

BRIDGING LABORATORY AND FIELD STUDIES

Jan-Erik Källhammer

Autoliv Development

Sweden

Kip Smith

Cognitive Engineering and Decision Making

USA

Paper Number 11-0215

ABSTRACT

The method we present here - retrospective review and rating of Field Operational Test (FOT) data - is designed to capture both the rigor of the laboratory and the ecological validity of the field. It is tailored for studies of driver acceptance of active safety systems. The method makes it possible to leverage expensive FOT data within the confines of the laboratory.

INTRODUCTION

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 an overriding goal. System acceptance is linked to driver acceptance of the issued alerts. Designers of active safety systems therefore need a method that can help them determine the factors that influence driver acceptance of alerts. Measuring driver acceptance using 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 behavioral research community has long acknowledged the need for methods that capture both the rigor of the laboratory and the ecological validity of the field (Brehmer & Dörner, 1993). We therefore need a method to leverage FOT data by analyzing it under experimental control. The method we present here – retrospective review and rating of FOT data - is designed to bridge the laboratory and the field. It is tailored for studies of driver acceptance of active safety systems.

A large body of research on active safety systems has been conducted in driving simulators (e.g., Caird, Chisholm, Edwards & Creaser, 2007; Hancock & deRidder, 2003; Smith & Källhammer, 2010). Driving simulators have often been a main tool in this research for four important reasons. First, the simulator allows measurement of realistic driver

responses to accurately controlled situations. Driving simulators can reproduce the studied situations quite well and do so under full experimental control. They allow precise response metrics to be collected with high fidelity (e.g., Liang, Reyes, & Lee, 2007). Second, traffic context can be generated with high detail. Driver behavior has been found to be sufficiently natural and to conform to that in naturalistic situations (e.g., Lee, McGehee, Brown, & Reyes, 2002). Third, simulator studies allow evaluation of active safety systems in collision-likely situations that are too dangerous to be reproduced on the road (Lees & Lee, 2007). Finally, the actual system does not need to be implemented before a study can be undertaken.

Another body of research has focused on field or naturalistic studies. Naturalistic driving can be defined as normal driving occurring in its everyday context which is, by definition, not under experimental control. Field studies capture the full range of contextual information and in an environment where mistakes may have serious consequences. The “100-Car Naturalistic Driving Study” (Transportation Research Board of the National Academies [TRB], 2005) and the several FOT studies sponsored by NHTSA, the European Union, and others have triggered a lot of interest in naturalistic and FOT studies. Such studies are seen as a means for obtaining data with high ecological validity.

The increase in ecological validity associate with FOT studies (compared to simulator studies) comes at the cost of experimental control. A major challenge to naturalistic studies is that most of the observed events are unique in various ways. The everyday context makes it difficult to exert the control needed to repeat trials accurately and to identify sequences of observed events that truly replicate (Walker, Stanton & Young, 2008). Collecting a large set of similar naturalistic events (e.g., a single pedestrian crossing the road from left to right) is time consuming. The variability within each set is likely to reflect a number of factors that may differentially effect the

drivers' reactions. Assessing the sources and importance of those differences is difficult.

Even with a large set of naturalistic data, the base rate of a collision is so low that the likelihood of being able to capture an actual collision or even a close call is very low. The low base rate limits the utility of field studies at evaluating collision-likely situations. Even if it were technically possible to stage such situations, the studies would likely not be ethically acceptable (Kiefer, Flannagan, & Jerome, 2006).

Assessing active safety system performance from driving data is not straightforward. It can be difficult to establish whether the issued alert was timely or whether an initiated or pending driver action would have eliminated the potential collision (McLaughlin, Hankey & Dingus, 2008). The issue of what actually constitutes a signal (here correct alert) rarely arises in a laboratory setting, but in a naturalistic setting the definition of a signal often depends on contextual factors (Parasuraman, Masalonis & Hancock, 2000). Understanding the range of signal types and levels are essential to understand the driver's perception of a given signal, which is a challenge to capture in a simulator and to extract from naturalistic data.

Driving simulators represent the rigor of the laboratory while naturalistic and FOT studies represent the ecological validity of the field. Some limitations are in fact shared by both driving simulator and field studies. For example, 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). Therefore, neither driving simulator nor field studies have been found to be fully satisfactory for all important aspects of automotive research. Both have advantages and disadvantages and should be viewed as complementing each other, rather than competing.

The appropriate tools and their requirements have to be assessed by considering the aim and constraints on the study. A balanced approach would use elements from both approaches and combine naturalistic and simulator research. Taking the best of both worlds and combining them would therefore be a rational approach. The approach we advocate is to leverage FOT data in the laboratory. Naturalistic driving with its high ecological validity generates the stimuli used to elicit drivers' assessments and responses in a controlled setting like that provided by a driving simulator. Subsequent FOT studies may then provide verification for the issues being studied in the simulator (Walker et al., 2008). This method retains a high level of ecological validity by collecting actual incidents on the road, which are expensive and time-

consuming to collect. We then make efficient use of the recorded rare field incidents using within-subjects designs, categorical independent variables, and replicable, quantitative dependent measures.

METHOD

The approach is inspired by the hazard perception test used in U.K. driving tests (Jackson, Chapman, & Crundell, 2009). The method presents to observers a set of video recordings of events captured during a FOT. We have evaluated the method using a prototype Night Vision system with pedestrian detection that flags events (e.g., a pedestrian standing in the road) in the field. The system records a continuous 'video' to display to the driver and superimposes an alert icon when pedestrians may be at risk.

FOT data collection

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. A snapshot of a pedestrian alert is shown in Figure 1.



Figure 1. A typical alert issued by the system.

The system was installed in ten recent model year Volvo S80, Volvo V70, and SAAB 9-5 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.

The ten Night Vision systems flagged a total of 88 video clips with pedestrian encounters. Back in the laboratory, we selected clips of flagged events for review, excluding clips with possible ambiguity regarding which pedestrian(s) triggered the alert. Groups of pedestrians were allowed if and only if they were walking together or defined a common context. Clips with pedestrians visible at different locations within each clip were excluded. Bicyclists were also excluded as their patterns of motion and levels of perceived risk are likely to be different than those of pedestrians. The final set contained 57 clips of pedestrian events. Each clip shows approximately 30 s of images, roughly 20 s before and 10 s after the recorded alert. The 57 video clips were the stimuli used in the subsequent laboratory experiment.

Laboratory experiment

The experimental procedure consists of viewing and rating: The participants watch the replay of an event and then individually rate the level with which they would accept an alert from an active safety system to that event. Randomizing the order of presentation adds experimental rigor to the review. The laboratory setup consists of a PC laptop connected to a video projector that presents the set of video clips on the wall at a distance of approximately 3 m and a horizontal field of view of approximately 40 degrees.

Two groups of participants took part in the experiment. The first was the group of 10 drivers from the field study. The second was a group of 25 other volunteers whom we refer to as *non-drivers* even though they all were experienced, licensed drivers. None of the non-drivers had experience with the pedestrian alert system in their personal vehicles. Both the group of drivers and the non-drivers had considerable driving experience (drivers M 30.9 yr, range 22 to 41, non-drivers M 24.2 yr, range 10 to 46). Subject participation conformed to the ethical guidelines established by Vetenskapsrådet, the Swedish Research Council (2002).

Response measure

There does not appear to be a standard method for assessing driver acceptance of new automotive technology. The most widely used method may be that described by van der Laan, Heino, and De Waard (1997). It has been used to compare driver responses to a variety of systems. The method measures driver's acceptance by asking participants to rate the system output using a differential scale with opposing adjectives to anchor the scale with a neutral reference point.

Instead of the response time collected in the hazard perception test, our approach quantifies the relative level with which drivers are likely to accept an alert. We condensed the two components usefulness and satisfying used by van der Laan et al. (1997) into a single acceptance score using a scale from *completely reject* to *completely accept*. By using a single measure, we seek to avoid any confound posed by drivers' potentially varying interpretation of the different components of the van der Laan metric. The values can be entered either on paper or, preferably, directly on the computer using a slider bar, Figure 2.

Immediately following the presentation of each clip, the projector screen showed the frozen last frame of the video clip and the PC presented the response screen shown in Figure 2. The response screen contained a scale bar and two buttons labeled Replay and Next.

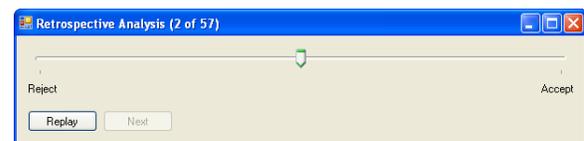


Figure 2. Response screen used in the experiment to elicit ratings of the level of acceptance of an alert.

The experiment was self-paced. The alternation between stimuli projected on the wall and the response screen reinforced the sequential but linked nature of the viewing and the rating. The participant could replay a clip by selecting the Replay button. Each of the 35 participants rated all 57 clips.

RESULTS

We have analyzed the collected ratings to assess their concordance across three groups of raters. The three groups are (1) the drivers who experienced an event in the field, (2) the other nine drivers who participated in the FOT, and (3) the non-drivers who did not drive in the FOT. The obtained ratings are ranked to control for individual differences in scale use.

The consistency between the drivers who experienced the events and the other two groups of raters is shown in Figure 3. The cross-plots compare the ranks of the ratings assigned by the driver who experienced the event and the average ranks assigned by the other raters. In Figure 3a, the other raters are the other nine drivers, while in Figure 3b; the other raters are the 25 non-drivers. 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 driver and the nine other drivers as well as between the driver and the non-drivers is linear and quite good, $r^2 = 0.59$, $F(1,55) = 78$, $p < .001$ for the nine other drivers and $r^2 = 0.52$, $F(1,55) = 59$, $p < .001$ for the non-drivers. The slope of the regression line is less than 1.0, reflecting a regression to the mean; the ratings from the observers in the laboratory are less extreme than those of the drivers who experienced the events.

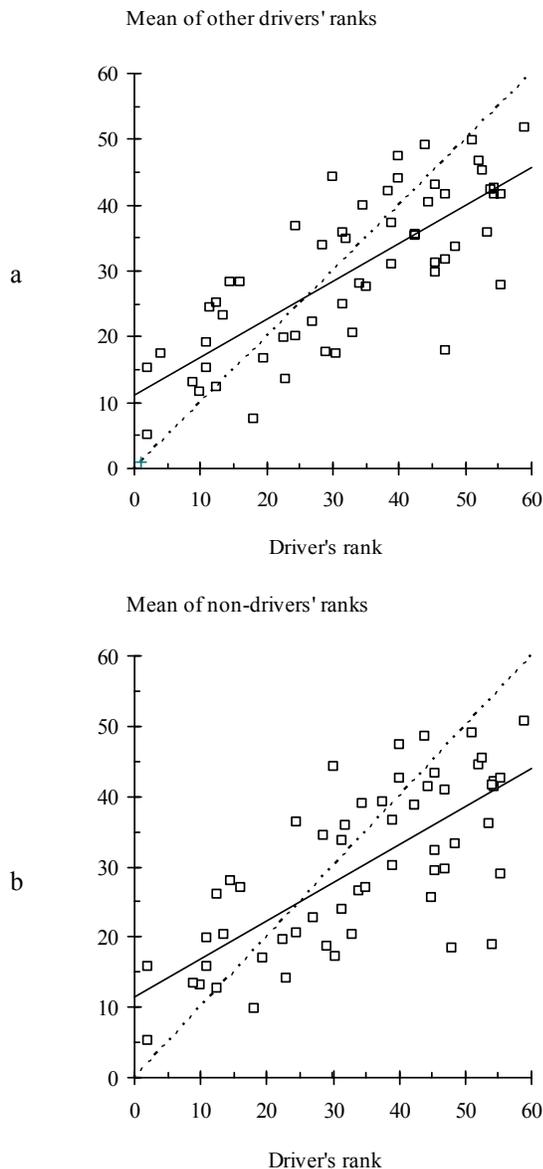


Figure 3. Cross-plots of the ranks of the driver's ratings and the average ranks of (a) the other nine drivers and (b) the 25 raters who did not participate in the FOT (non-drivers).

Inspection of the correlation analysis shown in Table 1 finds that, for nine of the ten drivers, the validity coefficients comparing responses from the original FOT driver and the responses of drivers who had not seen the event are significant at $p < 0.001$.

Table 1.

Drivers' rating consistency

Driver	r_a
1	0.87**
2	0.79**
3	0.72**
4	0.78**
5	0.76**
6	0.81**
7	0.88**
8	0.41*
9	0.82**
10	0.58**

Note. r_a Correlation the driver's rankings of the 57 events with the means of the other nine drivers' rankings. * $p < .01$. ** $p < .001$.

We tested the internal consistency of the ratings provided by the 35 laboratory raters by applying the Kendall coefficient of concordance to the ranks of 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 57 video clips) by N judges (Siegel & Castellan, 1988). It imposes no categorical dimensions of similarity on rated items. After correcting for the numerous ties in the intra-judge ranks, we found them highly consistent; $W = 0.55$, $\chi^2(56) = 1247$, $p < 0.0001$. This result encourages us to conclude that the laboratory results are highly consistent. On average, the raters, whether they had driving 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 align with the drivers who experienced the events in the field. The high level of concordance implies that the ranks may be aggregated in subsequent analyses of the influence of various parameters on the acceptance of alerts.

DISCUSSION

The experimental setup enables presentation of representative situations and should have good predictive validity of the environmental cues. The good quality of the FOT recordings retains much of the ecological validity of actual traffic events. The

ratings provide the rigor of the laboratory. Use of the recorded incidents in a laboratory environment provides experimental control of the stimuli while retaining much of the original ecological validity. Fully situated contextualization is, of course, achieved only in the moment.

The method produces reliable and reproducible data that align with the experience of drivers in the field. 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.

We are using this bridging of the field and the laboratory to inform our design of active safety systems. The level of acceptance for various situations rated can be used to define decision criteria for the active safety systems that should result in higher user acceptance of the safety system. Although the method was developed to address the analysis of field data, the method 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.

CONCLUSIONS

Our retrospective review and rating method produces reliable and reproducible data that align with the view of the drivers who experienced the situations in the field. By eliciting responses from a large number of observers, we leverage the high cost of the FOT data and generate sample sizes that are amenable to statistical tests of significance.

We offer our retrospective review and rating method as a cost-effective approach to bridging the laboratory and the field. Its findings are informing our system design and development.

Limitation

A major limitation of the method is the performance of the system used to record events in the field. False alarms can make the selection of video clips time consuming. Driver state measures such as driver fatigue are also difficult to assess using this method. Both the drivers in the FOT and the other raters in our study had considerable driving experience. All were Swedes. Additional studies with participants with less experience and other demographic backgrounds are needed to verify that the method is applicable to the global population of drivers. Further research may test whether the method can be

extended to other traffic situations and other types of active safety systems.

REFERENCES

- Brehmer, B., & Dörner, D. 1993. "Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study." *Computers in Human Behavior*, 9 (2-3), 171.
- Caird, J. K., Chisholm, S. L. Edwards, C. J., & Creaser, J. I. 2007. "The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior." *Transportation Research Part F* 10, 383-396.
- Hancock, P. A., & deRidder, S. N. 2003. "Behavioural accident avoidance science: Understanding response in collision incipient conditions." *Ergonomics*, 46, 1111-1135.
- 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.
- Kiefer, R. J., Flannagan, C.A., & Jerome, C. J. 2006. "Time-to-collision judgments under realistic driving conditions." *Human Factors*, 48(2), 334-345.
- Lee, J. D., McGehee, D. V., Brown, T. L., & Reyes, M. L. 2002. "Collision warning timing, driver distraction and driver response to imminent rear-end collisions in a high-fidelity driving simulator." *Human Factors* 44(2), 314-334.
- Liang, Y., Reyes, M. L., & Lee, J. D. 2007. "Real-time detection of driver cognitive distraction using support vector machines." *IEEE Transactions on Intelligent Transportations Systems*, 8(2), 340-350.
- McLaughlin, S.B., Hankey, J.M. & Dingus, T.A. 2008. "A method for evaluating collision avoidance systems using naturalistic driving data." *Accident Analysis and Prevention*, 40, 8–16.
- Lees, M.N. & Lee, J.D. 2007. "The influence of distraction and driving context on driver response to imperfect collision warning systems." *Ergonomics*, 50, 1264–1286.
- Parasuraman, R., Masalonis, A.J., Hancock, P.A. 2000. "Fuzzy Signal Detection Theory: Basic Postulates and Formulas for Analyzing Human and Machine Performance." *Human Factors*, 42, 636–659.
- Siegel, S., & Castellan, N. J., Jr. 1988. "Nonparametric statistics for the behavioral sciences", 2nd Ed. New York: McGraw-Hill.

Smith, K. & Källhammer, J.-E. 2010. "Driver acceptance of false alarms to simulated encroachment." *Human Factors*, 52, 466-476.

Transportation Research Board of the National Academies. 2005. "100-Car Naturalistic Driving Study." retrieved http://144.171.11.107/Main/Blurbs/100Car_Naturalistic_Driving_Study_155990.aspx.

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.] Vetenskapsrådet, Sweden.

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.