Influence of Expertise on the Judgment of Controllability of Advanced Driver Assistance Systems

Patrick Galaske[a]*, Mehdi Farid*[a] *and Klaus Bengler*[b]

[a]*Department for Integral Safety*
BMW Group
Knorrstraße 147
D-80788 München, Germany

[b]*Lehrstuhl für Ergonomie*
Technische Universität München
Boltzmannstraße 15
D-85747 München, Germany

ABSTRACT

Modern passenger vehicles are equipped with a rising number of advanced driver assistance systems (ADAS). The increasing complexity of these systems causes issues of controllability that need to be dealt with. The RESPONSE 3 Code of Practice (CoP) provides methods to assess the controllability of ADAS within the framework of ISO26262. Among the methods described in the CoP is the expert review (RESPONSE 3, 2009). However, no quantifiable requirements for such expert reviews are given. This paper describes a vehicle simulator study that aims to compare the judgment behavior of experts with that of naïve participants to draw conclusions on the applicability of expert reviews of controllability. The results of the study show that for the studied sample of experts there is no advantage in the variance of the obtained judgments for either group. The higher expertise however did exhibit itself in a trend towards more critical judgments of the observed situations. It is concluded that the application of expert reviews in the evaluation of controllability of ADAS should be studied in more detail. The results indicate that the conduct of expert reviews of controllability with high requirements of precision is not trivial and requires in-depth analysis.

Keywords: ADAS, RESPONSE 3, Controllability, Expert review

INTRODUCTION

The number and complexity of advanced driver assistance systems (ADAS) in modern passenger vehicles has been rising continuously in the last decades. Modern driver assistance systems not only temporarily inform or warn the driver about potential hazards in specific situations but can assist with or take over parts of the vehicle control for prolonged periods of time. Examples for such systems already in production are adaptive cruise control, active lane keeping assist or park assist systems with active vehicle control. It is foreseeable that in the future the number and complexity of such systems is going to increase even further.

The ability of modern ADAS to strongly influence the vehicle's dynamics is causing concerns about the controllability of these systems in the case of a system malfunction or when the ADAS reaches its functional limits.

https://openaccess.cms-conferences.org/#/publications/book/978-1-4951-2097-8

These concerns are going to increase even further with the ongoing development of ADAS and the shift towards partially automated driving. The RESPONSE 3 Code of Practice has laid down rules for the assessment of controllability in the framework of ISO 26262. Two of the provided methods for controllability assessment are simulated and real vehicle studies with naïve test subjects. These methods are difficult to apply however when the number of factors that need to be considered increases as the number of controllability-relevant situations that need to be evaluated increases in a combinatory fashion. This is especially true when interaction effects between several systems are being looked at. Furthermore such naïve test subject studies can only realistically prove controllability at the 90% level (Weitzel & Winner, 2012) and are therefore irrelevant when higher levels of controllability are required.

A different method for controllability assessment that is suggested in the RESPONSE 3 Code of Practice is the expert review. However the Code of Practice doesn’t provide details on how these must be conducted to warrant a sufficiently precise assessment of the relevant risks to ensure the intended level of controllability. The aim of this paper is to provide insight into how well expert reviews can perform in the assessment of the controllability of driver assistance systems.

In this article we describe a simulator-study performed using a static simulator of the BMW Group that compares the judgment of controllability for two different groups of subjects in several controllability-critical situations.

THEORY

Expert reviews have been utilized in other areas of research and development. Examples for areas of wide spread use of expert reviews are nuclear safety (Cooke & Goossens, 2000), biological safety (Burgman, Fidler, McBride, Walshe & Wintle, 2006), aviation (Harper & Cooper, 1986) and economics (De Bondt, 1991). The reason for this is that in these areas there are many influencing factors that interact in a non-trivial manner. Furthermore typically areas where expert reviews are commonly used are not easily available for detailed study. Therefore it can be difficult to empirically develop validated models or identify relevant parameters precisely for these areas. Under such circumstances other methods of analysis are difficult or even impossible to apply and thus expert reviews are used. Similarly, more empirical methods of assessment can be difficult to develop when no measure of accuracy for the method of assessment is available.

At least since the observations of biases (Tversky & Kahneman, 1974) the ability of experts to perform better than other forms of assessment has come under scrutiny. In some fields it was possible to identify situations where experts with significant subject matter expertise judged in a non-rational ways (Englich, Mussweiler & Strack, 2006). Attempts at debiasing experts have resulted in limited success (Fischoff, 1981). Where expert reviews are unavoidable it has been possible to develop elicitation methods that limit overestimation of the given judgments’ precision (Burgman, 2006). The experiences from other fields have shown that the decision for the application of human judgment should be a conscious one that weights the advantages and disadvantages of the available alternatives (Kahneman, 2011).

Whenever experts are discussed the question of definition of experts arises. One can generally distinguish absolute and relative expertise (Ericsson & Anders, 2006). In short, absolute experts measurably excel at a given task and mark the pinnacle of contemporary domain specific human ability while relative experts are defined by their relative capabilities compared with a certain point of reference. Given a suitably low point of reference it is therefore much easier to acquire relative experts than absolute experts. This article focuses on the analysis of the effect of relative expertise because this effect is more relevant for the practical application of expert reviews in controllability.

When determining the controllability of advanced driver assistance systems it is necessary to establish a requirement for the level of accuracy of the method of assessment being used. This is to ensure that the operator of the ADAS is not being subjected to intolerable risks when using the system. One agreed standard method of controllability assessment is the naïve test subject study for the 85% controllability level (RESPONSE 3, 2009). If expert reviews provide a lower risk of wrong decisions they can be used to replace such studies. The study described in this article aims to compare an expert review with such a study. The goal is to find clues about how well expert studies can realistically perform and how they can be utilized when assessing the controllability of driver assistance systems.

In the framework of the RESPONSE 3 CoP a naive test subject study used to prove C2-controllability consists of 20 participants. Binary pass/fail criteria are set up based on objective measures. The assistance system is said to pass the test if none of the participants failed any of the fail-criteria (RESPONSE 3, 2009). This binary measure of controllability is however unsuitable for the comparison of judgment behavior of experts and naïve test subjects. First, it will generally be difficult to identify 20 experts on the subject of controllability with sufficient subject-matter experience. Second, experts on the topic of controllability often have undergone advanced vehicle handling training and are therefore more likely to not fail the pass-criteria than naïve test subjects. Third, binary pass/fail criteria are unsuitable for comparison between the two methods of assessment because valid conclusions would necessitate an impractically large number of trials. Neukum, Lübbeke, Krüger, Mayser and Steinle (2008) provided a scale for the elicitation of disturbance that has been successfully applied to controllability-studies. The used scale has the advantageous property of approximate interval level of measurement. This allows for the comparison of naïve test subject studies and expert reviews on the same scale and makes it possible to evaluate the magnitude of the differences between the judgments.

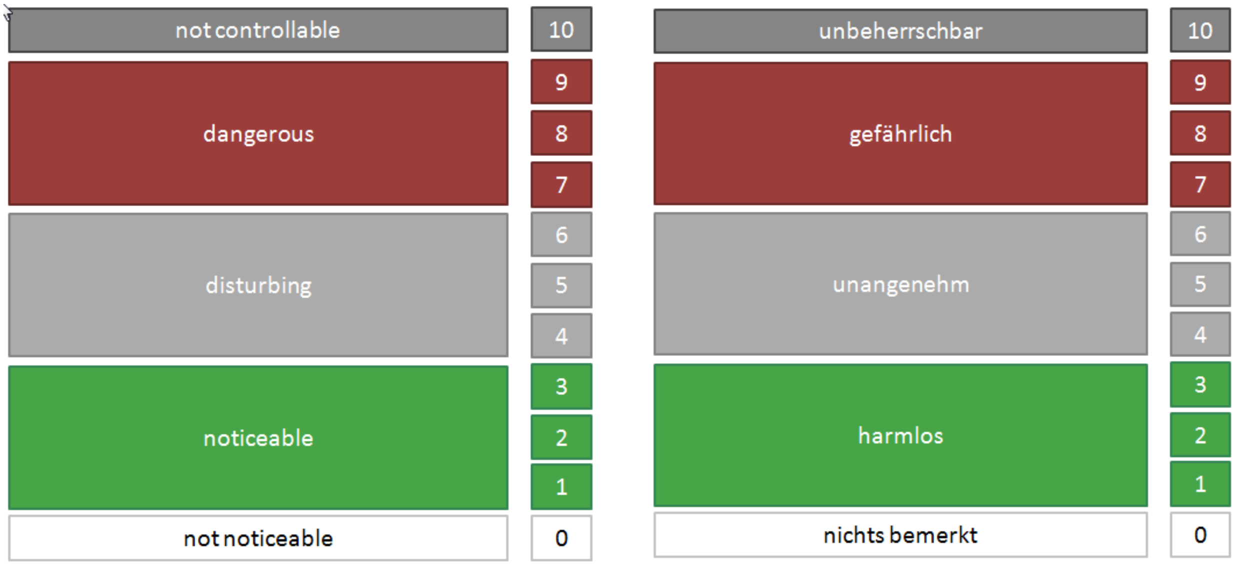

Figure 1 English and German adaptions of the scale of disturbance adapted from (Neukum et al., 2008).

When using experts to assess the controllability of driver assistance systems one would ideally hope for experts to make better judgments than naïve test subjects would. Otherwise the status of expert would likely have been attributed wrongly. Since there will only be one correct judgment high expertise should reveal itself as a lower distance of the judgments to this point of reference. Therefore one would expect a group of experts to cast more homogeneous judgments on the criticality of a scenario than a group with comparatively less subject matter expertise. Furthermore one would expect experts to be affected less by habituation to a specific issue. An ideal subject matter expert will likely have encountered a similar issue before and should therefore not be surprised but instead decisively cast a judgment and stick with it when the issue is evaluated repeatedly. Lastly one may expect that experts on the topic of security assessment will be aware of the consequences of their work and have a tendency to rather overestimate the criticality of a risk, if in doubt.

The following study was set up to test these expectations.

METHOD

The performed study aimed to compare two groups of participants in a setting similar to a naïve test subject study as per the RESPONSE 3 CoP. One group of participants consisted of 33 subjects with low to no experience with driver assistance systems from the environment of the BMW development center. The average age of this group was 33 years with a standard deviation of 12.7 years. 10% of the test subjects were female. All members of this "low expertise" group claimed to not be involved in the development of driver assistance systems as part of their daily work. In the context of the RESPONSE 3 CoP this provides a suitable sample of drivers to compare against the results of the experts. The high expertise group of participants was composed of 19 employees of the BMW Group. These participants were individually recruited based on their direct involvement in earlier controllability assessments. No other measure of expertise was employed. Members of this group were from a variety of departments and fields and had advanced knowledge of the driver assistance systems currently under development. The requirements for an expert panel set out in the RESPONSE 3 CoP have therefore been met.

The study was performed as a vehicle simulator study to minimize the variance caused by outside disturbances. The simulator environment is assumed to influence judgment behavior of both the experts and naïve test subjects identically. The utilized vehicle simulator was the static driving simulator 2 at the BMW Forschungs- und Innovationszentrum (FIZ) in Munich, Germany. The drivers were seated in mockup based on the front half of a BMW 5-series sedan. The 8-channel projection provided a 210° frontal field of vision and separate displays for each side and interior rear view mirror. The mockup vehicle was operated using an automatic gearbox. Figure 2 shows the mockup and the projection surface with the rear view mirror displays uninstalled for improved visibility.

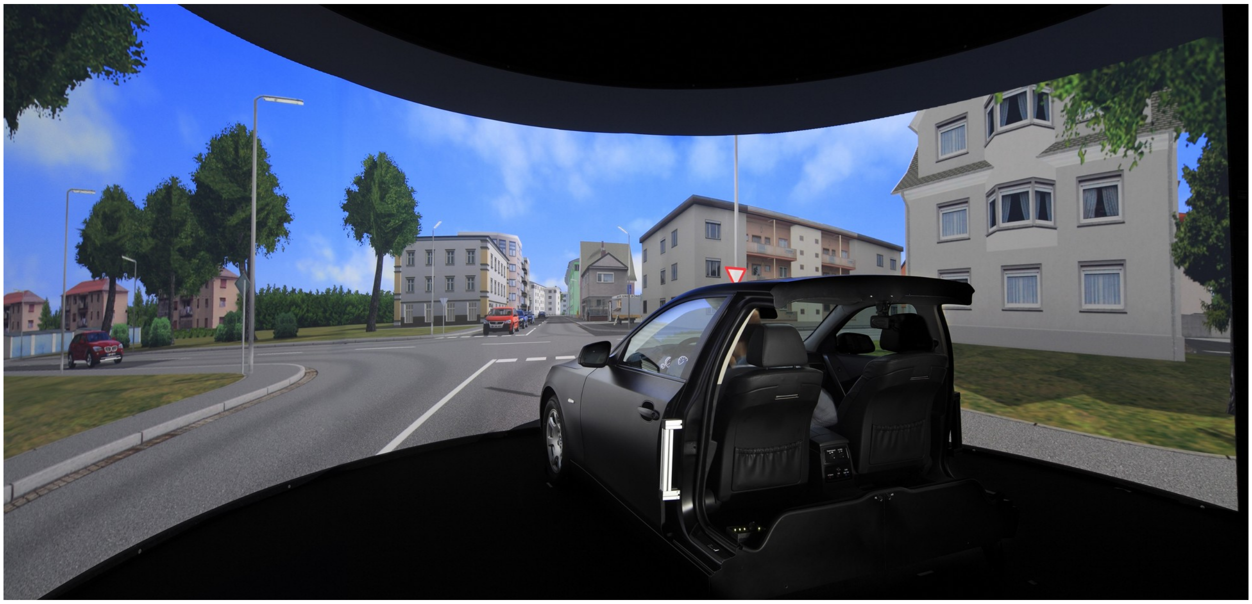

Figure 2 The static vehicle simulator at the BMW FIZ.

All participants, regardless of earlier experience or subject matter expertise, were subjected to the same introduction program. Each participant was introduced to the aim of the study, the used measurement scales and the simulator itself. Following the verbal instruction each participant was instructed to a 15 minute introductory drive to accommodate to the operation of the virtual vehicle. During the operation of the simulator the instructor was connected to the test subject via a two way intercom system.

The main body of the study was a non-permuted mixed design 2-factor-study. The two main influencing factors identified were expertise and repetition. Expertise was assumed to be constant and therefore modified by the recruitment of the low- and high-expertise groups. Repetition was manipulated by repeating the first 4 scenarios of the study for a total of 8 scenarios per participant. The four scenarios were laid out with equal spacing on a 20 minute closed loop road course. This way it was possible to repeat the 4 scenarios without announcing this to the participant.

It was not chosen to permute the order of scenarios to maximize the statistical power obtainable with the given number of participants. Permutation was not required because only the differences between the judgments of the two groups were to be analyzed.

The four presented scenarios consisted of two different scenarios for two different types of driver assistance systems each. System A assisted the driver with automated longitudinal and transversal control of the vehicle but required the hands of the driver to be on the steering wheel at all times. In the critical scenario of system A the transversal control was replaced with a static steering torque to the right hand side during a straight section of the road. No nearby traffic was present during this situation. System B allowed for highly automated driving, giving the driver the opportunity to not observe the traffic in front of the vehicle. In the critical scenario of system B there was an obstacle on the road ahead and a visual-acoustic warning was sounded to ask the driver to take over the vehicle control. This take-over-request had been practiced earlier during the introductory drive. The situation was set up so that oncoming traffic made an evasion of the obstacle impossible. This means the only possible reaction to avoid an accident was the application of the brakes. To limit the variance caused by differences in time spent looking at the road during the operation of system B each participant was asked to operate a secondary task on the mock-up's integrated infotainment system. The employed secondary task was the Surrogate Reference Task (SuRT). This secondary task requires the driver to identify a target circle among a number of distracters and then select whether the target is on the left or right hand side of the screen using the vehicles in-built controller. The size of the distracters was set to approximately 26 arcmin and the target's size was about 32 arcmin. There were 50 distracters. A minor price was offered for the best performance at the SuRT to standardize the motivation for each participant. Figure 3 shows a screenshot of the secondary task in its employed setup.

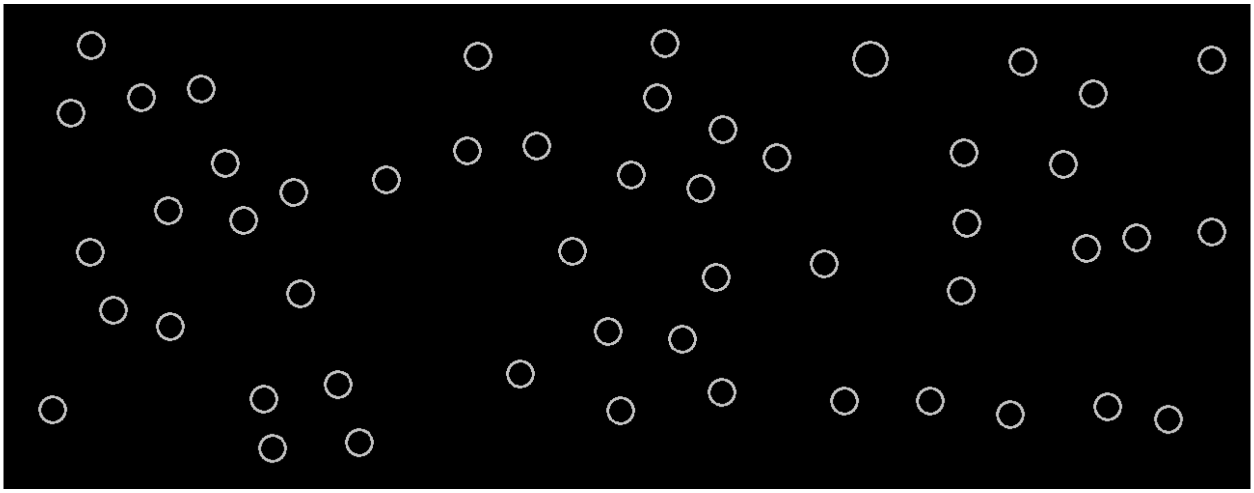

Figure 3 Screenshot from the employed Surrogate Reference Task.

For each system two critical scenarios of different severity were presented. These four situations were experienced twice for a total of 8 situations. After each situation the driver was asked to rate the situation on the discussed scale of disturbance. Following the elicitation the participant was instructed to commence the travel along the road towards the next situation until all eight situations had been rated. Table 1 shows the order in which the scenarios were presented to the participants. It shows that the order of the first 4 scenarios is identical to the second half.

Table 1 Order of presented critical scenarios.

Number	1	2	3	4	5	6	7	8
System	B	A	A	B	B	A	A	B
Severity	High	Low	High	Low	High	Low	High	Low

With the study set up like this it is now possible to formulate hypothesis to test:

1. The variance of the judgments of the high-expertise group is lower than for the low-expertise group in each scenario.
2. The mean of the judgments of the high-expertise group is higher than for the low-expertise group in each scenario.
3. The hedges-g strength of effect of repetition for the high-expertise group is lower than for the high-expertise group.

This adds up to a total of 20 hypotheses. Hypotheses 1 and 2 are to be tested against all 8 scenarios and hypothesis 3 will be tested for each repetition of a scenario, thus 4 times.

MAIN RESULTS

Using the Lilliefors-test for normal distribution it was determined that in 4 of the 8 scenarios presented to the low-expertise group the assumption of normal distribution had to be dropped. For the high-expertise group this was the case in 2 of the 8 scenarios. To make conservative assumptions all results were treated as non-normally distributed. Table 2 reports the results for the test for normal-distribution. Non-significant results are reported as "n.s.".

Table 2 Results of Lilliefors-test for normal distribution.

Scenario	1	2	3	4	5	6	7	8
Low-expertise	$p < 0.05$	n.s.	n.s.	$p < 0.05$	n.s.	$p < 0.05$	n.s.	$p < 0.05$
High-expertise	$p < 0.05$	n.s.	$p < 0.05$	n.s.	n.s.	n.s.	n.s.	n.s.

Afterwards the Levene-test for equality of variances was used to analyze whether the high-expertise group did in fact make judgments with a lower variance than the low-expertise group. The test resulted in only one of 8 scenarios yielding a significant result. The null-hypothesis of equal variance was therefore not discarded. Table 3 shows the results obtained from this test.

Table 3 Results of the Levene-test for equal variances between the low-and high-expertise groups.

Scenario	1	2	3	4	5	6	7	8
Result	$p \sim 0.03$	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

With the assumption of equal variance it was chosen to use the one-sided Mann-Whitney-Wilcoxon-ranksum-test to check the null-hypothesis of the means of the high-expertise groups' judgments being lower or equal to those of the low-expertise group on the scale of disturbance. Table 4 shows the results of this test as well as the Hedge's g strength of effect with the 95% interval of confidence computed numerically using bootstrapping.

Table 4 Comparison of the judgments of the low- and high-expertise group.

Scenario	1	2	3	4	5	6	7	8
MWW-ranksum-test	n.s.	n.s.	$p \sim 0.03$	n.s.	n.s.	n.s.	$p \sim 0.03$	$p \sim 0.03$
Hedges-g upper bound	0.88	0.85	1.17	1.32	1.05	0.93	1.25	1.27
Hedges-g median	0.31	0.26	0.52	0.67	0.41	0.31	0.57	0.61
Hedges-g lower bound	-0.18	-0.31	-0.03	0.11	-0.17	-0.25	0.04	0.08

These results indicate that it can't generally be assumed that a group of 20 participants with high expertise will give judgments with a higher median than a group of lower expertise with 33 participants. The medians of the Hedge-g measure however indicate a consistent trend in the data. In fact in all 8 scenarios the median of the judgments of the high-expertise group was higher than that of the low-expertise group. Assuming normally distributed means the likelihood of receiving a lower mean judgment from the high-expertise group than from the low-expertise group was calculated. Table 5 reports the results of this calculation for each of the 8 scenarios.

Table 5 Likelihood of the high-expertise group giving a lower mean judgment than the low-expertise group.

Scenario	1	2	3	4	5	6	7	8
Result	11%	18%	3%	1%	8%	14%	17%	14%

To analyze the effect of repetition on the two groups for the third hypothesis the Hedge's-g strength of effect parameter and the 95% boundaries are calculated for each of the repeated scenarios.

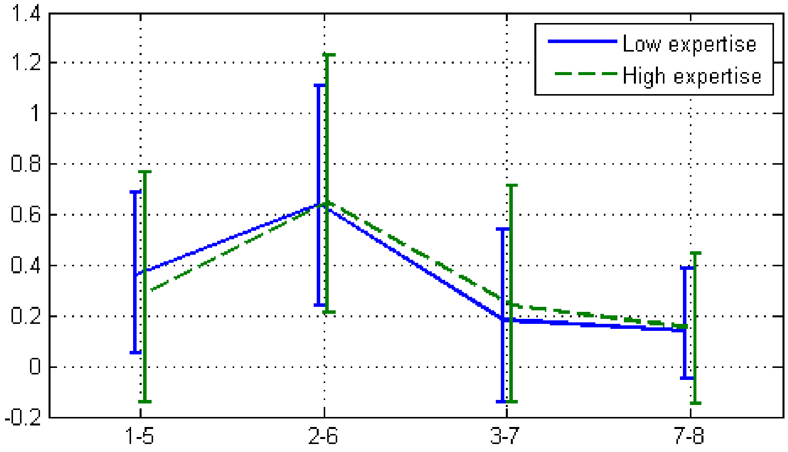

Figure 4 Hedges-g strength of effect of repetition for the four repeated scenarios for the low- and high-expertise group.

The results in Figure 4 indicate that there is no considerable difference between the low- and high-expertise groups regarding the strength of effect of repetition of the scenarios on the judgments on the scale of disturbance.

CONCLUSIONS

The performed study indicated no significant effect of expertise on the variance of judgments on the chosen scale despite the comparatively large number of participants involved. This means that expert reviews of the chosen format can’t be used to decrease the number of participants necessary to achieve a result with the same level of security as with naïve participants. Improved methods of conducting expert reviews must be used if the probability of error is to be equal or smaller than when conducting a naïve test subject study as described in the RESPONSE 3 CoP.

The study identified a trend that indicates that higher expertise may cause higher judgments on the used scale of disturbance. If confirmed this effect could be used to design a test-procedure that can help identify severely uncontrollable situations reliably, thereby eliminating unsuitable system designs early in the development process and increasing the likelihood of obtaining a safe system design.

The results obtained from the study indicate that the low- and high-expertise groups are affected very similarly by repetition. That means higher expertise doesn’t diminish the effect of habituation in a manner that is relevant in practice. This fact should be considered when designing expert reviews of controllability. The order of presented scenarios appears to be relevant for participants with high expertise just as for naïve participants.

In summation these results indicate that the execution of expert reviews isn’t trivial. If no measures are taken to reduce the likelihood of error underestimation of the criticality of a situation is likely even when ample high-expertise participants are available. The presented results show that it is necessary to inspect how expert reviews must be conducted to yield the same level of performance as naïve participant studies as described in the RESPONSE 3 Code of Practice.

REFERENCES

Burgman, M., Fidler, F., McBride, M., Walshe, T., & Wintle, B. (2006). “*Eliciting Expert Judgments: Literature Review*”. University of Melbourne.

Cooke, R., & Goossens, L. H. J. (2000). “*Procedures guide for structured expert judgement in accident consequence modelin*g”. Radiation Protection Dosimetry, 90(3), 303–309.

De Bondt, W. F. M. (1991). “*What do economists know about the stock market?*” Journal of Portfolio Management, 17(2), 84–91.

Englich, B., Mussweiler, T., & Strack, F. (2006). "*Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making*". Personality and Social Psychology Bulletin, 32(2), 188–200.

Ericsson, K. A. (Ed.) (2006). "*The Cambridge Handbook of Expertise and Expert Performance*". Cambridge University Press.

Fischoff, B. (1981). "*Debiasing*".

ISO, 26262-3 (2011, November 14).

Kahneman, D. (2011). "*Thinking fast and slow*". Macmillan.

Neukum, A., Lübbeke T., Krüger H. P., Mayser C., & Steinle J. (2008). "*ACC-Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen*". In M. Maurer & C. Stiller (Eds.), 5. Workshop Fahrerassistenzsysteme (pp. 141–150). Karlsruhe.

RESPONSE 3 (2009). "*Code of Practice for the Design and Evaluation of ADAS*".

Tversky, A., & Kahneman, D. (1974). "*Judgment under Uncertainty: Heuristics and Biases*". Science, 185(4157), 1124–1131.

Weitzel, A., & Winner, H. (2012). "*Ansatz zur Kontrollierbarkeitsbewertung von Fahrerassistenzsystemen vor dem Hintergrund der ISO 26262*". In K. Dietmayer (Ed.), 8. Workshop Fahrerassistenzsysteme. Darmstadt: Uni-Das.