NEED FOR REVISED TOTAL EYES-OFF-ROAD CRITERION IN THE NHTSA DISTRACTION GUIDELINES: TRACK RADIO-TUNING DATA

Richard A. Young
School of Medicine and College of Engineering, Wayne State University
Detroit, Michigan USA
richardyoung9@gmail.com

Summary: This study re-analyzes participant-level glance data from a NHTSA-sponsored test track study of nine radio-tuning tasks in five radios. NHTSA stated that in its judgment, all nine tasks met the definition of traditional manual radio tuning, and so collapsed the data across all participants to estimate an 85th percentile. NHTSA further stated that it combined that track percentile with the 85th percentile from a radio-tuning task in a separate simulator study, to set its total-eyes-off-road time (TEORT) acceptance criterion. Given NHTSA's statements, individual radio-tuning tasks should, in general, meet the criteria created from them. This study performed such an analysis, and found that this expectation was not met. Four out of nine radio-tuning tasks did not meet the criterion. One problem is that NHTSA did not allow for variability in its 85th percentile estimate. Additionally, TEORT values were higher in the simulator than track for the same task with age-matched data, meaning that if the track tasks had been run in the simulator then the 85th percentile TEORT may have been higher. These issues illustrate the need for revising the criteria based on an improved analysis of the data that NHTSA used to set those criteria. Without doing so, many commonly-accepted secondary tasks (including manual radio tuning in many vehicles) would not meet the current NHTSA Guidelines glance criteria. Revised criteria should be derived in a way that would provide the needed consistency with age-balance requirements of task-acceptability testing, as well as allowing robustness for variability in the percentile estimates.

INTRODUCTION

The National Highway Traffic Safety Administration (NHTSA) issued nonbinding, voluntary Driver Distraction Guidelines (hereafter referred to as "Guidelines") to "promote safety by discouraging the introduction of excessively distracting devices in vehicles" (NHTSA, 2013, p. 24818). In a NHTSA-sponsored and co-authored study, Perez et al. (2013) state that their team analyzed visual demand acceptability criteria based upon two sets of data, "Results of the test track radio running experiment were evaluated along with experimental data for radio tuning obtained in a driving simulator by NHTSA… The data suggest the following visual demand acceptability criteria based upon driver 85th percentile radio tuning performance." These results were later used by NHTSA as the threshold values for glance criteria in the subsequently-developed Guidelines, as cited by NHTSA in the Guideline document (p. 24864). The three final NHTSA-specified acceptance criteria for assessing visual-manual tasks were based on Total Eyes-Off-Road Time (TEORT), Long Glance Proportion (LGP), and Mean Single Glance Duration (MSGD). Young (2015) documents these and re-evaluates NHTSA simulator data, while the current study re-examines track data (Perez et al., 2013).

One of NHTSA's (2013, p. 24820) "fundamental principles" for the Guidelines is, "The distraction induced by any secondary task performed while driving should not exceed that associated with a baseline reference task (manual radio tuning)." It follows from this statement that traditional manual radio-tuning tasks performed on representative radios should meet the Guidelines criteria. Ideally, they should do so in a reliable manner, so that repetitions of the test using the NHTSA protocols would reliably give rise to the same result, with different participants or a different simulator.

NHTSA's Radio Tuning Reference Task

NHTSA (2012a, p. 11225) explains its use of radio tuning as a reference for setting its criteria:

> *Since there is no agreed upon absolute level at which distraction becomes unacceptably high, a relative limit can be developed by comparing the distraction level associated with a driver performing an "acceptable" reference task with the distraction level associated with a driver performing new tasks.…NHTSA has chosen traditional, manual radio tuning as its recommended reference task.*

If mean glance values were used for acceptance testing under this reasoning, then it would be expected that, on average, about 85% of reference radio-tuning tasks would meet an 85th percentile criterion value established for the glance metric, if tested according to the NHTSA procedures. However, NHTSA's acceptance test is not based on means, but on 21 of 24 participants, which is effectively a percentile criterion. Therefore, a task whose glance metric was near the acceptance criterion would randomly meet or not meet the criterion about half the time, just due to random participant variability. Indeed, those tasks with glance metrics nearest the acceptance criteria, are, in fact, those used to set the criteria. Thus, one way to verify whether the method used to set the currently-published criterion values is robust to test variability is to determine whether the individual radio-tuning tasks used to set the criterion, meet that criterion.

NHTSA-Sponsored Test Track Study

The Perez et al. (2013) study was conducted between October and December 2010 on a closed 4-lane test track built to highway specifications at Virginia Tech Transportation Institute. It collected glance data from 43 participants assigned in various groups from n = 19 to 40 that separately performed nine radio-tuning tasks in five radios. Table 1 summarizes the vehicle and test protocols derived from descriptions in Perez et al. (2013). The 2010 Toyota Prius premium navigation radio reflected in columns 8 and 9 was also tested in the NHTSA simulator (see Young, 2015).

Table 1. Protocols for the 9 radio-tuning tasks tested on the test track. The variables studied were vehicle type (rows 3-6), tuning method (row 7), and constant or varied speed of lead vehicle (row 9). The letter at the end of the ID in row 2 codes the tuning method (K = Knob, B = Button, S = Seek, T = Toggle)

Task	1.	2.	3.	4.	5.	6.	7.	8.	9.
ID	1CadillacK	2ImpalaB	3ImpalaK	4ImpalaB	5ImpalaK	6InfinitiS	7MercedesT	8PriusK	9PriusK
Year	2006	2010	2010	2010	2010	2006	2005	2010	2010
OEM	GM	GM	GM	GM	GM	Nissan	Mercedes	Toyota	Toyota

Task	1.	2.	3.	4.	5.	6.	7.	8.	9.
Make	Cadillac	Chevrolet	Chevrolet	Chevrolet	Chevrolet	Infiniti	R-Class	Prius	Prius
Model	STS	Impala	Impala	Impala	Impala	M35	R350	Premium	Premium
Tuning	knob	button	knob	button	knob	seek	toggle	knob	knob
Phase	1	1	1	2	2	1	1	2	2
Speed	constant	constant	constant	varied	varied	constant	constant	constant	Varied

The stated purpose of the track study by was to "evaluate the radio tuning reference task" described in the Alliance (2006) guidelines by assessing driver performance and glance metrics during the task. As such, the tests were conducted according to the Alliance (2006) protocols. Indeed, Perez et al. (2013, p. 4) state that all five tested radios, "met the apparatus specifications contained in the Alliance Guidelines," and give details of the faceplates, task steps, and data collection methods.

It was not the purpose of the Perez et al. (2013) study to set glance criteria for the NHTSA (2013) Guidelines (even though NHTSA has stated it eventually used them for that purpose), or conduct the study according to the NHTSA (2013) test protocols, which did even exist at the time of data collection in 2010. As a result, there were some substantial differences between the test protocols and the NHTSA Guideline protocols. For example, a test track is not an acceptable test venue in the Guidelines, which defines only simulator and occlusion goggles tests. It is also well established that the absolute values of driving performance metrics during secondary tasks in track or road tests are not the same as in simulator tests (Young et al., 2005, 2009). In addition, all 43 participants were in the age range 45-64 as specified by the Alliance (2006) Guidelines, and did not include the age ranges 25-34 and 35-44 as required by the NHTSA (2013) Guidelines. Nonetheless, NHTSA (2013, p. 24861) explicitly stated that it used these track data in part to set its glance criteria. To estimate the 85th percentile TEORT for the track data, Perez et al. (2013, Table 8) simply collapsed all 218 trials across all participants and all 9 radio-tuning tasks, but did not state why. The *objective* of the current study is to re-analyze the TEORT track data on a task-by-task basis in order to test whether the radio-tuning tasks used to set the NHTSA TEORT criterion, actually meet it.

METHOD FOR DATA ANALYSIS

NHTSA (2012b) publically released participant-level TEORT data for each of the nine radio-tuning tasks (the test trials are separate from the preceding practice trials). All other glance variables in the released dataset were glances to the radio itself (the Alliance, 2006 requirement), which are a subset of the NHTSA TEORT glances, which also include, for example, glances to mirrors or out the left and right windows. The glance-level data from which off-the-road MSGD and LGP metrics could be calculated were separately requested by the author in an email to NHTSA on February 19, 2014, but were not received, so this study is limited to analysis of the TEORT data only.

There were 43 participants, who each performed anywhere from 2 to 6 of the test scenarios, for a total of 218 lines of data in the dataset (i.e., each consisting of two test trials of four data collection runs on the track). The individual ages of the participants were not identified in the released data, although the overall age range was specified as 45 to 65 years old (Perez et al.,

2013, p. 27). The current analysis used data only from test trial 1, and not test trial 3, because NHTSA stated (and the current analysis verified) that it used only the first test trial in setting its glance criteria. Indeed, using only the first test trial is a requirement of the NHTSA (2013, p. 24888) Guideline protocols, "Following the completion of training, each test participant should drive the driving scenario one final time while performing a single instance of the testable task being studied."

Perez et al. (2013) attempted to replicate with each of the 9 tasks, the benchmark radio-tuning task described in the Alliance (2006) Guidelines, which required powering on the radio (or switching from another "function" to the radio), switching from the AM to the FM band (or vice versa), and tuning to a specified frequency that was at least 40 steps above or below the starting frequency. The method of tuning differed depending upon the specific radio and type of control – knob, button, seek, or toggle (see Perez et al., 2013, pp. 5-23). NHTSA considered all these tuning controls to be representative of traditional radio tuning as defined by the Alliance (2006, Principle 2.1B). Perez et al. (2013, p. 23) state, "However, because basic tuning controls (using a knob or a set of buttons) are still available in these newer vehicles, meaningful departures from the specification in the Alliance document for the setup of a reference radio-tuning task appear to be limited. This means that all of the radios could be used to configure the reference task for testing." Perez et al. (2013) then collapsed all 218 trials across all 9 tuning tasks and participants for their final TEORT percentile estimate (see Perez et al., Table 8) – which was later used by NHTSA to set its 12-s TEORT criterion.

RESULTS

Table 2 gives the new test results for the nine tasks. Four tasks (1, 2, 6, and 7, red) did not meet the NHTSA TEORT criterion for at least 21 out of 24 participants; i.e., the "% not meet" (row 4) is greater than 12.5%. This result is confirmed by the "85th percentile TEORT" estimates from Perez et al. (2013, Table 8), reproduced in the next to last row in Table 2 − the same four tasks had 85th percentile TEORT scores > 12 s.

Table 2. Test results for the nine radio-tuning tasks listed in Table 1

Task	1.	2.	3.	4.	5.	6.	7.	8.	9.
# not met	7	8	0	2	0	8	7	0	2
Total n	19	39	40	19	20	20	22	19	20
% not meet	36.8%	20.5%	0.0%	10.5%	0.0%	40.0%	31.8%	0.0%	10.0%
Test Result	not meet	not meet	meet	meet	meet	not meet	not meet	meet	meet
85th percentile TEORT*	15.8	12.6	8.0	11.4	8.0	15.2	13.6	9.5	10.9
85th percentile TGT†	15.3	11.7	7.8	11.2	7.6	14.9	11.9	9.5	10.9

* NHTSA Total Eyes-Off-Road Time (TEORT) 85th percentile data from Perez et al. (2013, Table 8, second data column)
† Alliance Total Glance Time (TGT) 85th percentile data from Perez et al. (2013, Table 8, first data column)

The fact that four of the radio-tuning tasks used by NHTSA to set its 12-s TEORT criterion did not meet that criterion was an unexpected outcome – given that these radio-tuning tasks contributed to the setting of the 12-s criterion. This finding also seems to contradict NHTSA's claim that the Guidelines criteria are benchmarked to traditional manual radio tuning. However, all 9 tasks met the Alliance (2006) Total Glance Time (TGT) criterion of 20 s, which counts glance time only to the device (Table 2, last row). TGT is a subset of TEORT (which counts *all* glances off the forward roadway). In the NHTSA simulator test, there were few glances other

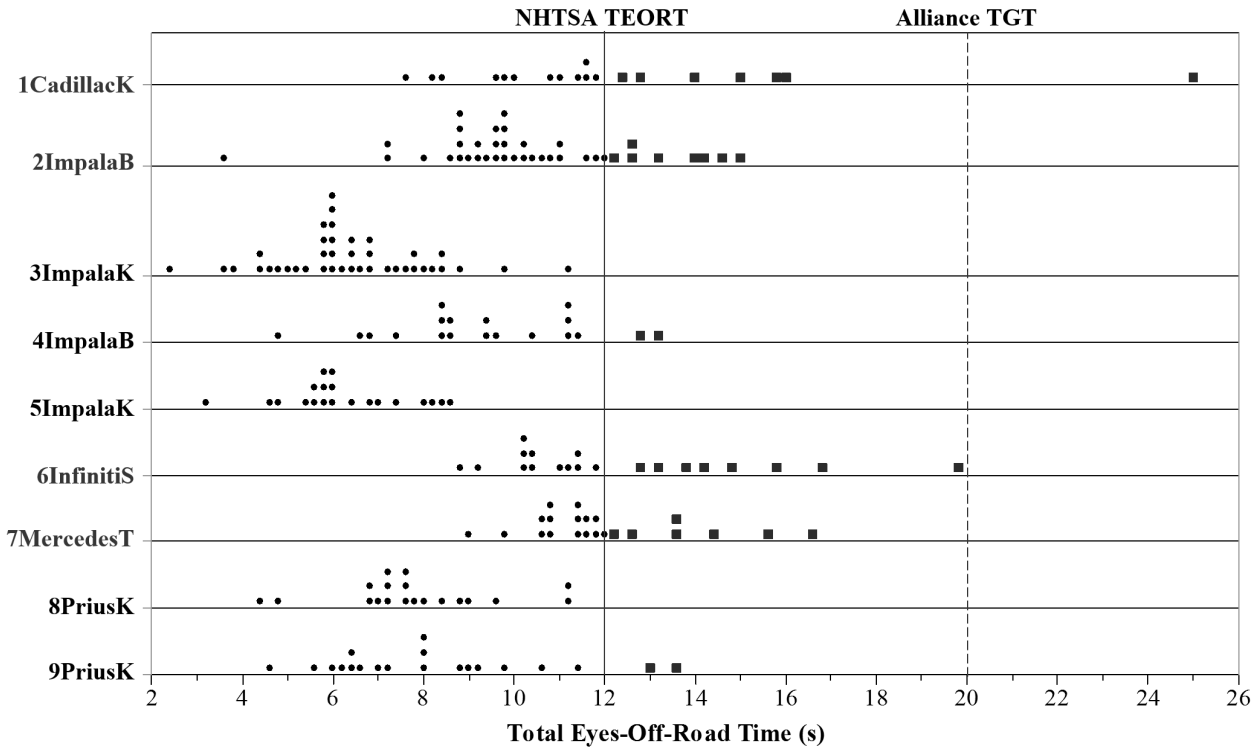

Figure 1. Track test TEORT data. Each dot is a participant's result for the first test trial

than to the device, yielding close values for the last two rows of Table 2. In naturalistic driving, glances to mirrors or out the left and right windows would increase NHTSA's TEORT but not the Alliance's TGT.

Figure 1 (next page) plots the TEORT data for Prius radio tuning for the 218 first test trials. Red symbols indicate trials which did not meet NHTSA's 12-s TEORT criterion (vertical solid red line). All tasks met the Alliance's 20-s TGT criterion (vertical dashed red line). Note that Figure 1 illustrates heterogeneities in the tested tuning tasks, despite the fact that they all were all supposed to represent "traditional radio tuning." For example, as noted by Perez et al. (2013, p. 65) in the Chevrolet Impala vehicle, the button tuning TEORT (10 red dots for 2 and 4 in Figure 1) was significantly worse than knob tuning (0 red dots for 3 & 5 in Figure 1). In addition, knob tuning in the Cadillac (top row in Figure 1) was worse than button tuning in the Impala (rows 2 & 4). In general, the newer 2010 vehicles (Impala and Prius), which were oversampled in 72% of the collapsed trials, had lower TEORT values than the 2005-2006 model years for the other 3 vehicles.

All tasks in both the simulator (Young, 2015) and track study met the MSGD criterion (not shown for space reasons), but the LGP criterion could not be evaluated for the track data because NHTSA has not yet reported individual glance data for each participant, nor subject-level LGP values. The LGP criterion should be re-examined according to those data, and also revised as necessary.

DISCUSSION

Perez et al. (2013, Table 8) collapsed all 218 trials across tasks and participants, to derive their

85th percentile TEORT estimate of 12 s, which NHTSA (2013) used in conjunction with its simulator data (Young, 2015) to set its final 12-s TEORT criterion. Acknowledged heterogeneities in the track data call this method into question. The present study assessed all 9 tuning tasks against the NHTSA TEORT criterion that had been derived from these same data. In spite of NHTSA's premise that its glance criteria permit traditional manual radio tuning, the track results showed that four of the five tested radios had at least one manual tuning task that did not meet NHTSA's TEORT criterion. The present analysis suggests that the likely reason is methodological issues in NHTSA's application of the reference task concept for setting the NHTSA Guidelines criteria. In particular, the Perez et al. (2013) track study was designed for different purposes than setting the final NHTSA (2013) Guidelines criteria, and hence its protocols were not consistent with the final Guidelines test protocols, giving rise to limitations in using that dataset for setting Guidelines glance criteria.

Unaddressed Potential Limitations of Track Study

Questions about the use of only one test trial per participant and equivalency of these tasks to traditional radio tuning were addressed by Young (2015). Other potential limitations are:

1. *The participants using the two most modern radios with the lowest TEORT scores were oversampled*. Perez et al. (2013, Table 8) shows that 247 (80%) of the 347 total participants tested, from which the final NHTSA criteria were derived after pooling, used the two most modern radios (the 2010 Toyota Prius and 2010 Chevrolet Impala) among the 5 tested. These two radios also had the lowest TEORT scores, which, because of being oversampled, biased the TEORT percentiles downwards.
2. *Not age balanced.* Only older participants (45 to 65) were used in the track study, contrary to the NHTSA (2013) Guidelines protocol that also require younger participants (25 to 44 years old).
3. *"Test track" is not an accepted Guidelines test venue* (only occluded goggles and simulator test venues are). Indeed, the metrics collected on a test track versus a simulator are equivalent on a relative but not absolute scale (Young et al., 2005, 2009). In fact, for the one task that matched on the track and simulator (Toyota Prius tuning), the 85th percentile simulator TEORT for participants with matched ages to the track participants (45 to 65), was 12.85 s, or 18% higher than the track 85th percentile (10.5 s) for the combined Toyota tasks 8 and 9 for TEORT for the identical task in the identical vehicle. This result suggests that if all the radios tested on the track had been tested in the simulator, that the TEORT percentiles would have been substantially higher. In other words, NHTSA may have considerably underestimated the TEORT criterion for its required stimulator or occlusion test by using unadjusted track data.
4. *The LGP criterion has not been validated*. The track study was performed in 2010 before NHTSA had defined the LGP metric, so NHTSA (2013) would have to re-analyze or provide the requested track glance data to verify that the radio-tuning tasks it tested met its LGP criterion.
5. *Does TEORT predict relative crash risk or not?* A deeper issue is that NHTSA has not established whether the task scores on its TEORT metric predict relative crash risk that may arise from secondary task physical and cognitive demand (Young, 2012). In fact, Liang et al. (2012) predicted that TEORT would have little relationship to relative crash risk, and Victor

et al. (2014) confirmed this prediction for lead vehicle crashes in the SHRP 2 naturalistic driving data. Both studies indicate that the duration of off-road glances is the more important variable for predicting crash risk, but Victor et al. (2014) finds the increase in risk is a function of longer glances away from the road that occur concurrently with lead vehicle braking. Prediction of crash risk from experimental studies remains an important open issue for further research.

The results of the present analysis indicate the need for NHTSA to revise its glance criteria. Despite these many limitations, this could be done based on a more valid analysis of NHTSA's present data. For example, the track data can be transformed from the track to the simulator venue, using the simulator/track ratio for the Prius task matched for 45-65 years old. Also, predicted track values for the younger ages in the simulator could be made using young/old ratios in the existing simulator dataset. Setting the criteria at the upper confidence limit of the 87.5 percentile would allow the tested tasks to reliably meet the resultant criteria. A substantial upward revision of NHTSA's TEORT criterion is consistent with the on-road experimental studies conducted by Reimer et al. (2014), who found that manual radio tuning had an 85th percentile closer to 20 s than 12 s, which would extend even higher after allowing for variability by using the upper confidence limit.

Unintended Consequences

Automakers who attempt a good faith effort to adhere to the NHTSA (2013) Guidelines will likely face having to decide whether to unnecessarily lock-out or redesign radio-tuning tasks, as well as many other secondary tasks in the vehicle that have glance scores equivalent to or even better than those for radio tuning, unless the TEORT criterion is revised appropriately upward to resolve the task-acceptability-testing issues raised here.

More importantly for safety, however, severely restricting the types of tasks drivers can perform using in-vehicle devices is likely to result in drivers resorting to use of portable and hand-held devices (Strassburger, 2012) whose interfaces have not been designed for use while driving (Young and Zhang, 2015), thus likely causing a net increase in crash risk, rather than a net decrease.

ACKNOWLEDGMENTS

I thank Bruce Papazian, Barbara Wendling, and anonymous reviewers for helpful comments.

REFERENCES

Alliance (2006). S*tatement of principles, criteria and verification procedures on driver-interactions with advanced in-vehicle information and communication systems, June 26, 2006 version*. Wash., DC: Alliance of Automobile Manufacturers Driver Focus-Telematics Working Group.

Liang, Y., Lee, J.D. and Yekhshatyan, L. (2012). How dangerous is looking away from the road? Algorithms predict crash risk from glance patterns in naturalistic driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 54(6):1104-1116.

NHTSA (2012a, February 24). *Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices* [proposed]. (Docket No. NHTSA-2010-0053). Washington, DC: NHTSA. http://www.gpo.gov/fdsys/pkg/FR-2012-02-24/pdf/2012-4017.pdf

NHTSA. (2012b, November 5). Supporting tables 5, 6, and 7 from the initial notice for NHTSA's driver distraction. (Sheet_2.xls). (Docket No. NHTSA-2010-0053). Washington, DC: NHTSA. http://www.regulations.gov/#!documentDetail;D=NHTSA-2010-0053-0126

NHTSA. (2013, April 26). *Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices* [final]. (Docket No. NHTSA-2010-0053). Washington, DC: NHTSA. http://www.regulations.gov/contentStreamer?documentId=NHTSA-2010-0053-0135&disposition=attachment&contentType=pdf

Perez, M., Owens, J., Viita, D., Angell, L., Ranney, T. A., Baldwin, G., Parmer, E., Martin, J., Garrott, W. R., & Mazzae, E. N. (2013, May). *Radio tuning effects on visual test track and driving performance measures–Simulator and test track studies*. Washington, D.C.: NHTSA.

Reimer, B., Mehler, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., & Rumpold, A. (2014). Effects of an 'expert mode' voice command system on task performance, glance behavior, and driver physiology. AutomotiveUI '14, *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 1-9.

Strassburger, R. (2012). *Comments on Visual-Manual Driver Distraction Guidelines for In-Vehicle Electronic Devices*. http://www.regulations.gov/#!documentDetail;D=NHTSA-2010-0053-0104

Victor, T., Bärgman, J., Boda, C.-N., Dozza, M. et al. (2014) *Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk*, Transportation Research Board.

Young, R. A. (2012). Event detection: The second dimension of driver performance for visual-manual tasks. *SAE Int. J. Passeng. Cars - Electron. Electr. Syst., 5*(1).

Young, R. A. (2015). Need for revised total eyes-off-road criterion in the NHTSA distraction guidelines: Simulator and track radio-tuning data. *Poster, Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Salt Lake City, Utah.

Young, R. A., Angell, L. S., Sullivan, J. M., Seaman, S., & Hsieh, L. (2009). Validation of the static load test for event detection during hands-free conversation. *Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Big Sky, MT.

Young, R. A., Aryal, B., Muresan, M., Ding, X., et al. (2005). Road-to-lab: Validation of the static load test for predicting on-road driving performance while using advanced information and communication devices. *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Rockport, Maine.

Young, R A. & Zhang, J. (2015, April). Safe interaction for drivers: Driver behavior metrics and design implications. *Proceedings of the SAE*, Detroit, Michigan, USA, in press.