendations under Section 4 may provide a cost-effective way to test the value to be extracted from this class of crashes.

4. Robustness and Depth of Analysis Methods Employed

Comments provided in item 2 earlier in this review address some of the limitations of the data analyses. It should be noted, however, that the statistical tests chosen in this study are valid and reasonable choices. In particular, the Wilcoxin-Mann-Whitney and Negative Binomial Regression statistical tests are good choices given the nature of the data. The analyses could be enhanced with additional tests (as suggested under Item 2).

I would recommend those additional tests, but no changes in the statistical analytic methods used to evaluate significance. However, and throughout the paper, the values of the statistical test results are *not* presented in situ (they are presented in Appendix D). This is not standard practice and it is equally not standard to say something reached statistical significance without reporting the alpha level (or the probability of the test statistic in the body of the paper. In some cases the data are not summarized in a table either so the reader has no idea what the test showed or what the summary values were without reference to the Appendix.

Recommendations:

- For all statistical test results, provide the value of the test statistic and the associated p value or indicate the alpha level in the body of the paper. This should be true for all comparison values as well.

5. Appropriateness of Methods for the Hypotheses Being Tested

This issue can be addressed generally with a note of praise for the researchers in showing resourcefulness and clever use of the available data. The use of FARS and NMVCCS data to assess the reliability and sufficiency of PARs (Section 3 of the Report) and in assessing the reliability of the chosen coding/assignment/weighting process is noteworthy. Ultimately, I think the decision to use a dataset mostly comprised of fatal crash records hampered the study's utility, but it does by no means destroy the value of the study. Far from it. The researchers had a choice to make between demonstrable reliability of their coding choices and the effect sizes they could expect in statistical testing. They didn't have MORE data sources that could have been used to provide valid prior coding. They also didn't have easily accessible resources of PARs for non-fatal crashes. The better choices would have been much more costly.

The current study leaves some room for further study or improvement. However, the hypotheses stated at the outset were testable with empirical methods and the research did result in valid quantifiable measures of the cost and benefit of weighting crashes by the role the carrier's driver and vehicle played in the event.

6. Extent to Which the Conclusions Follow the Analysis

The study is presented in a logical sequence and generally the conclusions do not "go beyond" the strength of the analysis, the data, or the results. In particular, I agree with the frequently stated caveat that the analyses depended to a great extent on fatal crashes. The caveat regarding generalizing the SMS analysis presented in Section 4.3 is also well stated and true.

I believe, however, that the researchers stopped short of findings that may well have been within their grasp. In particular, they did not pursue as much as I would have hoped the notion that single-vehicle preventable crashes have some predictive value. They didn't include this as a factor in the analysis in Section 4.3. Nor did they pursue the logical step of trying to predict future single-vehicle preventable crashes as a "best case" test of the value of crash weighting. I believe the additional analyses would be valuable and not too difficult to conduct given the already available datasets. The only thing that would have to be accomplished would be to create the post-period database of preventable single-vehicle crashes for use in additional analyses in Sections 4.2 and 4.3. I hope there is sufficient budget to accomplish this because, frankly, I think there is an important point to be quantified: *what is the most improvement we could expect from the*

process of coding ALL crashes? That datum is needed in order to do a complete job of comparing benefits to costs.

7. Strengths and Limitations of the Overall Product

My previous comments address this point-by-point for each analysis and overall for the report. Overall I would say that the strengths of this report are in its methodical and thorough approach to quantifying the improvements to be gained from assigning crash weights based on the carrier's role in the crash event. The methods are valid and reliability of the coding assignments was demonstrated adequately. While I have recommendations there, for the most part they relate to providing more information to the reader.

As far as weaknesses, the biggest one is the use of fatal crashes as the primary source of crash weighting assignment. It is completely understandable and, from a cost perspective, entirely reasonable. There are unfortunately consequences of this choice. Rather than mar the overall utility of the report for decision makers, however, the choices must be understood in the context of how costly and difficult it would be to conduct a better study and, ultimately, how uncertain the gains in explanatory power might be.

That is why I urge the completion of additional analysis in Section 4.2 and 4.3 that build on the insightful finding that single-vehicle preventable crashes are logically assignable. Section 3 showed that the automated assignment of these crashes agreed with the coders to a greater extent than the post-hoc review did. In other words, there is one clear case where automated assignment was roughly as good as the human coders. Thus, this is one portion of the data where the value of coding can be explored more deeply for crashes at all levels of severity, not just fatal crashes.

8. Specific Recommendations for Improvement of the Product

Section 3 of the report:

- Add in the specifics of how the match between PAR data and the FARS records was conducted. If automated, state that clearly. If it involved human judgment specify how the people doing the match were trained and tested for inter-rater agreement.
- Say more about the data collection tool. In a typical FARS case, the coder is well trained to identify items that this research says are largely "missing" from the PARs. In fact, FARS coders get a great deal of this information from the PARs but they do so by reading the narrative and diagram, and interpreting the

facts collected throughout the crash report form. They have other sources of information as well. A key question here is: Did the researchers make use of any FARS coders to advise them on how to translate from PAR to FARS?
This is not to say that an automated coding scheme would work—it will not or we wouldn't need FARS analysts. The point is, however, that this analysis may have failed for uninteresting reasons such as the failure of the data collection tool designers to understand the nuances of each state's crash report form.

- Make the larger point much more explicitly. An attempt was made to automate the coding of crash reports to test for reliability of key fields that would be part of a determination of "assignability" was made. It failed to match what trained humans (the FARS analysts) can do with the same data resource. Therefore, an automated tool for assigning crashes is unlikely to work. If the best quality crash reports (PARs from fatal crashes) can't be reliably interpreted through automated means, we shouldn't even bother trying with reports of less severe crashes.
- Describe the coders' level of experience. How many coders were there? What is their background and experience with LTCCS? Describe the training in more detail. Was it criterion-based training?
- Describe the quality control processes, specifically any inter-rater reliability testing that preceded the assignment phase of this analysis. Additionally, were there any inter-rater reliability checks during the assignment (coding) phase? Present the values for inter- rater reliability in the text.
- Explain the choice of using post-hoc review rather than criterion-based training of coders and testing for inter-rater reliability.
- Address the apparent bias (in comparison to NMVCCS coders) in the current studies' coders as evidenced in Table 10. How does this affect the overall interpretation of the results (if at all)?
- No recommendation. This analysis had to be attempted, but its results were predictable and so it serves a purpose of showing that an automated method based on MCMIS data would not be sufficient.

Section 4 of the report:

- Restate the findings (in 4.2) to show that there is only one fair test of the value of knowing crash assignment—that is the test using the single-vehicle preventable crashes. The analysis of "coded" crashes is not a fair test of the value of knowing that a crash was assigned to a carrier for the reasons stated.

- Run a second analysis (in 4.2) predicting ONLY single vehicle crashes (crash rate or involvement) in the post-period. Examine the value of knowing pre-period single- vehicle preventable crashes on the ability to predict post-period single-vehicle crashes. **That analysis should give FMCSA the *upper bound* on the value of crash weighting.** If it adds a great deal of strength to the statistical model that incorporates "all crashes" then that should be the basis for determining the value of crash weighting (assuming one could do universal crash weighting for all crash types and severity).
- Provide a more complete explanation of the analysis and meaning of Tables 23 and 24. The description is not sufficient to ensure that the reader fully understands what the analysis is showing.
- Run another analysis where a similar weighting scheme is used but for single-vehicle preventable crashes. This was the truly important result of Analysis 4.2 (that these assignments were potentially valuable). It isn't clear why they weren't the subject of an alternative weighting analysis in 4.3.

Section 5 of the report:

- Address the start-up and recurring costs of training under various models such as (a) States code all the crashes; (b) a centralized staff (contractor?) codes all the crashes for the entire US.

Overall reporting of analytic results:

- For all statistical test results, provide the value of the test statistic and the associated p value or indicate the alpha level. This should be true for all comparison values as well.

Author: Eric Teoh, Insurance Institute for Highway Safety

Below is my review, in the requested format, of "Crash Weighting Analysis", as dated September 2013.

1. Clarity of Hypothesis

I think the report states the research questions, objectives, and hypotheses clearly.

2. Validity of Research Design

I think the research design, at a high level, is valid.

3. Quality of Data Collection Activities

Data collection in this study largely focused on obtaining PARs and coding information contained in them. The authors seem to have done this well, especially given the complicated nature of variation in states' policies and coding practices (including variation over time within a state). The authors' use of existing federal databases (MCMIS, FARS, NMVCCS) seems appropriate as well.

4. Robustness and Depth of Analysis Methods Employed

The level of depth of the analyses is appropriate, and I see no immediate concern in terms of robustness.

5. Appropriateness of Methods for the Hypotheses Being Tested

I have two main concerns here.

Firstly, while I agree that it is important to prevent all types of crashes, it seems odd to use crashes assigned to the carrier to predict all types of reportable crashes. The idea of assigning responsibility for a crash to the motor carrier, as stated in the report, is to focus on crashes that could have been avoided by actions of the carrier or CMV driver. Implicitly, crashes not assigned to the carrier (ignoring the small number of unassignable crashes) occurred largely due to exposure to other vehicles or adverse environmental conditions. So it seems like assigned crashes should be a good predictor of future assigned crashes. Any effect beyond this seems to be tantamount to a situation where assigned crashes predict exposure.

Secondly, dividing carriers by crash existence (those having only assigned crashes, only non-assigned crashes, both types, or neither) does not fully inform the question of to what extent assigned crashes are a better predictor of future crashes than are all reportable crashes. As the number of power units of a carrier increases, all else being equal, so does the likelihood of having a crash of any type. So large carriers have a much higher chance of being included in the

group that experienced both assigned and non-assigned crashes. This is evident in Table 13 on page 31 of the report in which the average number of power units for carriers with both types of crashes greatly exceeds that of the other three groups (965 vs. 67, 36, and 21). This partitioning method, while possibly accounting for carrier size to some extent, considers only the presence of assigned crashes, and not their number or rate, as a predictor of future crashes.

I think the analyses in section 4.3, implementing crash weighting in the SMS, were stronger in this regard since each crash was weighted and contributed to the overall crash indicator measure and carrier rating. Also the authors investigated the future crash rate by safety rating using a modified crash indicator measure, absent any intervention, relative to baseline (current crash indicator measure formula) to minimize the effect of unrelated time trends. Benefits in this analysis appeared primarily for fatal crashes, which are the most serious despite being relatively rare.

6. Extent to Which the Conclusions Follow the Analysis

Aside from the limitations stated above, I believe the conclusions are supported by the results.

7. Strengths and Limitations of the Overall Product

One major strength of this study is its overall research design. Since the carrier safety rating system did not change during the study period, this design provides an opportunity to evaluate the extent to which changes in how carriers are rated would be predictive of future crash rates. Had changes been implemented, it would be less certain that alternative rating systems predict future crashes differently since observed changes could be due also to changes in enforcement actions taken as a result of the alternative rating results.

Some limitations have been outlined above. Also, as discussed above and in the manuscript, variation in coding practices of PARs creates difficulty in assigning responsibility for crashes and in implementing a safety rating system based on such results. This is not to be taken as a limitation of the study efforts, but of the concept as a whole. However, the possibility of assigning crashes based on contributing factors directly (ie. instead of determining critical event/reason) does not appear to have been explored in the study.

Overall I think Volpe and the agency have made a reasonable effort to study the feasibility of implementing such a system.

8. Specific Recommendations for Improvement of the Product

Consider studying the rate of assigned crashes as an outcome measure.

Instead of separating carriers into groups based on presence of assigned crashes, perform a similar analysis that categorizes carriers based on the rate of assigned crashes per power unit.

Describe the negative binomial regression models a little more fully. For instance, it would be helpful to know exactly what variables were included in the models. AIC and BIC were used as measures of predictive ability, but perhaps it would be helpful to present the models' parameter estimates. If not, it would be helpful to explain why this would not be helpful.

Investigate whether contributing factors could be used to assign responsibility in situations where the critical event/reason could not be assigned.