

DEVELOPMENT AND APPLICATION OF AN EXPERT ASSESSMENT METHOD FOR EVALUATING THE USABILITY OF SAE LEVEL 3 ADS HMIS

Naujoks, Frederik
Hergeth, Sebastian
Keinath, Andreas
BMW Group
Germany

Wiedemann, Katharina
Schömig, Nadja
Wuerzburg Institute for Traffic Sciences (WIVW)
Germany

Paper Number 19-0026

ABSTRACT

With the Federal Automated Vehicles Policy, the U.S. National Highway Traffic Safety Administration (NHTSA) has provided an outline that can be used to guide the development and validation of Automated Driving Systems (ADS). Acknowledging that the Human-Machine-Interface (HMI) – identified as one of the 12 priority safety design elements in this voluntary guidance – will be crucial for the success of ADSs, we developed a two-step iterative test procedure that serves to evaluate the conformity of SAE level 3 ADS HMIs with the requirements outlined in NHTSA's Automated Vehicles policy. The aim of this assessment is to evaluate whether minimum HMI requirements are met that facilitate a safe and efficient use of AVs. The present contribution describes the development of an expert-based checklist, how it was compiled from existing literature, how its content and application were refined in simulator and real-world studies, and how it can be employed as a complimentary or stand-alone tool to assess the conformity of SAE Level 3 ADS HMIs with NHTSA's AV policy. It also discusses boundary conditions for the application of the method and the generalization of findings. The described method can be employed in a variety of settings to evaluate SAE Level 3 ADS HMIs, therefore making it a valuable tool for both researchers and practitioners alike.

INTRODUCTION

Conditionally automated driving (SAE L3; [1]) will change how vehicles are used. Depending on the Operational Design Domain (ODD), user of ADS may no longer be required to monitor the driving situation continuously when the system is engaged in automated mode. However, the driver still needs to take back control over the vehicle as soon as a Request to intervene (RtI, also called take-over request) is issued. Therefore, the Human-Machine Interface is of crucial importance to enable a safe and efficient use of the ADS. The ADS has to inform the user through HMI indicators about the current system mode and support the user's awareness about their responsibilities corresponding with the respective mode. Therefore, the NHTSA has proposed that an AV HMI at minimum shall inform the user that the system is (NHTSA, [2]):

- (1) Functioning properly
- (2) Engaged in automated driving mode
- (3) Currently 'unavailable' for use
- (4) Experiencing a malfunction and/or
- (5) Requesting a control transition from ADS to the operator

A suitable design of mode indicators should effectively support the driver in using an ADS and prevent a false understanding of the current driving mode. This is especially important when considering that a given vehicle may be equipped with different driver assistance systems as well that may be confused with ADSs. As this may produce undesired consequences, there is an urgent need to establish test and evaluation methods that can be applied during product development to ensure that these basic HMI requirements are met.

We developed a heuristic evaluation method that can be used by Human Factor and Usability experts to evaluate and document whether an HMI [3] meets the above-mentioned minimum requirements. In Usability Engineering, such heuristic assessment methods are commonly applied during the product development cycle [4] and can be used as a quick and efficient tool to identify and correct potential usability issues associated with the HMI. The heuristic assessment method consists of a set of AV HMI guidelines together with a checklist that can be used as a systematic HMI inspection and a problem reporting sheet. This paper describes the background and application of the checklist.

METHOD DESCRIPTION

Evaluators

The method should be conducted by a pair of HMI experts. Experts should have received formal training in Human Factors and Usability Engineering and have demonstrable practical experience in HMI assessment and evaluation.

Procedure

The HMI inspection is conducted in an on-road assessment of a production vehicle or a high-fidelity prototype. The aim of the assessment is to evaluate whether a set of pre-defined HMI principles (the “heuristics”) are met. Therefore, each of the two evaluators completes a set of fixed use-cases, observes the visual, auditory and haptic HMI output and records potential usability issues arising from non-compliance with the HMI heuristics that have been compiled into a checklist (see [3] for a detailed description of the checklist). The use-case set depends on the specific design of the ADS with respect to the available levels of automation (e.g., whether only manual or conditional automation are available, or if driver assistance is also available within the same vehicle). For an extensive assessment, the use-case set presented in Table 1 should be completed (for a detailed description, see [5]). The aim of the heuristic assessment is twofold:

- (1) For the minimum HMI requirements to be fulfilled, each of the use-cases presented in Table 1 should be reflected in a mode indicator or the change of a mode indicator that must be present in the in-vehicle HMI. The mode indicator can be presented visually, auditory and/or tactile.
- (2) The design of the respective mode indicator should be in accordance with common HMI standards and best practices that are the basis of the checklist (see Table 2; an extended version of the checklist with corresponding examples and background literature can be found in [3]).

Reporting and documentation

Checklist compliance and identified usability issues should be initially documented independently by each of the raters. Each of the checklist items should be answered using the following rating categories:

- “*major concerns*”: non-compliance with guideline
- “*minor concerns*”: partial fulfillment of guideline, but some aspects of the HMI are non-compliant
- “*no concerns*”: compliance of all HMI aspects with guideline
- “*measurement necessary*”: no definite conclusion can be given on the basis of the checklist and empirical testing is needed; this may be the case when very innovative designs are used that are not covered by current standards and best practices.

Reasons for “major” and “minor” concerns should be documented. A problem reporting sheet can be found in [3]. After the individual assessment, the results should be discussed between the evaluators to come to a joint assessment that should also be documented. Figure 1 summarizes the rating procedure.

Table 1: Use-Case set (adapted from Naujoks et al., 2018). Note that some use cases might not be applicable if a vehicle is not equipped with a respective system.

Minimum HMI requirement	Use Case	Description
Functioning properly	L3	Steady driving in L3 mode
Engaged in AD mode	L3 → L2	Driver voluntarily switches from L3 to L2
	L2 → L3	Driver voluntarily switches from L2 to L3
	L2	Steady driving in L2
Currently unavailable for use	L3 _{unavailable}	Driving outside the system’s ODD, L3 is not available; this use case applies to all lower levels of automation (i.e., L0, L1, L2)
Experiencing a malfunction	L3 _{degraded}	Driving in or outside the ODD, L3 is not available because of a malfunction such as a sensor degradation; this applies to all lower levels of automation (i.e., L0, L1, L2)
Requesting a control transition from ADS to operator	L3 → L2	System initiated transition to L2
	L3 → L1	System initiated transition to L1 (either longitudinal or lateral assistance)
	L3 → L0	System initiated transition to L0

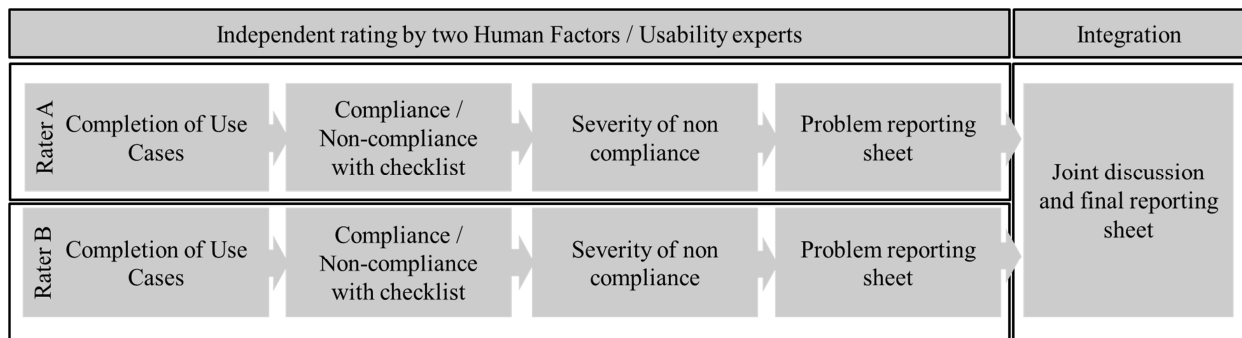


Figure 1: Rating procedure.

Table 2: Checklist items (adapted from [3]).

#	Item
1	Unintentional activation and deactivation should be prevented.
2	The system mode should be displayed continuously.
3	System state changes should be effectively communicated.
4	Visual interfaces used to communicate system states should be mounted to a suitable position and distance. High-priority information should be presented close to the driver's expected line of sight.
5	HMI elements should be grouped together according to their function to support the perception of mode indicators.
6	Time-critical interactions with the system should not afford continuous attention.
7	The visual interface should have a sufficient contrast in luminance and/or colour between foreground and background.
8	Texts (e.g., font types and size of characters) and symbols should be easily readable from the permitted seating position.
9	Commonly accepted or standardized symbols should be used to communicate the automation mode. Use of non-standard symbols should be supplemented by additional text explanations or vocal phrase/s.
10	The semantic of a message should be in accordance with its urgency.
11	Messages should be conveyed using the language of the users (e.g., national language, avoidance of technical language, use of common syntax).
12	Text messages should be as short as possible.
13	Not more than five colours should be consistently used to code system states (excluding white and black).
14	The colours used to communicate system states should be in accordance with common conventions and stereotypes.
15	Design for colour-blindness by redundant coding and avoidance of red/green and blue/yellow combinations.
16	Auditory output should raise the attention of the driver without startling her/him or causing pain.
17	Auditory and vibrotactile output should be adapted to the urgency of the message.
18	High-priority messages should be multimodal.
19	Warning messages should orient the user towards the source of danger.
20	In case of sensor failures, their consequences and required operator steps should be displayed.

METHOD EVALUATION

The method has been evaluated and refined with various approaches. The use of expert assessments may be practical and efficient, but it also comes with limitations. Expert raters might differ in their assessment, resulting in an unreliable outcome of the assessment. Furthermore, the validity of the assessment depends on the capability of the checklist items to predict the usability issues that would arise from non-compliance with them. Therefore, a series of validation experiments were conducted by the authoring team.

Study I: Inter-rater agreement [6]

The aim of the first evaluation study was to assess the *reliability* of the rating outcome in a realistic setting. Demonstrating inter-rater agreement is crucial to the generality of the findings generated from the heuristic assessment, as it is inherently influenced by the raters' subjective experiences and opinions. Therefore, it should be ensured that the ratings were not merely based on idiosyncratic judgements, but that different evaluators

would arrive at similar conclusions when using the method. Three teams of raters (i.e., six individual raters in total) conducted the heuristic assessment in an on-road setting. The employed checklist included two additional items¹. As L3 systems are not yet available to consumers, a L2 system was used to validate the checklist instead. Each of the evaluators drove a section of a German motorway while switching between different automation levels (A70/A71 Schweinfurt/Bamberg; 2 lane-motorway with mainly unrestricted speed limit, including sections with partially missing lane markings and a tunnel). The heuristic evaluation including the final discussion took about six hours per rater pair. All evaluators were employees of the Wuerzburg Institute for Traffic Sciences (WIVW GmbH). They hold a university degree in Psychology or Computer Science and had several years of experience in Human Factors and Usability research.

Table 3: Use cases driven in the on-road evaluation study. $L1_{Long}$ = ACC, $L1_{Lat}$ = Steering Assistance. The use-cases were adapted to the available automation levels in the test vehicle.

Category	Use Case
Activation (driver initiated)	$L0 \rightarrow L1_{long} \rightarrow L2$ $L0 \rightarrow L1_{Lat} \rightarrow L2$ $L0 \rightarrow L2$
Deactivation/ transition to lower level (driver- or system initiated)	$L2 \rightarrow L1_{Long} \rightarrow L0$ $L2 \rightarrow L1_{Lat} \rightarrow L0$ $L2 \rightarrow L0$
Driving steady in a system state	$L0, L1_{long}, L1_{Lat}, L2$
Higher level not available (e.g., sensor failure)	$L0, L1_{long}, L1_{Lat}$
Re-activation of passive system state (system-initiated)	$L0 \rightarrow L1_{Lat}$ $L1_{long} \rightarrow L2$

During and after the test drives, the evaluators recorded their individual assessment before discussing with the other rater. After the team discussion, a final rating was given by every rating team. The main interest of the study was to assess the inter-rater agreement between the individual raters and rater pairs before and after the joint discussion of the rating outcome. Brennan & Prediger's Kappa κ was used to evaluate the reliability of the ratings ([7]; for more details on differences to Cohen's Kappa κ , see [8]).

Table 4: Inter-rater agreement with an evaluation of the quality of the rating according to [9]. Rater pairs were Rater 1/2, Rater 3/4 and Rater 5/6.

	Brennan's κ	R1	R2	R3	R4	R5	R6
Pre	R1	-	0.29	0.36	0.08	0.14	0.13
	R2		-	0.55	0.37	0.48	0.42
	R3		"fair"* = $\kappa > 0.21$	-	0.21	0.37	0.48
	R4		"moderate" = $\kappa > 0.41$		-	0.45	0.12
	R5		"good" = $\kappa > 0.61$			-	0.48
	R6		"very good" = $\kappa > 0.81$				-
Post		R1	R2	R3	R4	R5	R6
	R1	-	0.79	0.48	0.48	0.40	0.40
	R2		-	0.40	0.40	0.48	0.48
	R3		"fair"	-	0.86	0.38	0.50
	R4		"moderate"		-	0.36	0.36
	R5		"good"			-	1
	R6		"very good"				-

As can be seen in Table 4, the inter-rater agreement was not sufficiently high on an individual level before the joint discussion. However, after the discussion among the rater pairs, agreement levels within each rater pair and between different rater pairs increased. This finding demonstrates that different rater pairs come to comparable

¹ The checklist used included two more items in addition to the initial item-set: "Instructions and information of the user manual facilitate the interaction with the HMI" (item #21) and "Interaction with the system is easy" (item#22). Note that these items do not directly pertain to the minimum HMI requirements as proposed by NHTSA.

conclusions using the heuristic evaluation approach, showing that it is a reliable tool to assess the HMI of AVs. However, the findings also highlight that the heuristic evaluation should always adhere to a four-eyes principle to ensure the quality of its outcome.

Study II: Predictive validity [10]

The usefulness of the heuristic HMI assessment not only depends on the reliability of the method, but also on its ability to *predict* usability problems that arise when the heuristics are violated. To test the predictive validity of the heuristics, we constructed two HMIs that are either compliant or non-compliant (“high-compliance” and “low-compliance” HMI) with several checklist items and ran a simulator study with N = 57 participants in the BMW Group’s simulator facilities. A fixed-based driving simulator was used. A detailed description of the study is provided in [10].

The simulated ADS had four modes: (1) manual driving, L3 unavailable for use, (2) manual driving, L3 available for use, (3) L3 engaged, (4) system-initiated take-over request in L3 mode due to system limits. The mode indicators were presented in the instrument cluster. The high compliance HMI (see Figure 2, left) communicated information redundantly by means of pictograms and a textbox. Textual information was displayed in German language. During the approach of the system limits, the HMI announced system limitations through a take-over cascade in form of an announcement, a cautionary take-over request (“*cautionary TOR*”) and an imminent take-over request (“*imminent TOR*”). The request to intervene was shown by animated hands grasping a steering wheel in both HMI variants.

The low-compliance HMI differed from the high-compliance HMI in various aspects (see Figure 2 and Table 5) of non-compliant colour coding, symbol size and labelling. Use-cases included driver initiated activations and deactivations of L3 mode, steady driving in L3 mode and two take-over requests resulting in a transition from L3 to manual driving. The ADS under investigation did not contain L2 or L1 driving assistance. One drive lasted approximately 15 minutes. The study results support the predictive validity of the heuristics in several ways:

- *Perceived usability*: Participants rated the usability of the low-compliance HMI to be statistically significantly lower than the high compliance HMI on the System Usability Scale (SUS, [11]).
- *Observer usability ratings*: Trained observers rated the frequency and severity of usability problems during interactions with the ADS from video footage on a five-point scale ranging from “no problems” to “help from experimenter needed”. Observed usability problems were significantly higher with the low compliance HMI.
- *Take-over time*: Participants reacted significantly slower to RtIs in the low compliance condition compared with the high compliance condition.

Study III: Predictive validity [12]

The predictive validity of the heuristics was further tested in another simulator study at the facilities of the WIVW GmbH. Again, two HMIs were designed that were either compliant or non-compliant with some of the checklist items (e.g., with regard to prominence of task responsibility in L2 assisted driving mode (item #2), color contrast coding (item #7 and item #14), readability of icons and text (item #8), additional explaining text (item #9), usage of understandable language (item #11), multimodality of urgent warnings/take-over requests (item #18) and button labeling consistent to functionality (“additional” item #22)). The HMI variant was varied as a between-subject factor. Twelve drivers completed a simulator drive either with the low- or high-compliant HMI. The participants experienced the HMI in a 30-minutes-driving course containing several use-cases, including driving in each available automation mode (L0 vs. L2 vs. L3), driver initiated-upwards and system-initiated downwards transitions between these levels. The results revealed that the classification of the HMI variants as low vs. highly compliant based on the heuristic evaluation was also reflected in participants’ behavior and subjective ratings of the system and the HMI. The results further support the predictive validity of the heuristics. Differences between the two HMI variants were observed in the following measures:

- *Take-over reaction times*: Participants of the low compliance condition reacted significantly slower to a RtI (hands-on times and take-over times)
- *Usability problems in activating either L2 or L3 system*: participants in the low-compliance condition required more frequent support by the experimenter to successfully activate/reactivate the L3 system
- *Number of handsoff-warnings*: the number of participants experiencing at least one hands-off warning during L2 driving was higher in the low compliance condition
- *Perceived understandability and difficulty in system usage*: Participants in the low-compliance condition reported worse system understanding and perceived it as more difficult to activate the L3 system, to react to a take-over requests in L3 and to react to a system-initiated transition from L3 to L2

- *Global evaluation of the HMI:* Global ratings of the acceptability of the HMI by participants into three categories (very good, acceptable or not acceptable) showed a higher percentage of non-acceptable ratings for the low compliant condition after experiencing the HMI in the driving scenarios.









Mode	High-compliance HMI	Low-compliance HMI
L3 ADS active		
Cautionary TOR		
Imminent TOR		
L3 ADS not available for use		

Figure 2: HMI for high-compliance (left) and low-compliance (right) during normal functioning (top) cautionary TOR (2nd row), imminent TOR (3rd row) and L3 ADS not available (bottom). Figure adapted from [10].

Table 5: Variations for low compliance HMI for the two components with respective criterion and reference to heuristics; adapted from [10].

Variation of low-compliance HMI	Guideline violation
Activation and deactivation through long-press (i.e., 0.8 seconds)	System state changes should be effectively communicated.
Pictograms are 60% of the original size	Texts (e.g., font types and size of characters) and symbols should be easily readable from the permitted seating position.
No text information except for L3 ADS availability	The system mode should be displayed continuously System state changes should be effectively communicated. Commonly accepted or standardized symbols should be used to communicate the automation mode. Use of non-standard symbols should be supplemented by additional text explanations or vocal phrase/s.
No color coding for cautionary and imminent TOR	System state changes should be effectively communicated. The visual interface should have a sufficient contrast in luminance and/or colour between foreground and background. The colours used to communicate system states should be in accordance with common conventions and stereotypes.
No blue color coding for active L3 ADS	System state changes should be effectively communicated. The visual interface should have a sufficient contrast in luminance and/or colour between foreground and background. The colours used to communicate system states should be in accordance with common conventions and stereotypes.

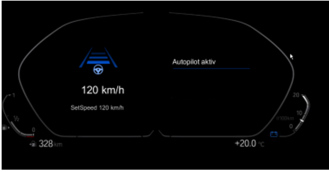







Mode	High-compliance HMI	Low-compliance HMI
L3, ADS active		
L2 assisted driving active		
L2 Handsoff-warning		
Take-over request in L3		

Figure 3: HMI for high compliance (left) and low compliance (right) in selected system modes. Figure adapted from [12].

SUMMARY

This paper presented a heuristic method for the assessment of in-vehicle HMIs for automated vehicles. The aim of the heuristic assessment is to provide a quick but reliable and valid tool that can be used during the product development cycle. It was developed to include common standards and practices and apply them to the in-vehicle interface of AVs [3]. In a series of studies, the reliability and predictive validity of the heuristic assessment was investigated and demonstrated. In view of the minimum HMI requirements proposed in NHTSA's automated vehicle's policy, the method can be used to verify compliance on an analytical level.

It should be noted, however that the method should be applied with care and thought. A thorough application of the method requires (1) the selection and adequate training of HMI evaluators and (2) quality control by periodically checking the agreement between rater pairs as demonstrated in this paper. Otherwise, the outcome of the heuristic assessment might suffer from subjectivity of evaluations and resulting low reliability. It must also be emphasized that the heuristic assessment should be combined with empirical test methods such as simulator or test track studies involving potential users of AVs. The combination of expert evaluations and empirical user tests has a long and successful history in the general Human Factors and Usability context, but has not seen wide-spread application to the domain of AV HMIs in the scientific and technical literature so far.

REFERENCES

- [1] Society of Automotive Engineers International J3016 (2018). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Warrendale, PA: SAE International.
- [2] National Highway Traffic Safety Administration (2016). Automated Driving Systems 2.0: A Vision for Safety. Washington, DC: NHTSA.
- [3] Naujoks, F., Wiedemann, K., Schömig, N., Hergeth, S., & Keinath, A. (2019). Towards guidelines and verification methods for automated vehicle HMIs. Transportation research part F: traffic psychology and behaviour, 60, 121-136.
- [4] Nielsen, J. (1994). Usability inspection methods. In Conference companion on Human factors in computing systems (pp. 413-414). ACM.

- [5] Naujoks, F., Hergeth, S., Wiedemann, K., Schömig, N., & Keinath, A. (2018, September). Use Cases for Assessing, Testing, and Validating the Human Machine Interface of Automated Driving Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 62, No. 1, pp. 1873-1877). Sage CA: Los Angeles, CA: SAGE Publications.
- [6] Wiedemann K., Schömig N., Naujoks F., Hergeth S., Neukum A. & Keinath, A (2018). Expert evaluation of automated driving HMI – does a checklist-based method work? Paper presented at: HFES Europe Chapter Annual Meeting. Berlin, Germany.
- [7] Brennan, R. L. & Prediger, D. J. (1981). Coefficient λ : Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41, pp. 687-699.
- [8] Umesh, U. N., Peterson, R.A., & Sauber M. H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement*, 49, pp. 835-850.
- [9] Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), pp. 159-174.
- [10] Forster Y., Hergeth S., Naujoks F., Krems J.F. & Keinath A. (2019). Empirical Validation of a Checklist for Heuristic Evaluation of Automated Vehicle HMIs. In: *International Conference on Applied Human Factors and Ergonomics*. Cham: Springer.
- [11] Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
- [12] Schömig N., Wiedemann, K., Naujoks, F., Hergeth, S. & Keinath, A. (in preparation). The heuristic evaluation of HMI requirements for Automated Driving Systems – a validation study.