A TECHNIQUE FOR EVALUATION OF PEDESTRIAN WARNING CONDITIONS WITH HIGH DRIVER ACCEPTANCE.

Jan-Erik Källhammer
Autoliv Development AB
Sweden
Hanieh Toghyani
Linköping University
Sweden
Kip Smith
Naval Postgraduate School
United States
Paper Number 13-0229

ABSTRACT

This paper discusses a research method used to evaluate the design of the alerting logic of automotive active safety systems and presents an example of how this method can be used to support the design of pedestrian collision warning alerts. Three questions therefore arise: First, how can we collect a measure of the acceptability of an alert to a wide variety of situations? Second, how consistent is that measure across contrasting samples of drivers? Finally, how to use the measure in designing alerting logic?

We describe an empirical approach to quantifying the relative level with which drivers are likely to accept pedestrian alerts by a night vision system. The study had two parts: a field operational test (FOT) that gathered a set of 302 video clips of pedestrian alerts with a night-vision system, and a post-hoc or retrospective ratings experiment in which volunteers viewed the clips and rated the relative acceptability of the alerts. We document the consistency of these subjective ratings across groups of raters with different levels of experience with the system. This finding supports the argument that laboratory reviews of FOT data are likely to generalize across the population of drivers.

The derived measure of acceptance was then used to investigate a range of contextual and quantitative factors likely to influence driver acceptance of alerts to pedestrians issued by a night vision active safety system. Least squares regression revealed that nominal characterization of pedestrian location and motion and two quantitative measures – minimum separation and time to closest approach - explain almost 70% of the variance in driver ratings and do not interact. We discuss the implications of this finding for the specification of the system's alerting strategies.

INTRODUCTION

A pre-requisite for a good pedestrian alerting system is reliable detection of pedestrians in a wide variety of ambient conditions, but reliable detection is not sufficient to ensure a system is widely accepted by drivers. The system must also implement alerting strategies that determine whether an alert should be issued to warn the driver about a potentially dangerous situation. The number and level of errors caused by unreliable detection will likely shrink with the evolution of technology, while the levels of other unwanted alerts for accurately detected pedestrians are likely to remain, since they are dependent on factors such as the driver's attention and predisposition.

The safety benefits of any active safety system can materialize if and only if the system will be used. Promoting system acceptance must therefore be a major goal. Accordingly, designers of pedestrian alerting systems need a method that can help them determine the factors that influence driver acceptance of alerts. However, the development of active safety systems has often taken an engineering perspective that emphasizes system accuracy, rather than a human factors perspective that emphasizes concordance with the driver's expectations of system performance.

We have taken the human factors approach to develop a metric of driver acceptance of system alerts to traffic situations. Our approach to evaluating system performance is predicated on the belief that system design should seek to maximize driver acceptance of the system, as driver acceptance is likely to improve if the system activates when the driver finds it reasonable.

Driver acceptance or lack of acceptance of a type of alert should never override concerns raised by the basic physics of the situation. An alert should likely always be issued whenever the situation appears to

be in the process of precipitating a collision. On the other hand, an alert should not be issued when the risk of a collision is sufficiently low to lead drivers to disregard the alarm. Accordingly, the key to promoting adoption of an active safety system is driver acceptance of system response in situations that lie between the extremes, that is, between situations where there is a clear need for alerts and situations where no alert is justified. Due to the low base rate of actual collisions, most of the alerts issued will be in such in-between situations.

Risk perception is an inherent part of the decision-making process. The quality of the risk assessment depends on the adequacy of the available information (Williams & Noyes, 2007). Risk perception does not have to be associated with the actual risk of collisions. Sheridan (2008) defined safety as conditions where risk is acceptable and points out that risk is partly a subjective factor. An active safety alerting system must therefore take into consideration the driver's assessments of the situations where alerts are being issued.

Wiese and Lee (2004) found a strong correlation between rated annoyance and subjective workload and suggested that perceived annoyance should be used in alert design. They pointed at the important trade-off between improving driver response time through increased alert urgency and the annoyance related to highly urgent alerts. The annoyance trade-off is affected by both the alert type and the frequency of the issued false alarms.

Measuring driver acceptance using field operational tests (FOT) data is a reasonable approach, but is faced with the scarcity and expense of FOT data. A second obstacle is that all FOT data are to some extent unique as their collection is not subject to experimental control. The approach we advocate is to leverage FOT data in the laboratory. This method retains a high level of ecological validity by collecting actual incidents on the road. We then make efficient use of the recorded (and rare) field incidents using within-subjects designs, categorical independent variables, and replicable, quantitative dependent measures.

METHOD

At issue here is the consistency, across groups of drivers with differing levels of driving experience, of the subjective ratings of acceptance obtained in the laboratory. We elicited ratings of the relative acceptability of alerts to pedestrian encounters in a set of FOT video recordings. Volunteers rated the acceptability of an alert to each pedestrian encounter.

FOT data collection

Eight drivers drove instrumented vehicles equipped with a Night Vision system with pedestrian detection software. The drivers were recruited at Autoliv in Vårgårda, Sweden, and applied voluntarily to the study. The system records a continuous 'video' to display to the driver and superimposes an alert icon when pedestrians may be at risk.

The Night Vision system consists of a Long Wave, or Far Infrared (FIR) night vision camera mounted in the grille of the vehicle and a video display mounted on the upper part of the center console. The system contains integrated pedestrian recognition software. The display screen is updated at 30Hz with a black and white FIR image. The image is augmented by a flashing yellow alert symbol and by red rectangle(s) that highlight the pedestrian(s) whom the system has detected. The system was installed in eight recent model year Volvo and SAAB vehicles. A PC mounted in the trunk of the car recorded the video clips in a time window before and after an alert. Each car was used for everyday driving by its owner. Subject participation conformed to the ethical guidelines established by Vetenskapsrådet, the Swedish Research Council (2002).

The eight Night Vision systems flagged a large number of video clips with pedestrian encounters like that shown in Figure 1. Back in the laboratory, we selected clips of flagged events for review. The criterion for selection of clips to be reviewed was that there should not be any ambiguity regarding which pedestrian(s) the volunteers were to rate. Groups of pedestrians were allowed, if they were in the same context, for example walking together. Clips with pedestrians visible at different locations were excluded. Presenting the video clips in a random order contributed experimental rigor to the review.

Figure 1. A typical alert issued by the system.

The product of this selection process was a series of 302 video clips of traffic incidents and situations that were the stimuli used in the table-top laboratory experiment. Each clip shows approximately 5 s of images from the FIR camera. The duration of each video clip was kept short to allow each rater to rate a large number of clips within a reasonable time. Each clip ended a few seconds before the pedestrian encounter at approximately the time that an alert would have to be given to provide the driver enough time to take appropriate action. The amount of time between the end of the clip and the pedestrian encounter varied between 0.9 and 12 seconds. The time to pedestrian encounter was measured as either the time to collision or, if the vehicle and pedestrian were not on direct collision path, to the time when the vehicle position would be adjacent to the pedestrian.

Laboratory experiment

Our method is inspired by the hazard perception test used in U.K. driving tests (Jackson, Chapman, & Crundell, 2009). The experimental procedure consists of viewing and rating: Observers in the laboratory watch the replay of a video clip and then rate the level with which they would likely accept an alert from an active safety system to that event.

A laptop computer connected to a video projector presented the video clips on a wall at a distance of 3 m and a horizontal field of view of 40 degrees. No information about the traffic context other than the FIR video clips was provided to the raters. To avoid response bias, we did not query them on their thoughts regarding their criteria for acceptance.

Immediately following the presentation of each clip, the projector froze the last frame of the clip and the laptop display presented the response screen and its scale bar, Figure 2. The experiment was self-paced. The raters used the Next button in the response screen to queue the next clip. The flashing alert symbol was suppressed to avoid any indication whether an alert was issued and if so, the timing of the alert.

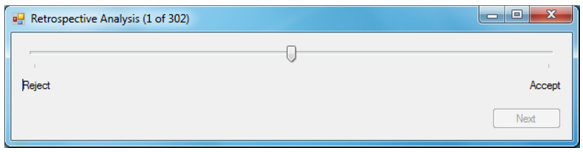

Figure 2. The Response Screen Used in the Experiment.

Instead of the response time collected in the U.K. hazard perception test, our approach quantifies the relative level with which drivers are likely to accept an alert using the approach proposed by van der Laan, Heino, and De Waard (1997). To simplify and clarify the participants' task and to achieve a single measure as in the hazard perception test, we condense the nine scales advocated by van der Laan et al. into a single acceptance score using a continuous scale from 'Reject' to 'Accept'.

The instructions given to the raters were that the scale was linear and that the position on the scale should reflect the degree of rejection or acceptance they had towards an alert to each situation. The raters were asked to provide a rating further to the right on the scale as their acceptance to an alert increased. When they had no objections to an alert they were asked to place their marker at the 'Accept' end of the scale (100%). Likewise, decreasing acceptance to an alert should generate ratings further to the left and would end up at the 'Reject' end of the scale (0%) when they under no circumstances would want an alert to the situation. They were further told that the middle of the scale (50%) was the threshold, above which they would, all things being equal, accept an alert and below which they wouldn't accept an alert. Each rating therefore reflects a judgment by the rater of the relative level at which the situation warrants an alert from the system.

Three groups of participants took part in the experiment. The first was the group of eight drivers from the field study (age: M 53.1 yr., SD 5.8, range 46 to 61). All of the 8 drivers had considerable driving experience (M 34.9 yr., range 27 to 43). They all worked with automotive safety in various capacities. Each of them had experienced some of the reviewed pedestrian encounters. Between them, they had experienced them all.

The second was a group of 42 volunteers (age: M 42.6 yr., SD 10.7, range 24 to 67) recruited from the same facility as the drivers. Most of the 42 had considerable driving experience (M 19.0 yr., range 3 to 48). None of the 42 had experience with the pedestrian alert system in their personal vehicles.

The third was a group of 24 volunteers (age: M 33.4 yr., SD 6.6, range 26 to 52) recruited from a sister company at another location in Sweden. Most of the 24 had considerable driving experience (M 14.2 yr., range 4 to 34). All had experience with the development of the pedestrian detection system. Some were involved with testing and some involved with system and algorithm development. Their experience with the system and inside knowledge of its functionality made them possible candidates for judging the various pedestrian encounters differently.

All participants lived in Sweden and had similar professional backgrounds. The main differentiation is

their experience with the system. To make the rating process less of a burden, the video clips where divided into two subsets and each rater was randomly assigned to one of the two subsets. An exception to the randomization was made for the eight drivers who participated in the field study. They wanted to review the events they had experienced on the road. We therefore arranged the two subsets such that all of the events experienced by a given driver were placed in the same subset (together with other clips collected by other drivers).

RESULTS

Here we present analyses of the consistency of the ratings across the participants. With the consistency established, we use the average rating for each event as the response variable (dependent measure) in a regression analysis. The predictor variables in the regression analysis are nominal characterizations of pedestrian location and motion and a pair of quantitative variables previously found to influence ratings.

Consistency or ratings

We have analyzed the collected ratings to assess their concordance across three groups of raters. The three groups are (1) the laboratory participants without direct experience with the system, (2) the drivers who drove the cars in the FOT, and (3) the laboratory participants who had experience developing the pedestrian detection system. We expected that the three groups would reflect increasing experience with the functionality of the tested system.

We found driver acceptance of alerts to be highly consistent across groups with differing exposure to the system. The consistency of ratings between the raters without direct experience with the system and the other two groups of raters is shown in Figure 3. The cross-plots compare the mean ratings assigned by the raters without direct experience with the system and the mean ratings assigned by the other raters. In Figure 3a, the other raters are the eight drivers who drove the cars in the FOT. In Figure 3b, the other raters are the 24 participants who had experience developing the pedestrian detection system.

The graphs in Figure 3 also show the best-fit least-squares regression equations for the rating data. The agreement between the ratings by the drivers and the other participants are linear and quite good, $r^2 = 0.79$, $F(1, 302) = 1127$, $p < 0.0001$ for the drivers in the FOT and $r^2 = 0.89$, $F(1, 302) = 2462$, $p < 0.0001$ for the 24 engineers. The larger spread among the drivers who drove the cars in the FOT may be due to the small sample size.

We tested the internal consistency of the ratings by applying the Kendall coefficient of concordance to their ratings. This non-parametric test of inter-judge reliability assesses the degree of agreement in the rank ordering of a set of items (e.g., the 302 video clips) by N judges (Siegel & Castellan, 1988). It imposes no categorical dimensions of similarity on rated items. We found them highly concordant; $W = 0.5702$, $\chi 2\ (301) = 6350$, $p < 0.0001$.

a

b

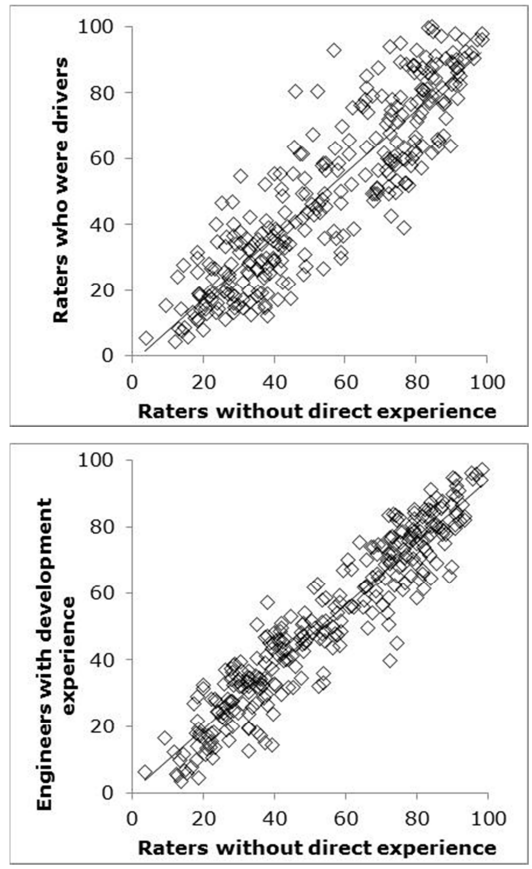

Figure 3. Cross-plots of the average rates of the raters without experience with the system and the average rates of (a) the eight drivers who drove the cars in the FOT and (b) 24 engineers involved in developing the system.

This result encourages us to conclude that the laboratory results are highly consistent. On average, the raters, whether they had experience with the system or not, differentiated among the events in a similar way. This finding supports the contention that the laboratory method of review and rating of events recorded during an FOT study produces data that are consistent between groups of raters with different experience with the system. The high level of concordance implies that the ratings may be aggregated in subsequent analyses of the influence of various parameters on the acceptance of alerts.

Regression analysis

To understand better the influence of pedestrian location and motion on driver acceptance of alerts, we used regression to ascertain the degree to which they explained the observed variability in driver ratings. We created nominal variables that combined categories of pedestrian location and motion, Table 1. Each of these composite categories (e.g., a pedestrian on the left edge of the street who is moving into the street) represents an incursion into the driver's field of safe travel (Gibson & Crooks, 1938). The field of safe travel is an indefinitely bounded field consisting of all unimpeded paths that a vehicle can take at any moment. In a previous study (Källhammer & Smith, 2012), we found some of these incursions to be statistically significant predictors of driver acceptance of alerts.

Table1.
Factors describing combinations of pedestrian location and motion analyzed in this study

Pedestrian location	Combined with pedestrian motion
In Street	None
Right edge	None
Left edge	Same or Opposite Into street
Left side (beyond curb)	Same or Opposite Standing
Right side (beyond curb)	Same or Opposite Standing

We expect ratings to be higher when pedestrians are seen to be within or moving toward the driver's field of safe travel (Gibson & Crooks, 1938). The relative likelihood of pedestrian incursion into the field supports the hypothesis that the acceptance of alerts would be higher when the pedestrian is In the street or on its Right edge than when on the Left edge of the roadway or beyond it on either side.

With regard to Pedestrian Motion, we expect alerts to receive higher ratings when the pedestrian is walking into the street than when standing or walking parallel to the street. The category 'Into street' implies that the pedestrian was walking essentially perpendicularly to the direction of vehicle travel and into its field of safe travel. The categories 'Same' and 'Opposite' are reserved for pedestrians walking in a direction predominately parallel to the vehicle's path in either the same or opposite direction as the vehicle.

When Right edge was used as a separate variable, it interacted with combinations with motion variables and was thus excluded from the combinations. The combinations of Left edge and Standing, Left side and Into street, and Right side and Into street were excluded due to too few cases.

Each of the 302 clips from the FOT was reviewed and categories determined for these factors. Each category was partitioned into a single nominal (dummy) variable at two levels 1 and 0. The categories were the predictor variables in the analyses. The response variable was the average ratings of acceptance of an alert to the events in the video clips.

In addition, the analysis of the motion data used the observed distance in meters between the pedestrian and the car. The observed distance was measured in the direction of the vehicle's travel. By convention, X is positive in the direction of travel. Because the sensor was directed ahead of the car, X distances to pedestrians are always positive. Time to Impact (TTI), measured as time to collision or time to a vehicle position adjacent to the pedestrian if the vehicle and pedestrian were not on direct collision path, was used to describe the time-distance relationship.

The least squares regression equation is shown in Equation 1, the summary statistics in Table 2, and the fit of the model in Figure 4a. The ten parameters explain nearly 70% of the variability in the drivers' ratings of the alerts issued in the 302 video clips. None of their interactions are significant. The ability to explain 70% of the variability in human judgment is both unusual and, we believe, impressive.

$$\text{Rating} = 74.0 + 18.9*L1 + 12.6*L2 - 18.8*M1 - 36.1*M2 - 25.9*M3 + 16.7*M4 - 25.9*M5 - 20.0*M6 - 0.2*\text{Min } X - 0.6*TTI \quad (1).$$

As expected, factors that indicate that pedestrians are either within or moving toward the driver's field of safe travel prove to be highly significant predictors of alert acceptance. A positive beta weight increases the predicted rating of pedestrians in the path of the approaching vehicle (In street or Right edge) and for pedestrians moving from the Left edge into the street. A negative beta weight decreases the predicted rating of pedestrian motion parallel to the street or pedestrian standing on either side of the street. The weights become increasingly negative the further away the pedestrian is located from the path of the vehicle. The further away the pedestrian is located from the on-coming car and the longer the TTI also reduce the predicted rating.

Table2.
Summary of the Regression Model with 10 Predictors

Variable	Equation symbol	B	t	p
Intercept		74.0	14.6	<0.001
In Street	L1	18.9	3.7	<0.001
Right edge	L2	12.6	2.7	0.008
Left edge and (Same or Opposite)	M1	-18.8	-3.6	<0.001
Left side and (Same or Opposite)	M2	-36.1	-6.8	<0.001
Right side and (Same or Opposite)	M3	-25.9	-5.5	<0.001
Left edge and Into Street	M4	16.7	2.8	0.006
Left side and Standing	M5	-25.9	-4.7	<0.001
Right side and Standing	M6	-20.0	-3.9	<0.001
MinX	Min X	-0.2	-3.6	<0.001
TTI	TTI	-0.6	-4.1	<0.001

To illustrate the importance of pedestrian location, we tested a reduced model with three predictor variables, one categorical variable that combined two categories of pedestrian location - In street and Right Edge - and two quantitative variables, MinX and TTI. The simplified model explains almost 54% of the variability in the drivers' ratings. As shown in Figure 4b, the reduced model produces two layers or groups of predicted values.

The consistency of model and the average rate by all participants is shown in the two graphs of Figure 4. The graphs also show the best-fit least-squares regression equations for the rating data. The agreement between the predicted and observed average ratings by all the participating raters is quite good, $r^2 = 0.70$, $F(1, 302) = 696$, $p < 0.0001$ for the 10 variable model and $r^2 = 0.54$, $F(1, 302) = 347$, $p < 0.0001$ for the three variable model.

DISCUSSION

A key finding is the consistency of subjective ratings of the acceptance of alerts between groups with different experience with the system. Raters without experience in the field produce reliable and reproducible data that align with the experience of drivers in the field. This lends credence to the method, its reliability, and its application to events recorded by FOTs.

a

b

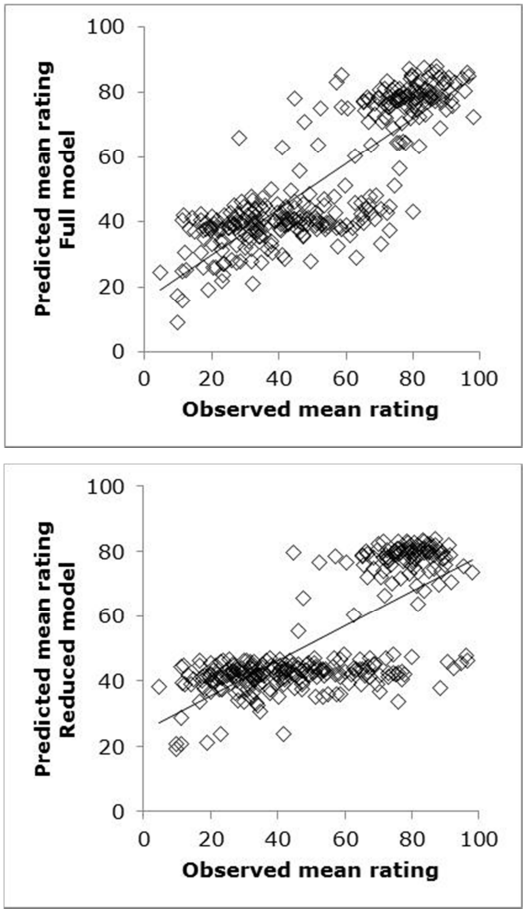

Figure 4. Cross-plots of the average rates of all participants and the modeled rates of (a) the full 10 variable model and (b) reduced three variable model.

A major challenge to FOT studies is that most of the observed events are unique in various ways. The everyday context makes it difficult to experimentally control and accurately repeat trials (Walker, Stanton & Young, 2008). Using recordings of these events in a laboratory environment provides experimental control of the stimuli while retaining much of the original ecological validity. The subjective rating method presented here leverages the scarce and expensive FOT data and help bridges the field and the laboratory. By eliciting responses from a large number of observers, we leverage the high cost of the FOT and generate sample sizes that are amenable to statistical tests of significance.

Although the method was developed to address the analysis of field data, it is applicable to simulator studies as well. Smith & Källhammer (2010) used it in a simulator study to assess the risk posed by

intersection encroachments and how that level varies across situations. The method has also been used by Källhammer, Smith, Karlsson, and Hollnagel (2007) to elicit drivers' assessments of a variety of naturalistic traffic situations.

By enhancing our understanding of when and why drivers accept system alerts, we are better able to develop warning strategies that will likely lead to higher levels of driver trust and system acceptance.

CONCLUSIONS

In this paper we have described an empirical approach quantifying the relative level with which drivers are likely to accept pedestrian alerts by a night vision system. Analyses of the ratings data from participants found consensus across volunteers with different levels of system experience about how they rate the acceptability of system alerts to pedestrian encounters. The method will therefore enable us to leverage expensive FOT data. The work demonstrates the utility of subjective driver acceptance criteria as a tool to inform the development of the system's alerting criteria.

System performance optimized based on driver acceptance rather than objective performance criteria should better match driver expectations and thereby promote system acceptance. System use and safety benefits from the systems should increase with higher user acceptance. The derived metric of acceptance can be used to improve alerting criteria and help uncover factors that influence when alerts should be issued.

A nominal characterization of pedestrian location and two quantitative measures of pedestrian location with respect to the approaching vehicle explain almost 54% of the variance in driver ratings. About 70% of the variance in driver ratings can be explained by extending the model with additional factors that refine the description of a pedestrian within or moving toward the driver's field of safe travel (Gibson & Crooks, 1938). Designers of pedestrian crash warning systems need to be aware of the contextual sensitivity of driver expectations and assessments.

Limitations

Participants are known to develop expectations for staged events or alerts not only during the course of a simulator study but also when they are exposed to those events in the field (Vogel, Kircher, Alm, & Nilsson, 2003). It is unclear how that may affect the ratings elicited in our experiments.

All of the video clips were collected in Sweden and all participants were Swedish. Additional studies using material collected in other countries and with participants with no experience of Swedish traffic are needed to verify that the method is applicable to the global population of drivers.

Driver state measures such as driver distraction and fatigue are difficult to assess using this method. Further research may test whether the method can be extended to other traffic situations and other types of active safety systems.

REFERENCES

Gibson, J. J. & Crooks, L. E., "A theoretical field-analysis of automobile driving", The American Journal of Psychology, 51, 453-471, 1938.

Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. Ergonomics, 52, 154 – 164.

Källhammer, J.-E. and Smith, K. (2012). Assessing Contextual Factors that Influence Acceptance of Pedestrian Alerts by a Night Vision System. Human Factors, 54, 654-662.

Källhammer, J. E., Smith, K., Karlsson, J., & Hollnagel, E. (2007). Shouldn't the vehicle react as the driver expects? Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design. Stevenson, WA, 9-15.

Sheridan T. B. (2008). Risk, Human Error, and System Resilience: Fundamental Ideas. Human Factors, 50, 418–426.

Siegel, S., & Castellan, N. J., Jr. (1988). Nonparametric statistics for the behavioral sciences, 2nd Ed. New York: McGraw-Hill.

Smith, K. and Källhammer, J.-E. (2010). Driver acceptance of false alarms to simulated encroachment. Human Factors, 52, 446-476.

van der Laan, J.D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. Transportation Research - Part C: Emerging Technologies, 5, 1-10.

Vetenskapsrådet. (2002). Forskningsetiska principer inom humanistisk-samhällsvetenskaplig forskning. [Ethical principles for research using human participants.] Stockholm: Vetenskapsrådet.

Vogel, K., Kircher, A., Alm, H., & Nilsson, L. (2003). Traffic sense—which factors influence the skill to predict the development of traffic scenes? Accident Analysis and Prevention, 35, 749–762.

Walker, G.H., Stanton, N.A., & Young, M.S. (2008). Feedback and driver situation awareness (SA): A comparison of SA measures and contexts. Transportation Research - Part F, 11, 282–299.

Wiese, E. E. & Lee, J. D. (2004). Auditory alerts for in-vehicle information systems: The effects of temporal conflict and sound parameters on driver attitudes and performance. Ergonomics, 47, 965 - 986.

Williams, D. J. & Noyes, J. M. (2007). How does our perception of risk influence decision-making? Implications for the design of risk information. Theoretical Issues in Ergonomics Science, 8, 1-35.