

June 8, 2011

Via email

The Honorable Anne Ferro
Administrator
Federal Motor Carrier Safety Administration
1200 New Jersey Ave., S.E.
Washington, D.C. 20590

RE: Hours of Service of Drivers - Docket Number FMCSA-2004-19608

Dear Administrator Ferro:

Thank you for the opportunity to review and comment on several new research reports identified and/or funded by FMCSA and placed in the Hours-of-Service docket on May 9, 2011. The purpose of this letter is to forward to you a peer review critique of the study performed by Penn State University's Larson Transportation Institution entitled *Hours of Service and Driver Fatigue: Driver Characteristics Research*. We have also placed a copy of this peer review, conducted by Dr. Ron Knipling, in the hours of service docket.

As you know, Dr. Knipling is exceptionally well-qualified to assess these studies and the relevance of their findings to the proposed hours of service rule. He has more than 30 years of experience in traffic safety with emphasis on driver human factors and motor carrier safety. For six years he served as the Chief of FMCSA's Research Division before accepting a position as a Senior Research Scientist and Senior Transportation Fellow with the Virginia Tech Transportation Institute (VTTI). He has authored nearly 250 technical reports, papers, and conference presentations, as well as the first and only comprehensive textbook on large truck safety, *Safety for the Long Haul; Large Truck Crash Risk, Causation, & Prevention*. In recognition of the book, he received the International Road Transport Union's (IRU) Order of Merit award, the first given to an American scientist.

At ATA's request, Dr. Knipling reviewed the Penn State study and prepared the attached report. In his report, Dr. Knipling finds that the study has significant limitations such as:

- There was no description of crash characteristics (other than their associated work schedules) provided;
- There was no distillation of the crash dataset to exclude non-preventable crashes;
- The researchers did not perform validation tests of study conclusions via disaggregation of the crash dataset by prominent fatigue-related factors; i.e., single-vehicle vs. multi-vehicle crashes; other crash characteristics;
- Inadequate attention was paid to time-of-day as a looming confound;
- The study employed an inter-subject design rather than intra-subject design;
- There were a relatively small number of 11th hour crashes and exposure hours;
- The study sample may be unrepresentative due to apparent inclusion of truck tractors not equipped with sleeper berths;

- The report included minimal acknowledgements of study design limitations and minimal acknowledgements of the presence and probable influences of confounding and competing fatigue and other crash risk factors.

Based on these and other concerns, Dr. Knipling finds that “it would be erroneous and unwarranted to accept Penn State’s principal findings and conclusions without extensive re-analysis, internal validation, and external replication.”

We trust that Dr. Knipling’s report will be read and reviewed by Agency staff and very carefully considered in this process. Again, thank you for the opportunity to comment on these studies.

Sincerely,

A handwritten signature in black ink, appearing to read "D. Osiecki". The signature is fluid and cursive, with a large initial "D" and a long, sweeping underline.

David J. Osiecki
ATA Sr. Vice President for Policy & Regulatory Affairs

Enclosure

**Peer Review Critique of Penn State Study:
*Hours of Service and Driver Fatigue: Driver Characteristics Research***

Prepared for the American Trucking Associations
June 4, 2011

Submitted to Federal Motor Carrier Safety Administration (FMCSA)
Docket FMCSA-2004-19608

Ronald R. Knipling, Ph.D.
Safety for the Long Haul Consultancy
5059 36th N. Street; Arlington, VA 22207-2946
(703) 533-2895; rknipling@verizon.net

This is a review and critique of a research report by Pennsylvania State University placed on the FMCSA website and in the commercial driver Hours-of-Service (HOS) rulemaking docket (FMCSA-2004-19608) on May 6, 2011. The citation follows:

Jovanis, P. P. Wu, K-F., Chen, C. *Hours of Service and Driver Fatigue: Driver Characteristics Research*, Report No. FMCSA-RRR-11-018, Contract #19079-425868, Task Order #6, May 2011.

The Penn State report was a case-control comparison of HOS logs for drivers on truck trips where a crash occurred to those of different drivers where no crash occurred. Parameters of principal interest in the study included driving hours (time-on-task or TOT), multiday work patterns, and duty breaks. Time-of-day (TOD) was examined secondarily. The study incorporated 2004-2005 crash and control data previously analyzed and published (Jovanis et al., 2005). The 2004-2005 data were from three fleets. To this dataset it added new crash and control data from 2010, including data from one of the original fleets and two new fleets. The data were disaggregated by operational type (truckload vs. less-than-truckload) but not by crash type, preventability, conditions of occurrence, or causal factors.

The criticisms below relate most strongly to Penn State's conclusions in regard to driving hours, and especially the 11th hour of driving. They also relate to the issues of driver breaks and multi-day schedules. Concerns are expressed regarding the study's case-control design, lack of information on crash characteristics, use of an unfiltered crash dataset, lack of validation checks to eliminate alternative explanations, statistical procedures, and the overall plausibility of findings.

Inter-Subject Case-Control Design

As used in traffic safety research, the case-control methodology is a quasi-experimental approach used to determine the associations of driver, carrier, vehicle, and environmental factors with crash involvement. The case-control method can be useful, but it is inherently less rigorous than a well-conducted experiment. In an experiment, one systematically varies an independent variable (e.g., hours-of-driving, breaks) while *controlling* all other major variables known to affect the outcome of interest (e.g., crashes or other driving performance failures). A well-designed experiment demonstrates a causal relationship between independent and dependent variables. In contrast, the case-control method may demonstrate associations, but causation can only be inferred. When there is good control of exogenous variables within the case-control design, causal inference can be strong. When exogenous variables are not controlled, causal inference is questionable.

Control or counterbalancing of exogenous variables is essential when those variables are known to affect the outcome of interest; i.e., crashes. Otherwise, any outcome differences between groups may, in fact, be due to those exogenous factors. Had the Penn State study used an *intra*-subject design, whereby case (crash) schedules were compared to control (non-crash) schedules for the same drivers, it would have controlled for the biggest single factor affecting crash risk; i.e., the driver. Use of an inter-subject design did not in itself invalidate crash/non-crash schedule comparisons, but it did introduce the possibility that personal factors contributed to differences in schedules and crash outcomes.

Studies of both commercial and non-commercial drivers consistently show that risk among drivers is neither evenly nor randomly distributed (Hickman et al., 2005; Barr et al., 2011; Knipling et al., 2004; Knipling, 2009a). Rather, within almost any group of drivers, some will show greater likelihood of involvement in at-fault incidents whereas most others will show relatively low involvement rates. Among 95 drivers in the Virginia Tech Transportation Institute (VTTI) "Phase I" large truck naturalistic driving study (Hickman et al., 2005; Knipling et al., 2005), the worst 19% of drivers had 53% of all at-fault safety-critical events (SCEs). Hour-for-hour, these high-risk drivers were 4.9 times riskier than were other drivers. Individual differences in fatigue susceptibility were even more extreme; 13% of the drivers had 66% of the high-drowsiness SCEs; their SCE odds were 13 times those of the other 87% of the drivers. Most of the drivers in the Phase I study drove for prominent, national fleets much like those in Penn State's study. The 1996 Driver Fatigue and Alertness Study (DFAS; Wylie et al., 1996) also employed drivers with national fleets and found that 14% of the drivers had 54% of observed drowsiness (odd ratio = 7.2). In the Blanco et al. (2011) naturalistic driving study released by the agency along with the Penn State study, just four (4) of 97 drivers accounted for 36% of all SCEs. No *post hoc* comparison of crash-involved and non-crash-involved drivers can rightly assume that the two groups are equivalent. In fact, it should assume the opposite. The statistics cited above suggest that any group of crash-involved drivers consists largely of high-

risk drivers, whereas any group of non-crash-involved drivers consists mostly of lower-risk drivers. The two groups will differ in their personal characteristics relevant to risk, and these differences may be manifested in their work schedules and patterns.

An intra-subject design was called for in the Penn State study. When their researchers chose to use an inter-subject design, the onus was on them to provide evidence that the crash- and non-crash-involved drivers were comparable in their personal safety-relevant characteristics. This includes driver age, experience, known medical conditions, average daily and weekly work hours, and past crash and violation rates. None of these driver trait and history comparisons was provided and differences in any of them could and likely should challenge study findings.

No Analysis of Crash Characteristics

The Penn State report stated that the subject crashes “involved either a fatality, an injury requiring medical treatment away from the scene of the crash, or a towaway.” This appears to be merely a statement that the subject crashes were reportable to DOT’s SAFETYNET crash file, as required by law. Since no breakdown of the crashes per these three severity levels was provided, it has to be assumed that the only criterion applied to the crashes was that they be DOT-reportable. Crashes were disaggregated by fleet operations type (TL vs. LTL), but the crashes themselves were never described. Nor were they analyzed by their characteristics *as crashes*. The only analysis was by HOS-related schedule.

Page 3 of the Penn State report states that “the study team requested a list of crash information along with details of the hours driving prior to the crash.” The “list” seems to have been limited to HOS variables; i.e., log data and duty status. No further “list” of crash information requested is provided in the report. A list of critical fatigue-relevant information about a crash might also include its severity, the number of vehicles involved (i.e., single-vehicle vs. multi-vehicle), the crash type (e.g., rear-end striking, rear-end struck, lane change, left turn across path), its judged preventability, specific driver factors cited in the police report or company investigation, and various conditions of occurrence. All of these factors are indicators of crash’s probable relevance, or lack of relevance, to driver fatigue.

If “DOT-reportable” was indeed the only criterion applied, it would mean that, at a minimum, certain required SAFETYNET crash data elements would be in carrier files (FMCSA, 2008a). This includes numerous SAFETYNET variables relevant to crash risk and cause:

- Crash severity (fatal, injury, or towaway)
- Trafficway type; e.g., divided vs. undivided road
- Access control
- Road surface
- Weather conditions
- Crash events (sequence of events)

- Number of crash vehicles
- Number of persons medically transported
- Citation issued?

All of the above variables could have been used to predict and test likely and unlikely associations of driver fatigue and HOS-related variables with the observed crash data. Some specific examples will be provided in the sections below.

No Disaggregation by Crash Preventability/Fault

Page 4 of the Penn State report states that, “All of the carriers in the study were large national-scale carriers.” Virtually all such carriers investigate each of their crashes to determine *preventability* (Stock, 2001, Knipling et al., 2003; FMCSA, 2008b). In *A Motor Carrier’s Guide to Improving Highway Safety*, FMCSA (2008b) “recognizes that not all accidents are preventable” (Page 123) and provides a methodology for assessing preventability “based on the facts furnished in the motor carrier’s recordable accident register, and from various other sources” (Page 124). While there may be subtle differences, crash preventability is largely synonymous with *fault*.

The Penn State crash dataset retained all crashes provided by its participating carriers for the periods of study. A more incisive analysis would have disaggregated the dataset by crash preventability, and used only preventable crashes for primary study. The non-preventable crash dataset could have been used as a secondary, “control” subsample. Dick et al. (2006) gathered DOT-reportable crash data from 23 large fleets, all of which determined crash preventability as a standard procedure. Of 12,354 crashes reported, 6,181 (50.0%) had been classified as preventable by carriers. The two FMCSA-cited studies of Florida transit bus drivers (Sando et al., 2010a, 2010b) used a dataset consisting entirely of crashes reviewed and deemed preventable by its participating transit agencies. This was the scientifically correct procedure.

All crashes – preventable and non-preventable – are influenced by a multitude of ambient factors, including weather, lighting conditions, road surface, road type (i.e., divided vs. undivided), and traffic density. These factors confound all driver-related crash analyses, but especially those where the role of commercial driver error is minimal or non-existent. Commercial driver fatigue and factors reputed to cause fatigue *could* play a small role in not-at-fault or non-preventable crashes, but isolating and documenting that role amidst all the other interacting factors is highly problematic and easily subject to confounding.

The five Penn State carriers almost certainly made their own determinations of crash preventability. The study team could have used these determinations to distill its dataset, as was done in Florida. Alternatively, it could have made its own determination for analysis purposes

based on crash information submitted by the reporting State to SAFETYNET by using variables like Crash Events and Citation Issued.

Statistics from the Large Truck Crash Causation Study (LTCCS) illustrate the dramatic effect of preventability or fault determination on driver factors, including fatigue. In addition to coding proximal crash causes (CRs), the LTCCS coded *associated factors*. These were factors judged to be *present* in the crash; the LTCCS methodology (Blower and Campbell, 2005) specified that “No judgment is made as to whether the factors are related to the crash.” Nevertheless, the involvement patterns of associated factors illustrate the fundamental differences between at-fault (i.e., CR assigned to truck driver) and not-at-fault crashes. The following are comparisons based on two reports (Knipling and Bocanegra, 2008; Knipling, 2009b):

- Driver fatigue:
 - At-fault: 22%
 - Not-at-fault: 3%
- “Emotion/experience factors” (upset, under pressure, etc.)
 - At-fault: 46%
 - Not-at-fault: 22%
- “Speed/distance factors” (e.g., speeding, tailgating)
 - At-fault: 44%
 - Not-at-fault: 9%
- Roadway unfamiliarity (had rarely or never before driven roadway)
 - At-fault: 33%
 - Not-at-fault: 17%
- Not wearing safety belt (% of knowns)
 - At-fault: 15%
 - Not-at-fault: 6%.

Driver performance and behavior factors unquestionably play a far greater role in preventable/at-fault crashes than in non-preventable/not-at-fault crashes. Any exploration of driver factors in crashes should begin with that differentiation.

Single vs. Multi-Vehicle Crashes, and Other Missing Disaggregations

A particular disaggregation of importance in fatigue studies, and studies of driver error in general, is that of single-vehicle vs. multi-vehicle crashes. Simply disaggregating the sample by single-vehicle (SV) crashes vs. multi-vehicle (MV) ones would have provided a strong validation test of any driver performance-related hypothesis. SV crashes typically indicate a failure of driver *vehicle control*, whereas MV crashes reflect primarily a failure of *response to traffic events*. Compared to MV crashes (including both at-fault and not-at-fault crashes; based on Starnes, 2006, Knipling, 2009b; Knipling and Bocanegra, 2008), LTCCS SV crashes were:

- 7 times more likely to have “fatigue” as an associated factor (not necessarily a contributing factor);
- 32 times more likely to have truck driver asleep-at-the-wheel as the CR;
- 8 times more likely to have another physical impairment as the CR;
- 7 times more likely to have a performance/response execution error as the CR;
- 6 times *less* likely to have traffic as an associated factor.

Of all LTCCS crashes where truck driver fatigue was coded as an associated factor (not necessarily a contributing factor), 65% were SV crashes. For truck driver asleep-at-the-wheel as the CR, 91% were SV crashes. Corresponding percentages for other physical impairment and performance/response execution error were 76% and 74%, respectively. On the other hand, 93% of LTCCS crashes with the associated factor “traffic” were MV crashes. Any creditable analysis of the role of fatigue in large truck crashes would report both SV and MV statistics and perform separate analyses of them. If TOT were truly a major factor, it would be clearly evident in the SV crash subsample and much less evident, or not evident at all, in the MV crash sample. TOT effects on the entire crash sample are the “bottom line” concern, but SV vs. MV and similar disaggregations are necessary as validation checks.

The same criticism was filed in the 2007 HOS docket (Knipling and Engler, 2007) in response to the 2005 Penn State analysis (Jovanis et al., 2005). Yet the new Penn State work did not perform this validation test or otherwise disaggregate the sample for validation purposes. Distillation of the sample could have been performed in other ways as well. To reduce statistical “noise” and confounding factors, low-relevance crashes such as those listed below could have been eliminated from the sample. This is not to say that fatigue could not possibly play a role, but rather that the fatigue role is unlikely and/or small compared to other crashes. Low-relevance crashes might include:

- Parked or stopped truck
- Truck was struck in the rear
- Other driver was charged
- Severe road surface or weather conditions; e.g., ice, snow, fog, severe crosswinds
- Vehicle equipment or load securement failure.

If the Penn State study team had access to company log files and could take the time to enter those logs individually into computer software, they also could have reviewed the companies’ other records on the crashes.

No Disaggregation by Road Type

Another such co-variable is road type. Most notably, *trafficway flow* (i.e., divided vs. undivided roads) exerts a strong effect on crash risk. It varies systematically within many truck trips, which often end at locations away from Interstates and other divided highways. Many truck terminals

and customer locations are miles away from divided highways, and trucks must negotiate these miles to reach their end-of-trip locations. This late trip risk increase has been called “landing risk.”

In the same “Phase I” truck naturalistic driving study cited above (Hickman et al., 2005), analysts classified trafficway flow in each SCE video and in randomly selected driving periods. Only 10% of tractor-semitrailer driving was on undivided roadways, but 38% of incidents occurred there. This yields an SCE odds ratio of 5.3; in other words, driving on undivided roads had 5.3 times the SCE risk of driving on a freeway or other divided road. This finding is not surprising when one considers the design differences between divided and undivided roads. On divided roads, vehicles are all traveling in the same direction at relatively uniform speeds. On undivided roads, there are traffic signals, stops and starts, crossing vehicles, turning vehicles, pedestrians, many opportunities for distraction, and little margin-of-error. Penn State’s crashes were apparently all DOT-reportable, and their SAFETYNET reports would have contained a trafficway type designation. Trafficway flow could have been analyzed in relation to hours-of-driving and other study parameters. A disproportionate percentage of crashes on undivided highways late in truck trips would have suggested “landing risk” as a partial or alternative explanation for their observed late trip risk increases. Penn State’s alleged 11th hour effect was sharper for LTL than for TL fleets, consistent with the fact that LTL trips are relatively more likely to end at a terminal, whereas TL trips are relatively more likely to end at an Interstate rest area or nearby truck stop.

Unanswered Questions about TL vs. LTL Disaggregation

Penn State analyzed its TL and LTL data separately because, as stated in the Foreword, “previous research indicated differences in crash contributing factors for these two segments of the trucking industry.” Later, on Page 5, is a statement that, “previous research . . . indicated that the crash odds models for the two carrier types are significantly different.” No specifics were provided on either TL vs. LTL crash contributing factors or relative crash rates. The study did not report fleet mileages or overall crash rates for the periods of study. The case-control method (i.e., starting with a crash and then finding two controls) did not shed light on the fleets’ relative safety levels. In general, disaggregation of samples when there are known underlying subsample differences is a good practice to prevent masking of these differences within a whole-sample analysis. It raises the question, however, of why other, more important crash categories were not examined. Here are five different disaggregations which would have had stronger rationales and which would have provided more insight into driver fatigue and other crash risk factors:

- Preventable (more fatigue) vs. non-preventable
- SV (more fatigue) vs. MV
- Divided road (more fatigue) vs. undivided (more traffic effects)
- Night driving (more fatigue) vs. day driving (more traffic effects)
- Injury/fatal crashes (more fatigue) vs. no-injury towaways.

While the study classified both TL and LTL schedules into 10 multiday patterns, and examined their associations with crash risk, it never described or compared the TL and LTL samples in regard to TOD. TOD was entered into the TL model as a third factor after TOT and multiday pattern (as discussed above), but TOD was not addressed at all in the LTL analysis.

Although no specific data were provided, it appears that the Penn State TL subsample overwhelmingly involved day driving, whereas the LTL sample was overwhelmingly night driving. Seven of the ten TL patterns, representing 655 of 878 schedules (75%) involved mostly day driving. In contrast, six of the ten LTL patterns, representing 500 of 686 schedules (73%), involved mostly night driving. Day and night driving involve largely different crash risks. Traffic densities are much higher during the day and most crashes result from driver errors during vehicle-vehicle interactions. At night, traffic is sparse but there are greater risks associated with reckless and/or impaired other motorists, and from truck driver fatigue. The last hours of night driving trips often include both fatigue risks from early morning driving (i.e., 4:00 to 7:00am), overlapping with traffic risk from morning rush hours. The more significant 11th-hour rise for LTL than TL drivers is consistent with this. For both groups, however, many trips end during higher risk periods due to circadian factors and/or traffic density increases.

On Page 31, the authors themselves cast doubt on their TL findings: “This [TL] model shows no consistent trend relating crash odds to hours driving. The study team believes that the crash-odds increase in the last hour is in need of further analysis. At least a portion of the increase in odds may be attributable to the low sample size of observations in the last hour of driving (9 crashes of 318 TL crashes in the data . . .).”

Penn State’s current study incorporates data from its 2005 study (Jovanis et al., 2005). In that study, 50% of the crash sample and 46% of the non-crash sample was from “non-sleeper” operations. Apparently, “non-sleeper” operations involved trucks not equipped with sleeper berths. The new report does not explain the relationship between the dataset used in 2005 and the 2004-2005 portion of the current dataset, nor does it state the proportion of its trucks equipped with berths. It appears that roughly half of the current 2004-2005 subsample must also have involved “non-sleeper” operations, presumably on the LTL side. This raises a question about the representativeness of the current total LTL sample. Today, only a small percent of tractor-semitrailers in regional and long-haul use – TL or LTL – are not equipped with sleeper berths. Non-sleeper “day cabs” tend to be used in local/short-haul operations that are atypical of tractor-semitrailer operations in general.

Inadequate Statistical Analysis of Time-of-Day (TOD) Effects

TOD operates via two different mechanisms to strongly affect crash risk. The first association is with traffic density. In the “Phase I” VTTI large truck naturalistic driving study (Hickman et al.,

2005; Knipling et al., 2005), trucks drove in heavy-to-medium traffic 27% of the time (based on random control observations) but had 41% of their safety-critical events (SCEs) under these conditions. That's an odds ratio of 1.9 for SCE likelihood in heavy/medium vs. light traffic. In a large naturalistic driving study, Hanowski et al. (2008) found the correlation between truck SCE rate and average traffic density by TOD to be 0.83, meaning that 69% of the variance in SCE rate was associated with traffic density. A new analysis of freeway safety (Kononov et al., 2011) finds that crash rates per vehicle mile traveled rise slowly with increased traffic density *until a critical density is met*. Then they rise sharply, much like the sigmoid (S-shaped) dose-response curve often seen with medications. In the study, a 60% increase in freeway traffic beyond the critical density caused an 84% increase in crash *rate* per vehicle miles traveled. In other words, each vehicle's crash risk almost doubled. If trucks end their trips during rush hours and especially on undivided roads during rush hours, they will have higher crash rates during these times independently of driver performance factors. The authors acknowledged "higher traffic levels" as a possible confound for their TL findings (Page 64), but it was also an operative confound in their LTL sampling, which included many trips ending during morning rush hours.

The second major influence of TOD is via driver alertness. Circadian rhythms strongly affect human alertness every day. The deepest circadian trough is between 4:00am and 7:00am. Another, shallower trough occurs mid-afternoon. In the DFAS (Wylie, 1996), TOD was the "strongest and most consistent factor influencing driver fatigue and alertness." In the LTCCS, 62% of truck driver asleep-at-the-wheel crashes occurred in the *two-hour period* between 4:01am and 6:00am. Only 20% occurred in the 14-hour period between 6:01am and 10:00pm (Knipling, 2009a).

No clear descriptive statistics on crash TOD were presented in the Penn State study. Its Table 2 (Page 6) classifies crashes, non-crash exposure, crash exposure, total exposure, and crashes/exposure by driving hour/TOT, but there is no similar table for TOD. TOD values were known for every crash and hour of driving in the study datasets. A full and open presentation would have included these statistics.

The deficiency extends to the analytical modeling as well. The logistic regression method employed by Penn State was the correct *type* of analysis for differentiating TOT and TOD effects, but the specific *sequence* of procedures chosen by the researchers was inadequate for TOT-TOD differentiation. Their highly parameterized model likely had the effect of masking TOD associations with crash involvement in the TL fleet analysis. Their LTL analysis ignored TOD entirely, even though most of their LTL schedules involved overnight driving where trips ended in early morning period when risks from circadian troughs and/or morning rush hour could be operating.

In the TL analysis, Penn State first performed logistic regression for driving time (TOT) as a predictor, then added their multiday patterns as a predictor, then identified TOT-Pattern interactions, and then added TOD to the model as a final step. In other words, their sequence was:

TOT → Multiday Patterns → TOT-Pattern Interactions → TOD → Final Model.

After doing this, the following conclusion was drawn for TL fleets on Page 35: ““When time of day is added, many parameters change magnitude by a small amount . . . The study team’s judgment is that the effect of time of day is already addressed by the main effects and interactions of driving time and multiday patterns. Therefore, time of day was dropped as a predictor in subsequent models, as there is little need for an additional time-of-day variable.”

Given the order chosen for entering variables into the Penn State model, it is not surprising that TOD had no effect. No rationale was provided for choosing to add TOD to their model *after* adding the multiday driving pattern clusters. TOD differences were embedded in the multiday patterns, and thus findings attributed to the multiday patterns *per se* could mask underlying TOD influences. A clear alternative would have been the following sequence:

TOT → TOD → TOT-TOD Interactions → Multiday Patterns → Final Model.

This approach would differentiate TOT and TOD effects, as well as show their interactions. It would also show if adding multiday patterns improved the model, and their interactions with specific TOT-TOD effects. It’s unimaginable that TOD would not emerge as a stronger factor in this analysis approach than in the one chosen by Penn State. But, if it did not, that would justify the Penn State approach for this dataset. In the absence of it, there was no fair or adequate assessment of TOD in the study. If FMCSA and Penn State do not wish to perform this alternative analysis, they should consider making the dataset available for other interested researchers to do it.

Other Statistical Issues

The study team relied exclusively on the Akaike Information Criterion (AIC) statistic to describe the adequacy of their statistical modeling efforts. Yet AIC is appropriate only for determining which of two or more nested models is a superior fit to the observed study data. While it is correct to use AIC to say one model is better than another, the initial model (e.g., as seen in Table 12, Page 31) requires evidence that it is an adequate starting point for modeling the observed data. A non-significant model Chi-Square or Hosmer-Lemeshow test, for example, would lend support to model-predicted values closely resembling the observed data. Absent this information, there is no reason to believe that the initial model suitably fits the data. Therefore, comparing multiple models with AIC information loses much of its meaning, as it merely

assesses their *relative* goodness. AIC doesn't prove that any of them meet any external standard of goodness.

The level of significance $p \leq 0.20$ is used in the Penn State analysis, although some of their key findings reach higher levels. Rarely does one see this low level of significance in scientific work; sometimes the phrase “marginally significant” is used to describe findings at $p \leq 0.20$ or $p \leq 0.10$. While it may be appropriate for exploratory modeling, using a low significance threshold is a weak basis for policy decisions, especially in light of other statistical and design issues surrounding the study. A related concern is that of *post hoc* comparisons; when $p \leq 0.20$ is the benchmark, the effects of chance alone will make 20% of all comparisons appear significant.

While Penn State has added a large number of crashes and exposure points to its previous (Jovanis et al., 2005) analysis, the portion of crashes and exposure values in the critical 11th hour remain small compared to the overall dataset. Eleventh hour crashes were just 16 out of the total of 542 (2.95%), and 66 of 1,564 drivers (4.22%) drove in the 11th hour. When the 16 crashes were disaggregated, there were just nine (9) in the TL sample and seven (7) in the LTL sample. Given these small numbers and the many exogenous risk factors and crash causes that could have been operating, drawing conclusions about the effects of one risk factor seems questionable in spite of the derived levels of significance for 11th hour findings. FMCSA's own quality standards for HOS-related studies included the requirement that they are, “large enough to provide an accurate picture for individual hours, despite the rarity and randomness of crashes and the relatively small fraction of driving in the later hours . . . “ (December 2010 HOS NPRM, 75 Federal Register at 82179).

Study Findings on 34-Hour Recovery Periods

Penn State's studies of the effects of recovery periods found that crash odds went up after a recovery period. Time off work was followed by higher crash odds upon returning to work. A regular sequence of work days was better from the safety perspective. In other words, time off is bad for safety! This paradoxical finding does not relate meaningfully to the specifics of HOS rules, although it does suggest that the current daily HOS schedule successfully manages driver fatigue and sustains performance on a daily basis. Rather, the finding reflects driver activities and lifestyle choices during their off-duty days. Previous studies have also found increased incident rates after weekend breaks at the beginning of work weeks (e.g., Hanowski et al., 2000). A well-designed study of the safety effects of the 34-hour restart would parametrically compare crash odds in driving periods following recovery periods of varying lengths, with controls for TOD and other confounding factors.

As part of its 34-Hour Restart modeling, Penn State studied “pseudo-violations,” defined as schedules that would have violated the HOS rules had not the restart provision been in place. Their “pseudo-violation” findings were mixed. A “pseudo-violation” over two days was

associated with a crash odds increase, but one over one day appeared to decrease crash odds. The analysis “included crash-involved drivers only for one carrier over a limited time period” (Page 64). This preliminary, exploratory finding carries little weight in regard to current HOS deliberations.

Multiday Patterns

Each of the multiday patterns also involved a predominant daily TOD pattern. Based on a visual inspection of the 20 patterns (10 TL and 10 LTL), it appears that nine (9) involved mostly day driving, seven (7) involved mostly night driving, two (2) involved an early morning/day pattern, and two (2) involved mostly evening driving. Regardless of the exact breakdown, it is clear that TOD was a major factor in the classifications. As noted in the above critique of the logistic regression sequence, adding multiday patterns to the model before adding TOD had the effect of masking TOD effects. No one would deny that some multiday work patterns are intrinsically safer, or less safe, than others. However, a more logical and heuristic approach would be to assess their individual hourly elements first (i.e., TOD and TOT) and then determine further whether different multiday combinations of these elements further add to, or detract from, safety.

Consistent with this point, the study authors note on Page 64 that the TL multiday driving patterns with the highest late-hour crash odds all involved “extended hours occurring in late afternoon and early evening when higher traffic levels are possible. Of the four LTL patterns with the highest interactions with driving hours, two (Patterns 2 & 4) “result[ed] in increases in crash risk for driving near 6 a.m. This has been shown in previous studies to be a time of low circadian activity and elevated crash occurrence” (Page 54). The other two (Patterns 7 & 9) resulted in extended driving at mid-day or late afternoon, with “traffic” cited as a possible factor in each. A greater contribution to knowledge about truck safety would come from identifying multiday schedule associations with safety *apart from* late-shift TOD effects.

Fundamental Implausibility of Findings

Consider everything we know about driver fatigue risk factors and other risk factors and causes of crashes. Is it really plausible that one factor, driving TOT, could exert an influence as strong and clear as that claimed by this Penn State study? Consider the many other well-known factors affecting driver alertness and fatigue:

- Individual differences in fatigue susceptibility, which may be related to:
 - Sleep disorders or other medical conditions
 - “Normal” population variation (somewhat like normal variation in height).
- Time-of-day (circadian rhythms), with the following high-risk times:
 - Early morning (e.g., 4:00 to 7:00am)
 - Mid-afternoon.
- Hours of previous sleep:
 - Previous primary sleep period (most importantly)

- Prior nights
- Naps.
- Hours awake since last principal sleep; especially at 16+ hours, and independently of work or specific work activities.
- Health and wellness and recent related behaviors; i.e., diet and exercise.
- Caffeine intake.
- Prescription and over-the-counter drug use.

In a 2005 white paper, current National Transportation Safety Board (NTSB) Board Member Mark Rosekind highlighted three factors, as follows (from Page 12):

While there are a variety of complex factors that can affect fatigue, there are three primary physiological factors that have been scientifically demonstrated to affect alertness, performance and safety. These three factors are: a) sleep (specifically acute sleep loss and cumulative sleep debt), b) hours of continuous wakefulness, and c) circadian rhythms (time of day effects on sleep, alertness and performance).

On Page 7, the Rosekind report adds:

Given the physiological complexities associated with sleep and circadian rhythms, a regulatory approach or structure that relies on scientific findings for justification of each element will struggle with the data that are available to be fully supported. For example, while there is a large and consistent database of findings on sleep need and the effects of sleep loss, there are little to no relevant data to address work time (e.g., driving, flying) within a duty period.

Note in the above discussion that time *awake* is well established as an underlying physiological factor in alertness. Time awake above 16 hours is a principal element in most Sleep Performance Models (Krueger, 2004). One of these, the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model, has been used by FMCSA to assess its HOS rules. In almost any driver schedule, work hours and driving hours co-vary with time awake. Work and driving hours should not be treated as independent factors, however, unless it is proven that they operate independently.

The above suggests that the independent role of TOT in driver fatigue is limited. Next, consider the independent importance of driver fatigue as a cause of crashes. Only 3.8% of truck crash involvements in the LTCCS involved truck driver asleep-at-the-wheel as the CR. Thirteen percent (13%) were reported to involve fatigue as an associated factor, but this designation did *not* necessarily imply that fatigue played a contributory role (Blower and Campbell, 2005). The

LTCCS listed more than 40 specific proximal crash causes. Eight major LTCCS causal categories largely unrelated to fatigue are listed below:

- Errors of other motorists or other failures (e.g., vehicle) associated with them.
- Truck driver traffic violations or other misbehaviors (e.g., speeding, tailgating).
- Awake inattention, ranging from transient distractions associated with the driving task to egregious inattention associated with cell phone use or other non-driving behaviors.
- Information processing errors and misjudgments, such as misjudgment of cross traffic closing distances.
- Errors executing specific driving maneuvers, such as merges and turns. Two particularly difficult maneuvers for large trucks are merges/lane changes and 90° turns.
- Environmental and roadway factors, most notably adverse weather and roadway design factors (e.g., sharp curves and ramps).
- Vehicle deficiencies or defects
- Reduced driver alertness due to factors other than fatigue; i.e., illness.

In-depth investigation of the Penn State crashes would likely have revealed an even smaller role for fatigue, because their dataset included non-injury towaway crashes. The LTCCS dataset included only crashes with serious injuries or fatalities. The reported fatigue role in injury crashes is several times greater than its role in non-injury crashes (Knipling, 2009a, Knipling and Shelton, 1999).

Given all the “competing” fatigue factors and all the larger crash causes beyond fatigue, is it likely that a single fatigue factor, or a single driver performance factor of any kind, would be mathematically discernible in an unfiltered dataset of DOT-reportable crashes? Across the spectrum of road safety research, the strongest factors affecting whole populations of crashes are those which operate on all drivers and vehicles all the time. Notably, this includes road type and condition (e.g., divided vs. undivided; surface, weather) and traffic characteristics (e.g., density and speed).

Summary

The current Penn State study has added new 2010 crashes to its 2004-2005 dataset, and changed some aspects of its analysis and modeling approach. It has not, however, addressed fundamental criticisms of the approach raised four years ago by Knipling and Engler (2007) and filed in the HOS docket. This includes the following:

- No description of crash characteristics (other than their associated work schedules).
- No distillation of the crash dataset to exclude non-preventable crashes.
- No validation tests of study conclusions via disaggregations of the crash dataset by prominent fatigue-related factors; i.e., single-vehicle vs. multi-vehicle crashes; other crash characteristics.

- Inadequate attention to TOD as a looming confound.
- Use of inter-subject design rather than intra-subject design.
- Relatively small number of 11th hour crashes and exposure hours.
- Possible sample unrepresentativeness due to apparent inclusion of truck tractors not equipped with sleeper berths.
- Minimal acknowledgements of study design limitations.
- Minimal acknowledgements of the presence and probable influences of confounding and competing fatigue and other crash risk factors.

Among the first things taught about correlational methods is that they do not determine causation. For example, the following is from a primer entitled, *Statistics for Management; A Simplified Introduction to Statistics* (Mandel, 1977, Page 377):

It is not difficult by associating different pairs of variables to find a high degree of correlation among some of them. In such a situation it is tempting to conclude that one variable causes the other to change – to assume automatically a cause-and-effect relationship. It must be determined, however, whether the . . . association is not just a coincidence over a limited period of time, or whether the two variables are both associated with a third variable that causes both of them. . . . A common-sense approach is needed to tell whether two variables are, in fact, *causally related*.

Logistic regression is a type of correlational analysis, and thus the same fundamental principles, cautions, and pitfalls apply. Multiple “third variables” were likely operating within the Penn State dataset. Accordingly, it would be erroneous and unwarranted to accept Penn State’s principal findings and conclusions without extensive re-analysis, internal validation, and external replication.

Cited References

Balkin, T.J., Thorne, D., Sing, H., Thomas, M., Redmond, D.P., Wesensten, N., Russo, M., Williams, J., Hall, S., & Belenky, G.L., *Effects of Sleep Schedules on Commercial Motor Vehicle Driver Performance*, FMCSA Technical Report No. DOT-MC-00-133, U.S. Department of Transportation, Washington, DC, 2000.

Blanco, M., Hanowski, R. J., Olson, R.L., Morgan, J. F., Soccolich, S. A., Wu, S-C, and Guo, F. *The Impact of Driving, Non-Driving Work, and Rest Breaks on Driving Performance in Commercial Motor Vehicle Operations*. Report No. FMCSA-RRR-11-017, May 2011.

Blower, D.F. and Campbell, K.L. *Methodology of the Large Truck Crash Causation Study*. Large Truck Crash Causation Study Analysis Series. Report No. FMCSA-RI-05-035. February, 2005.

Dick, V., Hendrix, J.W., & Knipling, R.R. New Hours-of-Service rules: trucking industry reactions and safety outcomes. *Transportation Research Record*, No. 1966, Pp. 103-109, 2006.

Hanowski, R. J., Olson, R. L., Bocanegra, J. and Hickman, J.S.. *Analysis of Risk as a Function of Driving-Hour: Assessment of Driving-Hours 1 Through 11*. Report No. FMCSA-RRR-08-002, January 2008

Hickman, J.S., Knipling, R.R., Olson, R.L., Fumero, M., Hanowski, R.J., & Blanco, M. *Phase I - Preliminary Analysis of Data Collected In The Drowsy Driver Warning System Field Operational Test: Task 5, Phase I Data Analysis*, for the FMCSA under NHTSA Contract DTNH22-00-C-07007, TO #21, September 30, 2005.

Jovanis, P.P., Park, S-W., Gross, F., & Chen, K-Y. Crash Risk and Hours Driving: Interim Report II. Submitted to the FMCSA on August 19, 2005 under Project Number DTMC75-03-C-00011. Report dated April 15, 2005.

FMCSA. *SAFETYNET Crash System Coding and Validation Manual*. August, 2008a.

FMCSA. *A Motor Carrier's Guide to Improving Highway Safety*. FMCSA-ESO-08-003, October 2008b.

Hanowski, R. J., Wierwille, W. W., Garness, S. A., and Dingus, T. A. *Impact of Local Short Haul Operations on Driver Fatigue*. Final Report No. DOT-MC-00-203. Washington, DC: U.S. Department of Transportation, Federal Motor Carriers Safety Administration, September, 2000.

Knipling, R.R. *Safety for the Long Haul; Large Truck Crash Risk, Causation, & Prevention*. American Trucking Associations. ISBN 978-0-692-00073-1, 2009a.

Knipling, R.R. Three large truck crash categories: what they tell us about crash causation. Proceedings of the *Driving Assessment 2009* conference, Pp. 31-37, Big Sky, Montana, June, 2009b.

Knipling, R.R. & Bocanegra, J. *Comparison of Combination-Unit Truck and Single-Unit Truck Statistics from the LTCCS*. FMCSA & Volpe Center Project report. Contract No. DTRS57-04-D-30043. 2008.

Knipling, R.R., Boyle, L.N., Hickman, J.S., York, J.S., Daecher, C., Olsen, E. C. B., and Prailey, T.D. *Synthesis Report #4: Individual Differences and the High- Risk Commercial Driver*. TRB Commercial Truck & Bus Synthesis Program. ISSN 1544-6808, ISBN 0-309-08810-0, 2004a.

Knipling, R.R. & Engler, D. Peer review critique of Penn State Jovanis et al. report: Crash Risk and Hours Driving: Interim Report II. Prepared for the ATA and submitted to FMCSA HOS Docket, Document ID FMCSA-2004-19608-3069.1, September, 2007.

Knipling, R.R. Hanowski, R.J.; Hickman, J.S., Olson, R.L., Dingus, T.A. and Carroll, R.J. Exposure-risk analysis of large truck naturalistic driving data. *Proceedings of the 2005 Truck & Bus Safety & Security Symposium*, Alexandria, VA, November 14-16, 2005.

Knipling, R.R., J. S. Hickman, and G. Bergoffen, *Synthesis 1: Effective Commercial Truck and Bus Safety Management Techniques; A Synthesis of Safety Practice*. Transportation Research Board Commercial Truck & Bus Synthesis Program Project MC-02, ISSN 1544-6808, ISBN 0-309-08754-6, 2003.

Knipling, R.R. and Shelton, T.T. Problem size assessment: large truck crashes related primarily to driver fatigue. *Proceedings of the Second International Large Truck Safety Symposium*, E01-2510-002-00, University of Tennessee Transportation Center, Knoxville, Pp. 3-12, October 6-8, 1999.

Kononov, J., Lyon, C., and Allery, B.K. Relating flow, speed and density of urban freeways to functional form of an SPF [Safety Performance Function], Paper 11-2070, *Transportation Research Board Annual Meeting*, 2011.

Krueger, G.P. *Technologies and Methods for Monitoring Driver Alertness and Detecting Driver Fatigue: A Review Applicable to Long-Haul Truck Driving*. Unpublished report for ATRI and FMCSA. June 2004.

Mandel, B. J. *Statistics for Management; A Simplified Introduction to Statistics*, Dangary Publishing Company, Baltimore, 1977.

Rosekind, M.R. *Managing Safety, Alertness and Performance through Federal Hours-of-Service Regulations: Opportunities and Challenges*. Alertness Solutions. FMCSA rulemaking docket #FMCSA-2004-19608. 2005.

Sando, T., Angel, M., Mtoi, E., & Moses, R. Analysis of the relationship between operator cumulative driving hours and involvement in preventable collisions. Paper 11-4165 submitted to the TRB Annual Meeting, Session 665, Submitted November, 2010a.

Sando, T., Mtoi, E., & Moses, R. Potential causes of driver fatigue: a study on transit bus operators in Florida. Paper 11-3398 submitted to the TRB Annual Meeting, Session 665, Submitted November, 2010b.

Starnes, M. *LTCCS: An Initial Overview*. NHTSA National Center for Statistics & Analysis, DOTR HS 810 646, August 2006.

Stock, D. *I-95 Corridor Coalition Field Operational Test 10: Coordinated Safety Management; Volume I: Best Practices in Motor Carrier Safety Management*, PENNDOT Final Report, FHWA-PA-2001-020-97-04(8), Pennsylvania Transportation Institute, August, 2001.

Wylie, C.D., Shultz, T., Miller, J.C., Mitler, M.M., & Mackie, R.R., *Commercial Motor Vehicle Driver Fatigue and Alertness Study*, Federal Highway Administration, U.S. Department of Transportation, Washington, DC, 1996.

Author Bio:

Dr. Ronald R. Knipling is the author of the first and only comprehensive textbook on large truck safety, *Safety for the Long Haul; Large Truck Crash Risk, Causation, & Prevention*. In recognition of the book, he received the International Road Transport Union (IRU) Order of Merit award, the first given to an American scientist. Dr. Knipling has more than 30 years experience in traffic safety with emphasis on driver human factors and motor carrier safety. Specialty areas include crash data analysis, driver risk, crash causation, driver fatigue, hours-of-service, naturalistic driving, technology assessment, and carrier safety management. Accomplishments include nearly 250 technical reports, papers, and conference presentations.