Cognition, Technology & Work (2020) 22:703–720
https://doi.org/10.1007/s10111-019-00599-8

ORIGINAL ARTICLE

Self-report measures for the assessment of human–machine interfaces in automated driving

Yannick Forster[1,2] · Sebastian Hergeth[1] · Frederik Naujoks[1] · Josef F. Krems[2] · Andreas Keinath[1]

Received: 6 May 2019 / Accepted: 9 September 2019 / Published online: 16 September 2019

Abstract

For a successful market introduction of Level 3 Automated Driving Systems (L3 ADS), a careful evaluation of human–machine interfaces (HMIs) is necessary. User preference has often focused on usability, user experience, acceptance and trust. However, a thorough evaluation of measures when applied to ADS HMIs is missing. We investigated the appropriateness of nine self-reported measures in terms of reliability and validity. A sample of $N=57$ participants completed two 15-min simulator drives with a L3 ADS. They experienced two variations of a HMI that differed in the degree of complying with common guidelines. Consistency analysis identified scales that showed insufficient reliability. Validity examination revealed a three-factorial structure of self-reports for construct validity. These factors are design-orientation, usability-orientation and acceptance-orientation. All measures were sensitive to the HMI manipulation and therefore exhibited criterion-related validity. The present study provides researchers and practitioners in the area of ADS with a recommendation for self-report measure application.

Keywords Automated driving · Tools and methods · Human–machine interfaces · Human–computer interaction · Psychometrics

1 Introduction

Level 3 (L3) automated driving systems (ADS) are on the doorstep to the consumer market. These systems are characterized by taking over longitudinal and lateral vehicle control and the driver does not have to constantly monitor correct system functioning (Society of Automotive Engineers International J3016 2018). Instead, he/she has the possibility to engage in non-driving related tasks (NDRT) such as reading a newspaper or watching a movie. The driver has to be ready as fallback performer if the system function fails or the operational design domain of the function ends. Potential benefits of automating the driving task are increased comfort, safety and traffic efficiency (Nunes et al. 2018). However, there might be resistance of people to actually use the ADS (König and Neumayr 2017). To overcome resistance and for the potential benefits to become reality, it is important that people use the available technology and do not decide to disable the functions (van der Laan et al. 1997).

While ensuring the safe use of ADS is of primary importance, human factors researchers and designers need to accomplish two tasks related to usage intention. First, they have to enable HMIs that users evaluate positively to promote the adoption of this technology. Second, this requires development of a methodology on how to evaluate HMIs in the area of automated driving. Society of Automotive Engineers International J3016 (2018) describes the levels of driving automation from L0 (manual driving) up to L5 (full automation). The step from one level to the next higher level of automation is characterized by an incremental transfer of responsibility for certain task components (e.g., steering

✉ Yannick Forster
yannick.forster@bmw.de

Sebastian Hergeth
sebastian.hergeth@bmw.de

Frederik Naujoks
frederik.naujoks@bmw.de

Josef F. Krems
josef.krems@psychologie.tu-chemnitz.de

Andreas Keinath
andreas.keinath@bmw.de

[1] BMW Group, Knorrstr. 147, 80937 Munich, Germany

[2] Chemnitz University of Technology, Wilhelm-Raabe Str. 43, 09120 Chemnitz, Germany

and/or accelerating, monitoring, fallback performance). The SAE J3016R considers systems up to L2 as "driver support features" and starting at L3, systems are considered as "automated driving features". Here, conditions have changed from active driving (SAE L0) to being a passenger (SAE L3 and higher), who is relieved of the former primary task of driving.

In principle, both self-report (e.g., questionnaires) and behavioral measures (e.g., interaction performance) can be used for HMI evaluation (Hornbæk and Law 2007; Nielsen and Levy 1994). In that sense, it is necessary to investigate the suitability of the ample range of available self-report measures. Since the driver becomes the mere fallback performer, circumstances of HMI evaluation have fundamentally changed at this level of automation compared to manual driving or partial automation (Naujoks et al. 2019a). Testing scenarios for automated vehicle HMIs include not only voluntary transitions of control initiated by the driver but also system-initiated transitions from the ADS to the driver (so called Take-Over Requests, TOR). There is dearth of research concerning appropriateness of self-report application for L3 ADS evaluation despite the ample range of self-report measures. Scales that are applied without intensive investigation of its suitability for a particular context can lead researchers and practitioners astray in their decision of the quality of an HMI. Another problem arises when trying to compare results across studies. Without a consensus about methodological application, studies differ substantially in terms of user education, testing scenarios and dependent measures. There have been first efforts into standardization of testing scenarios (Gold et al. 2017; Naujoks et al. 2018). Concerning self-report measures, a recent study by Zoellick et al. (2019) outlined concerns about attitude measures for automated vehicles and brought forth empirical evidence that data structure and validity of attitude measures lack suitability for this context. Similarly, Forster et al. (2018b) have pointed towards the examination of self-report measure suitability. Therefore, the aim of the present methodological work is to thoroughly examine self-report measures for the evaluation of L3 ADS HMIs. The following paragraphs will give an overview of preliminary findings and constructs for the evaluation of HMIs and driving automation. Furthermore, psychometric measures as an evaluation criterion are outlined. From there, research questions are derived resulting in a study that eventually presents a comparative evaluation of different self-report measures.

1.1 Background

According to François et al. (2016), usability and acceptance are important evaluation criteria for HMIs. Additionally, the construct of trust has gained considerable research interest in the evaluation of automated driving lately (Lee and See 2004). Moreover, the construct User Experience (UX) became popular since the 1990s (Norman et al. 1995). The following paragraphs briefly outline those four constructs and relate them to automated driving research.

1.1.1 Usability

When it comes to design for automation, the human-centered design approach gains importance (International Organization for Standardization 2018). According to the ISO 9241, effectiveness, efficiency and satisfaction compose usability. The satisfaction component as a self-report measure refers to the user's attitude towards product use. A frequently applied scale to quantify self-reported usability is the *System Usability Scale* (Brooke 1996). It consists of ten items in total on the two subscales usability and learnability (Lewis and Sauro 2009). It was initially developed to serve as a usability measure that is applicable across a wide range of contexts. The SUS has previously been applied in research on automated driving (Forster et al. 2016, 2017; Hergeth 2016). The *Post Study System Usability Questionnaire* (Lewis 2002) was initially developed for the evaluation of speech dictation systems. Its structure with a total of 19 items can be described with the three subscales system usefulness, information quality and interface quality. Thus, it already bridges the gap to acceptance through its usefulness subscale (see Sect. 1.1.2) and to design-related interface features of user experience (e.g., attractiveness) through its interface quality subscale (see Sect. 1.1.3). The PSSUQ has been used by Walch et al. (2017) to evaluate a L3 ADS HMI. The present study thus examined the applicability of these questionnaires for the evaluation of L3 ADS HMIs.

1.1.2 Acceptance

In the automotive context, researchers have built upon acceptance theory (Davis 1985; Venkatesh et al. 2003) to develop models that predict usage of car technology. The Unified Theory of Acceptance and Use of Technology (UTAUT; Venkatesh et al. 2003) comprises the four subscales Performance Expectancy, Effort Expectancy, Social Influence and Intention to Use with a total of 13 items. The UTAUT combines eight different acceptance models within one generic framework and is a popular tool to evaluate acceptance.

Questionnaires on its basis have been adapted to automotive technology in general (Osswald et al. 2012), L1 driving automation (Adell et al. 2014), L2 driving automation (Rahman et al. 2017), L3 ADS (Rahman et al. 2017), L4 ADS (Nordhoff et al. 2016) and L5 ADS (Nees 2016). According to the acceptance framework by van der Laan et al. (1997), usefulness and satisfaction as two independent dimensions compose acceptance. The van-der-Laan scale consists of nine items on a 7-point semantic differential scale. Since its acceptance definition includes satisfaction as an integral component, it might also be linked with the definition of usability in International Organization for Standardization (2018). This framework has been applied for the evaluation of auditory HMI components for an L3 ADS by Bazilinskyy et al. (2017). Many studies on acceptance focus on acceptability of system functions without providing an experience of the respective technology (Forster et al. 2018a; Kyriakidis et al. 2015; Nees 2016; Payre et al. 2014). The present work fills this gap by examining and comparing acceptance measures for L3 ADS after an experience of the technology in a driving simulator.

1.1.3 User experience

Usability measures cover satisfaction with pragmatic aspects of interaction with a product (i.e., perception of interaction performance). However, they largely neglect non-pragmatic aspects such as interface attractiveness or joy during interaction. The lack of including such qualities into the evaluation of product perception led to the rise of *User Experience* (UX) in the 1990s (Norman et al. 1995). To quantify UX, Hassenzahl et al. (2003) developed and validated the AttrakDiff questionnaire. The 28 item questionnaire covers pragmatic aspects (pragmatic quality) and hedonic aspects (stimulation, identification). It was originally developed in a website and MP3-player context. Stating that the AttrakDiff puts too much emphasis on non-instrumental product aspects, Laugwitz et al. (2008) developed the User Experience Questionnaire (UEQ) over six different contexts such as cell-phones, statistical packages (SYSTAT) or SAP-tools (customer relationship management; CRM). Subsequently the authors report positive results of the UEQ in two validation studies using software products. The 26 semantic differentials describe six subscales (i.e., attractiveness, perspicuity, efficiency, dependability, stimulation, novelty). Minge et al. (2016) developed the modular evaluation of key Components of User Experience (meCUE) as a tool to measure UX. In their self-report measure, 33 items are allocated to 9 subscales (see Table 3) representing hedonic and pragmatic product qualities, emotions towards a product and usage intention. Hence, the meCUE includes aspects of acceptance through its intention subscale (see Sect. 1.1.2) and usability through its UX definition of pragmatic product qualities. There have been applications of the AttrakDiff (Frison et al. 2017), meCUE (Auricht et al. 2014) and UEQ (Häuslschmid et al. 2017) in the driving automation context. However, empirical support for the appropriateness of scale application is still missing.

1.1.4 Trust

Trust is an influential factor on acceptance of technology (Ghazizadeh et al. 2012). Consequently, low levels of trust lead to low acceptance and to rejection of a system (Eichinger 2011; Lee and See 2004). Among others, Jian et al. (2000) and Chien et al. (2014) have developed psychometric scales to measure the attitude trust in automation. The 12-item Automation Trust Scale (ATS; Jian et al. 2000) was explicitly developed for the automation context in computerized systems. From a three-phased experiment (i.e., word elicitation study, questionnaire study, paired comparison study) the authors report the development of a scale to assess human–machine trust. The Universal Trust in Automation scale (UTA; Chien et al. 2014) consists of two components which are “general automation” and “specific automation”. Each dimension includes the three subscales of performance, process and purpose. To evaluate a product or HMI in particular, the “specific automation” component is sufficient. It combines 18 items in total. One important aspect that the authors considered during the development process was inter-cultural differences in trust evolution. In the context of driving automation, these questionnaires or selected items have been frequently used in HMI evaluation (Beggiato et al. 2015; Forster et al. 2017; Gold et al. 2015; Hergeth et al. 2017; Naujoks et al. 2016; Verberne et al. 2012; Waytz et al. 2014). Up to now, it is not clear which scale fits best for evaluating HMIs for L3 ADS. In their work on the development of the Automation Trust Scale (ATS), Jian et al. (2000) recommend to examine the questionnaire in terms of validity and reliability. The current study follows this recommendation and thoroughly examines self-report measures for trust in automation.

The previous outline of constructs and measures has shown that there is a heterogeneity of constructs and measures that can theoretically be applied in HMI evaluation for L3 ADS. The constructs are not completely distinct, but overlap in certain parts. Hassenzahl (2001) describes usability in the sense of pragmatic product quality as one dimension of UX. Satisfaction can be found in both acceptance (van der Laan et al. 1997) and usability (International Organization for Standardization 2018) definitions. There are also links between usability and trust based on theoretical considerations (Hoff and Bashir 2016; Lee and See 2004) as well as on empirical research on ADS (Hergeth 2016). A recent study by Frison et al. (2019) found a link between interface aesthetics and trust in the ADS. Finally, Ghazizadeh et al. (2012) included trust as a precursor for technology acceptance in the framework of Davis (1989). Thus, the issue arises which constructs are necessary and suitable for L3 ADS HMI evaluation.

1.1.5 Psychometrics

The present study aims to examine and compare the psychometric properties of self-report measures for HMI evaluation in the context of L3 ADS. The quality of questionnaires is determined through psychometrics (Bühner 2011; Nunnally 1978). There are main quality criteria and side quality criteria. A high-quality measure adheres to the main quality criteria of objectivity, reliability, validity. The following paragraphs briefly outline these criteria.

Objectivity. Objectivity of a test refers to the degree to which test results are independent from the experimenter. If a test does not vary between experimenters, evaluators and interpreters, it conforms to this criterion. The present work focused on self-report measures that provide standardized instructions for participants when giving their ratings. They also provide instructions for researchers when scoring the questionnaire. Objectivity of conductors, evaluators and interpreters of these methods can thus be assumed and is not in the focus of this study.

Reliability. Reliability refers to the degree of accuracy that a test measures a certain trait with, independent whether the test claims to measure this construct or not. There are several different measures for reliability (Bühner 2011). Sijtsma (2009) describes Cronbach's alpha as the most frequently used measure. Here, each item is considered as an independent test. Accordingly, its accuracy is reflected in the average relationship between all single tests in consideration of the test length. Reliability is a necessary but not sufficient prerequisite for validity.

Validity. There are three types of validity, which are content validity, construct validity, and criterion validity (Bühner 2011). A test has sufficient content validity if its items are representative of the construct. A quantification of content validity is not possible. A closely related concept is face validity. A test has face validity if one can immediately form a connection between an item and a to-be-assessed behavior. Construct validity indicates whether an instrument measures the construct it intends to measure. Convergent (i.e., strong relationships between similar constructs) and divergent validity (i.e., weak relationships between dissimilar constructs) together determine construct validity (Campbell and Fiske 1959; Cronbach and Meehl 1955). A possible numerical method to evaluate construct validity is a factor analytical approach (Bühner 2011). Finally, criterion validity describes the relationship between the test and an external criterion. In HMI research, there is a wide range of guidelines for HMI design (Bubb et al. 2015; Green et al. 1994; Naujoks et al. 2019b). The degree of compliance of an interface to these guidelines can be an external criterion to self-report measures. Thus, the questionnaire should provide statistically significant results for different compliant and non-compliant HMIs.

Side quality criteria. Side quality criteria also add to the overall quality. These criteria are standardization (i.e., availability of norms), comparability (i.e., availability of parallel test forms), economy (i.e., brief and effortless administration) and usefulness (i.e., practical relevance of assessed criterion). These criteria are beyond the scope of the present study but should be considered individually when designing a study and using these questionnaires.

1.2 Research questions and study aim

In a review on usability measures, Hornbæk (2006) recommends the validation of self-report measures. For the constructs usability, acceptance, UX and trust there are several scales that have been applied in HMI research settings (see Sect. 2.7). To date, there exists no general recommendation and no study on the appropriateness of self-report measures for the evaluation of HMIs for driving automation. The scales that are frequently used such as the SUS, AttrakDiff or van-der-Laan certainly bear the advantage of flexibility and adaptability to many different contexts. However, once circumstances of human-technology interaction have changed with the step to L3 automated driving, it is no proper procedure to simply assume method suitability since it had been applied in other automotive contexts such as in-vehicle information systems (IVIS). The current study aims to fill this gap and provide researchers and practitioners with a recommendation for choosing an appropriate self-report measure. Hence, the primary aim is the investigation of the suitability of different self-report measures when evaluating automated vehicle HMIs. A possible criterion to evaluate such self-report measures is psychometrics (Bühner 2011; Nunnally 1978). Therefore, the questionnaires in this study were evaluated in regard to reliability and validity. We do explicitly not claim to conduct a rigorous psychometric evaluation of measures but rather use psychometrics as evaluative criteria to guide the quantification of self-report measure performance. The contribution of this work lies in methodological development for automated vehicle HMI testing. Since high-fidelity driving simulation experiments are time- and cost-consuming, particular circumstances of such setups (e.g., sample size, experimental duration, external validity of safety critical vehicle behavior) apply. Eventually, the goals of this study are (1) to find out whether the self-report measures would exhibit sufficient reliability and (2) meet the validity in terms of content-, construct- and criterion-related validity.

2 Method

2.1 Participants

In total, $N = 57$ (9 female, 48 male) participants took part in the driving simulation experiment. Mean age was 40.56 years (SD= 9.32, max= 60, min= 25). All participants were BMW Group employees, held a German driver's license, had normal or corrected-to-normal vision and had not previously partaken in a driving simulator study on L3 ADS. Thus, we ensured that there was no familiarity of any of the participants with HMIs for automated driving.

2.2 Driving simulation

The study was conducted in a high-fidelity static driving simulator (see Fig. 1). The integrated vehicle's console contained all necessary instrumentation and was identical to a BMW 5 series with automatic transmission. The front channels were displayed through three LED screens (each 1920 × 1080 pixels, 50′ size) providing a combined field of view of 120°. Three LED screens behind the vehicle displayed the rear-view for the mirrors. Driving simulation was rendered with a frequency of 60 Hz.

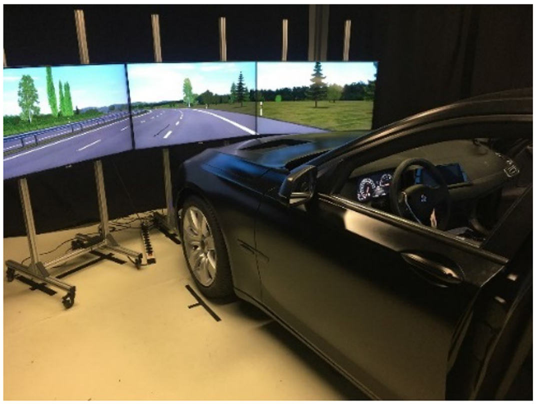

Fig. 1 Static driving simulator with mockup and three front channels used in the current study

2.3 Automated driving function

Once activated, the L3 ADS executed both longitudinal and lateral vehicle control. When the L3 ADS encountered a scenario that exceeded its operational design domain (see Sect. 2.6), a three-stage 20-s TOR was initiated and displayed to the driver (see Sect. 2.5).

2.4 Study design and procedure

There were two different HMIs in the present study. The study employed a one-factor within-subject design with two levels of HMI guideline compliance. Participants were randomly assigned to either the (1) high-compliance HMI or the (2) low-compliance HMI condition in the first drive and experienced the respective other condition in the second drive. The two HMIs and respective differences are outlined in the Sect. 2.5.

Upon arrival, participants were welcomed and gave informed consent. The experimenter explained that the study purpose was to examine two HMIs for automated driving and to evaluate different measures. To accustom themselves with the driving simulation, participants completed a 5-min familiarization drive. Prior to each experimental drive, the experimenter explained that, once activated, the L3 ADS would execute lateral and longitudinal vehicle guidance. Furthermore, the experimenter pointed out, that in case of exceedance of the system's limits, it would inform them with sufficient notice to take over manual control. Participants completed the first drive with all use cases (see Table 2) and subsequently evaluated the HMI in the first inquiry on all nine questionnaires. After this inquiry participants again completed the experimental drive with the respective other HMI. In the second inquiry, they evaluated the HMI again with the same scales as in inquiry 1. Participants received the questionnaires in a randomized order to counteract sequential effects. The experimental procedure is depicted in Fig. 2.

2.5 Human–machine interface

A HMI for automated driving, that had previously been used in studies by Jarosch et al. (2017) and Hergeth et al. (2017) served as the high-compliance HMI. It was

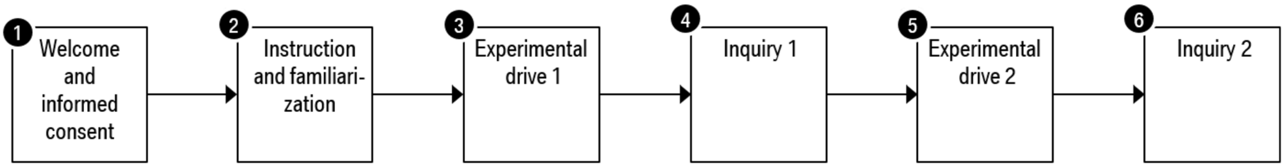

Fig. 2 Flowchart of experimental procedure

depicted in the instrument cluster. When activated, the blue colour of the lane symbols, the text and the steering wheel indicated that the system function carried out longitudinal and lateral vehicle guidance. This HMI (see Fig. 3) resembles that of existing HMI solutions for adaptive cruise control (i.e., longitudinal vehicle guidance; ACC) with additional steering assistance (Naujoks et al. 2015). Information was redundantly communicated by means of pictograms and a textbox above (Stevens et al. 2002). Textual information was displayed in German language. During the approach of the system limits, the HMI announced system limitations through three-staged TOR in form of an announcement, a conditional Take-Over request ('soft TOR') and an immediate take-over request ('hard TOR') (Forster et al. 2016). The stages lasted for 7 s (announcement and soft TOR) and 6 s (hard TOR), respectively. 20 s before reaching the limitation, a generic warning tone announced the upcoming limit. Additionally, the textbox displayed messages. The low-compliance HMI did not provide textual feedback. The soft TOR followed this notification after 6 s and the HMI colour switches from blue to yellow. The HMI shows hands that grab the steering wheel and additional information in the text. After seven more seconds, the hard TOR appeared with the HMI coloured in red and hands grabbing the steering wheel. A more critical warning tone accompanies the visual information. Drivers could activate the L3 ADS by pressing a button on the left side of the steering wheel with the label 'AUTO'. Deactivation was possible through either braking/accelerating, active steering input or pressing the 'AUTO'-button with subsequently putting hands on the steering wheel. During the hard TOR, a hands-on signal immediately deactivated the L3 ADS.

The development of a non-guideline compliant HMI and comparison with a compliant HMI is a mean for the purpose of investigating criterion-related validity. To create a difference between two HMIs, compliance with common HMI guidelines was systematically impaired in the low-compliance condition. A checklist for ADS HMI design by Naujoks et al. (2019b) served as the criterion for HMI compliance. From the high-compliance HMI, both the display component and the operation component (i.e., 'AUTO'-button) were changed by intentionally violating five items of the checklist. Table 1 provides an overview of variations in the HMI, the accordingly varied guideline and reference. The guideline items from Naujoks et al. (2019b) were the following:

- Item 3: System state changes should be effectively communicated.
- Item 7: The visual interface should have a sufficient contrast in luminance and/or colour between foreground and background.

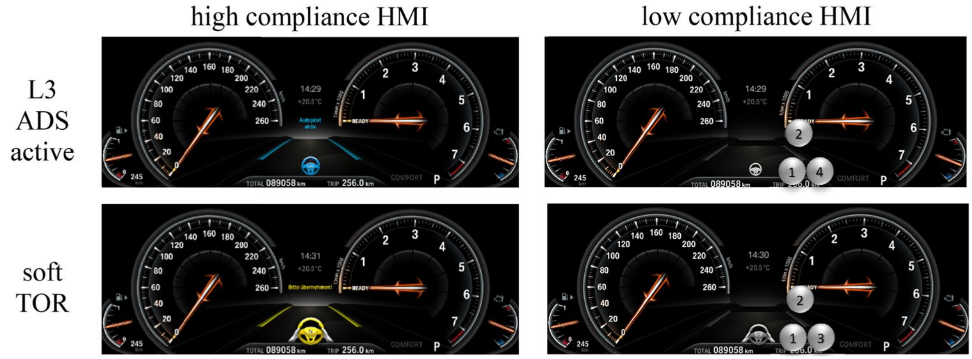

Fig. 3 HMI for high compliance (left) and low compliance (right) during normal functioning (top) and soft TOR (bottom). Numbers indicate HMI variations described in Table 1 column 2

Table 1 Variations for low-compliance HMI for the two components with respective criterion and reference

Component	Variation	Guideline number	Reference in guideline of Naujoks et al. (2019b)
Operation component	Activation and deactivation through long-press (i.e., .8 s)	Item 3	AdaptIVe Consortium (2017)
Display	(1) Pictograms are 60% of the original size	Item 8	Tullis et al. (1995)
	(2) No text information except for L3 ADS availability	Item 2, item 3, item 9	AdaptIVe Consortium (2017), CAMP (2016), Stevens et al. (2002)
	(3) No colour coding for cautionary and imminent TOR	Item 3, item 7, item 14	AdaptIVe Consortium (2017), ISO 15008 (2012), Stevens et al. (2002)
	(4) No blue colour coding for active L3 ADS	Item 3, item 7, item 14	AdaptIVe Consortium (2017), ISO 15008 (2012), Stevens et al. (2002)

- Item 8: Texts (e.g., font types and size of characters) and symbols should be easily readable from the permitted seating position.
- Item 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 phrases.
- Item 14: The colours used to communicate system states should be in accordance with common conventions and stereotypes.

Figure 3 depicts the high-compliance HMI (left) and the low-compliance HMI (right) when the L3 ADS is activated (top) and the soft TOR (bottom).

2.6 Use cases

Use-cases of the present study were chosen based on the HMI testing scenario catalogue for L3 ADS proposed by Naujoks et al. (2018). Use-cases included driver-initiated activations and deactivations, two TORs due to road works and the end of L3 ADS availability as well as independently executed system maneuvers (Naujoks et al. 2017). The eight use-cases were arranged in a fixed order. This is necessary in studies on automated driving since for example a TOR always requires a user-initiated transition to an automated mode as the subsequent UC. The TOR scenario itself lasted 30 s in total if the driver does not intervene. The driver initiated UCs took until the UC was completed successfully. At a maximum, the experimenter waited for 2 min. If the participant could not complete the respective transition, he/she was instructed by the experimenter. An analysis of the duration of the activation scenarios is reported in Forster et al. (2019). One drive lasted approximately 15 min. Table 2 shows the eight UCs with information about the initiator (i.e., driver vs. system).

2.7 Dependent variables

Table 3 summarizes the dependent measures used in the present study. Section 1.1 Background already provided information about the development of the respective scales. Questionnaires that did not exist in German language (i.e., UTAUT, PSSUQ) were translated and back-translated by a German and an English native speaker (Jones et al. 2001). To investigate face validity, participants were asked to additionally indicate when they struggled in answering a specific item due to unclear formulation of the item or inappropriateness in the automated driving context.

Table 2 Sequential order of use-cases for each experimental drive

Number	Use-case	Initiator
1	Initial activation	Driver
2	Lane change	System
3	Deactivation	Driver
4	Re-activation	Driver
5	TOR (road works)	System
6	Re-activation	Driver
7	Speed adaptation	System
8	TOR (end of L3 ADS)	Driver

2.8 Manipulation check

An expert evaluation served as manipulation check to ensure a successful variation of HMI compliance. It consisted of eight items on a 7-point Likert scale from 1 (“not at all”) to 7 (“very much”) concerning guidelines for HMI design (Naujoks et al. 2019b). Participants answered each item for both the high- and low-compliance HMI. The manipulation check for HMI guideline criteria was averaged into a composite. The items number with the wording are shown in Table 4.

2.9 Statistical procedure and data analysis

To ensure that no confounding factor (i.e., HMI guideline compliance) is present which could lead to an interaction between dependent measures and stages of the independent variable, reliability and validity are examined separately for both HMI conditions. Subscales were averaged into a composite as described in the original source. Mean ATS scores were calculated separately for trust and distrust as suggested by Spain et al. (2008). Hence, there are two separate mean scores (i.e., trust, distrust) with a value between 1 and 7 each.

Cronbach’s alpha was calculated as a measure of scale reliability (Cronbach 1951). To evaluate reliability coefficients in an absolute sense, the present values were compared to a minimum value of $\alpha = .7$ (Kline 1999; Nunnally 1978) as well as the coefficients in the original source (if reported).

We evaluated content validity by means of participants’ evaluation about whether they struggled giving their rating on the respective item (i.e., face validity). Here, we counted the total number of indications for each item.

To determine construct validity, the correlations of the entire set of subscales in the present study were evaluated by means of an exploratory factor analysis (EFA) approach (Bühner 2011). This approach allows combining subscales to a certain number of factors that assess a similar construct and distinguish them from other subscales that assess a different facet of user preference.

Table 3 Constructs, questionnaires, scales and original sources for self-report measures

Construct	Questionnaire (subscales)	Subscales	Scale	Source
Usability	SUS (2)	Usability Learnability	Likert [1–5]	Brooke (1996)
	PSSUQ (3)	System usefulness Information quality Interface quality	Likert [1–7]	Lewis (2002)
Acceptance	UTAUT (4)	Performance expectancy Effort expectancy Social influence Intention to use	Likert [1–7]	Rahman et al. (2017) adapted from Venkatesh et al. (2003) and Adell (2010)
	van-der-Laan Scale (2)	Usefulness Satisfaction	Semantic differential [1–5]	van der Laan et al. (1997)
User Experience	AttrakDiff (3)	Stimulation Identification Pragmatic quality	Semantic differential [1–7]	Hassenzahl et al. (2003)
	UEQ (6)	Attractiveness Perspicuity Efficiency Dependability Stimulation Novelty	Semantic differential [1–7]	Laugwitz et al. (2008)
	meCUE (9)	Usefulness Usability Status Aesthetics Commitment Positive affect Negative affect Intention Loyalty	Likert [1–7]	Minge et al. (2016)
Trust	ATS (2)	Distrust Trust	Likert [1–7]	Jian et al. (2000)
	UTA (3)	Performance Process Purpose	Likert [1–5]	Chien et al. (2014)

Table 4 Manipulation check item numbers, wording and respective guideline number in Naujoks et al. (2019b)

Item number	Item wording	Guideline number in Naujoks et al. (2019b)
1	The driver is supported by the HMI in his/her perception of system state changes	3
2	There is a sufficient contrast between foreground and background	7
3	The visual display and the background differ sufficiently by colour	7
4	Colour coding is according to urgency	14
5	The displayed symbols are easily readable from the permitted seating position	8
6	The displayed text is easily readable from the permitted seating position	8
7	The operating elements of the HMI is intuitive	3, 9
8	There is immediate feedback about user input on the HMI	3, 9

Criterion-related validity was evaluated by means of inferential statistical analysis of the within-subject factor HMI compliance. A repeated measures-ANOVA was calculated for all self-report measures. Thus, criterion-related validity represents sensitivity of the questionnaires to the experimental HMI variation. If a questionnaire is valid in this sense, it must be sensitive to the manipulation and reveal a statistically significant main effect of HMI condition.

3 Results

3.1 Manipulation check

$N = 8$ experts in the field of Human Factors (at a minimum Master's degree in psychology, human–computer interaction or related field) completed the manipulation check. The experts completed the manipulation check questionnaires (see Sect. 2.7) after experiencing both the high- and low-compliance HMI. Descriptive data for the fulfillment of HMI guidelines showed that the high-compliance HMI was considered superior ($M = 5.30$, SD= .57) compared to the low-compliance HMI ($M = 2.31$, SD= .42).

3.2 Missing data

Across all participants, only $n = 4$ missed to answer single items. This led to a total of $N = 13$ missing items. One participant did not complete the van-der-Laan scale. With every participant answering 334 items in total, the percentage of missing data is very low and equals .1%. Reliability analysis used list-wise deletion. For validity, however, an exclusion of $n = 4$ participants would be necessary. According to Tabachnick and Fidell (2007) 'Expectation–Maximization-methods sometimes offer the simplest and most reasonable approach to imputation of missing data, as long as your preliminary analysis provides evidence that scores are missing randomly' (p. 71). Since the loss of information looms larger than the overestimation of effects through the expectation–maximization (EM) approach, missing raw values as well as the van-der-Laan scale scores were estimated by an EM Algorithm (Lüdtke et al. 2007).

3.3 Reliability

The following section outlines reliability results for the subscales of self-report measures. Table 5 summarizes coefficients of the present study for both the high and the low compliance HMI overall. In addition, the reliability coefficient Cronbach's alpha of the original source is reported. Concerning the ATS, Jian et al. (2000) did not provide Cronbach's alpha for a two-factorial solution, so no comparison is possible.

Reliability analysis found that $n = 7$ subscales did not meet the minimum value of $\alpha = .7$ in at least one of the two experimental conditions. If a low internal consistency was observed in the present data, the original source also reported comparably low Cronbach's alpha values such SUS Learnability, UATUT Social Influence and UEQ Dependability. Conversely, there are also subscales (i.e., UTA Process, Purpose; meCUE Intention, Usefulness) that exhibited low internal consistency while the original source reports sufficient reliability.

Table 5 Reliability coefficients (Cronbach's alpha) for each subscale by HMI (i.e., high compliance, low compliance) and the original source (if reported)

Scale	Subscale	High-compliance HMI	Low-compliance HMI	Original
SUS	Learnability	**.571**	**.544**	.70
	Usability	.824	.863	.90
PSSUQ	System usefulness	.940	.940	.96
	Information quality	.898	.903	.96
	Interface quality	.783	.880	.92
UTAUT	Perf. expectancy	.774	.832	.87
	Effort expectancy	.907	.902	.86
	Social influence	**.194**	**−.133**	.48
	Intention to use	.900	.879	.91
VDL	Usefulness	.795	.811	.73–.87
	Satisfaction	.887	.879	.81–.90
AttrakDiff	Hed-Stim	.829	.868	.76–.90
	Hed-Ident	.872	.927	.73–.83
	Pragmatic	.832	.872	.83–.85
meCUE	Usefulness	**.588**	.718	.83
	Usability	.890	.898	.89
	Status	.789	.793	.83
	Aesthetics	.799	.857	.89
	Commitment	.806	.768	.86
	Positive affect	.892	.886	.94
	Negative affect	.851	.814	.92
	Intention	.716	**.571**	.86
	Loyalty	.749	.838	.76
UEQ	Attractiveness	.896	.954	.89
	Perspicuity	.786	.863	.82
	Efficiency	.796	.729	.73
	Dependability	**.669**	.866	.65
	Stimulation	.791	.799	.76
	Novelty	.906	.922	.83
ATS	Distrust	.732	.826	N/A
	Trust	.940	.946	N/A
UTA	Performance	.820	.828	.889
	Process	**.560**	.751	.870
	Purpose	**.691**	.781	.864

Cells failing to exceed a Cronbach's alpha value of .7 are in bold

Table 6 Items and respective scales with low face validity as indicated by the frequency of participants labelling an item as 'problematic to answer'

Scale	Item	Face validity concerns [n]
PSSUQ	The system gave error messages that clearly told me how to fix problem	12
	Whenever I made a mistake using the system, I could recover easily and quickly	17
UTAUT	People who influence my behavior would think that I should use the system	19
	People who are important to me would not think that I should use the system	12
AttrakDiff	Isolating-connective	15
	Alienating-integrating	17
	Brings me closer to people-separates me from people	26
	Cautious-bold	11
	Harmless-challenging	11
meCUE	The product would enhance my standing among peers	11
	Using the product, I would be perceived differently	12
	Compared to other products, this product seems incomplete	12
ATS	The system has integrity	17
UTA	The system uses appropriate methods to reach decisions	20
	I can always rely on the system to ensure my performance	14

3.4 Validity

3.4.1 Content validity

We approached content validity via participants' face validity ratings. Table 6 shows items with a minimum of $n = 11$ participants that considered an item as problematic to answer which equals close to every fifth participant (19.3%). This threshold was not chosen out of convenience but because from an expert's perspective, 20% or more of a sample indicating issues in answering a respective item is problematic. Analogous to the procedure to determine the number of factors for the EFA in Sect. 3.4.2, a Scree-plot of the number of indications across all 168 items led to the threshold of $n = 11$ that are worth mentioning here. When a smaller percentage (e.g., 3 out of 57 participants) indicated that they struggled in understanding, no consistent picture across the entire sample could be drawn. If a participant marked the same item at both times of measurement, he/she was counted as one. Results revealed $n = 15$ items with low face validity. PSSUQ ratings depend on whether the experiment included use cases with error messages and participants who had made a mistake and needed to recover from these. Both items of the UTAUT Social Influence subscale were considered problematic. The AttrakDiff included the highest number of items ($n = 5$) with low face validity. A large number of participants also considered meCUE items related to status as problematic. This is in accordance with the UTAUT Social Influence result. Another reason for face validity concerns were unclear content of expressions as observed for 'integrity' (ATS) and 'performance' (UTA).

3.4.2 Construct validity

Construct validity was investigated for the high compliance HMI condition by means of a factor analysis. The Kaiser–Meyer–Olkin (KMO) as a test for the appropriateness of the entire correlation table (relationship between all subscales) for factor analysis revealed a score of .878. This that indicates appropriateness of data for a subsequent factor analysis. Bartlett's test for sphericity became highly significant [$X^2(496) = 1955.340$, $p < .001$]. To determine the factor structure of the preference ratings, an EFA with principal-component factor extraction and Varimax orthogonal rotation was carried out. The Scree-criterion and Velicer's minimum average partial test (O'connor 2000; Velicer 1976) suggested a three-factor solution. The factors can explain 28.33%, 25.70% and 16.30% of total variance, respectively, adding up to a total variance explained of 70.33%. Table 7 shows factor loadings for a three-factor Varimax orthogonal rotated solution sorted by size. Loadings smaller than .5 are coloured in grey. Self-report measures that use semantic differentials such as all subscales of the AttrakDiff, four subscales of the UEQ and the van-der-Laan scale show high loadings on factor 1. This factor combines measures that evaluate the graphical interface design. Self-report measures that focus on interaction and pragmatic qualities of the interface such as the SUS, UTAUT Effort Expectancy and PSSUQ System Usefulness exhibit large factor loadings on factor 2. Finally, subscales that assess future intentions regarding the use of the system function accumulate on factor 3. Both ATS subscales load only weakly on factor two. This indicates that trust as measured by ATS rather forms a

Table 7 Matrix with factor loadings after Varimax rotation

Subscale	Factor 1	Factor 2	Factor 3
UEQ novelty	**.883**	.083	.090
UEQ stimulation	**.841**	.226	.315
meCUE aesthetics	**.811**	.176	.175
AttrakDiff pragmatic quality	**.793**	.339	.223
UEQ attractiveness	**.773**	.378	.330
AttrakDiff stimulation	**.771**	.454	.236
AttrakDiff identification	**.747**	**.510**	.119
PSSUQ interface quality	**.721**	.360	.076
UEQ efficiency	**.671**	.416	.270
VDL usefulness	**.644**	.396	.416
meCUE usefulness	**.589**	.304	.444
PSSUQ information quality	**.581**	**.553**	.109
UTAUT effort expectancy	.226	**.889**	.112
UEQ perspicuity	.269	**.849**	.053
meCUE usability	.283	**.841**	.082
SUS usability	.339	**.827**	.211
SUS learnability	.101	**.804**	.004
PSSUQ system usefulness	.291	**.803**	.207
UTA process	.411	**.714**	.159
meCUE negative affect	.378	**.692**	.170
VDL satisfaction	**.591**	**.606**	.322
UEQ dependability	**.524**	**.605**	.112
ATS distrust	.461	.497	.306
ATS trust	.038	.267	.116
meCUE positive affect	**.514**	.102	**.700**
UTA performance	.500	.138	**.676**
meCUE intention	**.510**	.212	**.664**
meCUE commitment	−.030	.029	**.655**
meCUE status	**.506**	−.092	**.610**
UTAUT intention	.498	.325	**.598**
UTAUT performance expectancy	.463	.096	**.594**
UTA purpose	.069	.305	**.547**
meCUE loyalty	.402	.377	.491
UTAUT social influence	−.098	.373	.378

Factor loadings are sorted by size and large values (i.e., > .5) are in bold

fourth component of user preference or is not suitable for interface evaluation.

According to the Fornell–Larker criterion (Fornell and Larcker 1981), factor loadings of a component on its factor needs to be at least .7 for sufficient convergent validity. To meet requirements of divergent validity, the item's factor loading on other factors must not exceed .3. The present results show that eight subscales on factor 1 show high convergent validity. This factor combines subscales that primarily assess an interfaces' design features and graphical appearance. Seven subscales show high convergent validity on factor 2. Scales that assess a user's interaction and ease of it load high on this factor. These subscales are all closely tied to the usability construct. One subscale of factor 3 showed sufficient convergent validity (i.e., UTAUT Intention). Subscales that assess usage intention accumulate on this factor. Scales that did not exhibit convergent validity (e.g., meCUE Usefulness) neither met the divergent validity goal. These subscales are represented by a combination of two or more factors rather than by one factor alone. Results of the EFA procedure for ratings of the low compliance HMI conditions also revealed a three-factor solution for user preference with similar factor loadings. Due to pagination constraints, these results are not additionally reported here.

3.4.3 Criterion validity

To determine each scale's criterion validity regarding guidelines for HMI design, 2-factorial repeated measures ANOVAs were calculated for all nine questionnaires. The within-subject factors were HMI compliance (high vs. low) and number of subscales. Table 8 shows descriptive (i.e., *M*, SD) and inferential results for the main effect of the HMI and the interaction between HMI and order of presentation (i.e., Wilk's λ). Statistically significant results are coloured in grey. Results of inferential statistics revealed that all scales could discriminate between the high and the low compliance HMI (significant main effects). Thus, the external criterion of HMI compliance is reflected in all self-report measures. Significant interaction effects (i.e., PSSUQ, UTAUT, meCUE Module 1, UEQ, UTA) indicate that the difference between the high and the low compliance HMI is not equally present at all subscales of the respective questionnaire. These questionnaires contain subscales that are highly sensitive to the experimental variation and subscales that are not as sensitive to the HMI criterion.

4 Discussion

The current study examined self-report measures that are frequently applied to evaluate L3 ADS HMIs regarding psychometrics. $N = 57$ participants completed nine questionnaires for the constructs usability, acceptance, user experience and trust once for a high compliance and once for a low compliance L3 ADS HMI. Cronbach's alpha served as an estimate of scale reliability. We applied an EFA approach for the investigation of construct validity. We furthermore followed an inferential analysis of the high compliance and low compliance HMI for the examination criterion validity. This section discusses the outcomes and methodological aspects for each requirement.

Table 8 Descriptive (i.e., M, SD) and inferential (i.e., main effect for HMI and interaction HMI × subscale) statistics for each scale

Scale	Subscale	High compliance *M* (SD)	Low compliance *M* (SD)	Main effect HMI	Interaction HMI × subscale
SUS	N/A	82.45 (14.01)	67.11 (19.14)	$t(56)=5.959, p<.001, d=.778$	N/A
PSSUQ	SysUse	6.20 (.89)	5.36 (1.13)	$F(1,56)=63.439, p<.001, \eta^2=.531$	$F(2,55)=8.463, p<.05, \eta^2=.235$
	InfoQual	5.48 (1.14)	4.08 (1.28)		
	IntQual	5.64 (1.09)	4.29 (1.57)		
UTAUT	Perf. expectancy	5.28 (.94)	4.79 (1.17)	$F(1,56)=38.583, p<.001, \eta^2=.408$	$F(3,54)=4.561, p<.05, \eta^2=.202$
	Effort expectancy	6.10 (.88)	5.23 (1.20)		
	Social influence	4.71 (1.03)	4.41 (.93)		
	Intention to use	6.22 (.87)	5.73 (1.11)		
van-der-Laan	Usefulness	4.24 (.52)	3.73 (.66)	$F(1,56)=40.736, p<.001, \eta^2=.421$	$F(1,56)=3.726, p=.059, \eta^2=.062$
	Satisfaction	4.24 (.59)	3.57 (.80)		
AttrakDiff	Hed-Stim	5.32 (.85)	4.47 (1.10)	$F(1,56)=38.951, p<.001, \eta^2=.410$	$F(2,55)=3.103, p=.053, \eta^2=.101$
	Hed-Ident	5.07 (.81)	4.28 (.98)		
	Prag.	5.13 (.86)	4.89 (1.08)		
meCUE	Usefulness	5.64 (.83)	4.95 (1.02)	$F(1,65)=41.276, p<.001, \eta^2=.424$	$F(3,54)=8.652, p<.001, \eta^2=.395$
	Usability	6.06 (.95)	5.05 (1.33)		
	Status	4.16 (1.26)	3.84 (1.35)		
	Aesthetics	4.74 (1.12)	3.61 (1.51)		
	Commitment	2.52 (1.11)	2.26 (1.15)		
	Pos. affect	4.65 (1.03)	4.07 (1.12)	$F(1,56)=22.793, p<.001, \eta^2=.289$	$F(1,56)=.035, p=.852, \eta^2=.001$
	Neg. affect	5.37 (1.00)	4.82 (1.06)		
	Intention	5.40 (1.04)	4.67 (1.07)	$F(1,56)=40.383, p<.001, \eta^2=.419$	$F(1,56)=3.678, p=.060, \eta^2=.062$
	Loyalty	4.44 (1.13)	3.49 (1.31)		
UEQ	Attractiveness	5.72 (.86)	4.73 (1.34)	$F(1,56)=48.355, p<.001, \eta^2=.463$	$F(5,52)=3.402, p<.05, \eta^2=.274$
	Perspicuity	6.02 (.89)	5.00 (1.25)		
	Efficiency	5.77 (.86)	5.23 (.97)		
	Dependability	5.79 (.79)	4.85 (1.23)		
	Stimulation	5.25 (.92)	4.50 (1.25)		
	Novelty	4.88 (1.37)	3.99 (1.61)		
ATS	Distrust	5.59 (.67)	4.70 (1.18)	$F(1,56)=34.551, p<.001, \eta^2=.382$	$F(1,56)=3.550, p=.065, \eta^2=.036$
	Trust	5.63 (1.07)	5.03 (1.23)		
UTA	Performance	3.78 (.76)	3.46 (.80)	$F(1,56)=58.848, p<.001, \eta^2=.512$	$F(2,55)=12.442, p<.001, \eta^2=.311$
	Process	4.11 (.56)	3.31 (.80)		
	Purpose	3.88 (.59)	3.41 (.76)		

4.1 Reliability

Measures of reliability were mostly sufficient in an absolute sense (Kline 1999) and comparable to the values reported in the original source (see Table 4). The PSSUQ, van-der-Laan scale, AttrakDiff, UEQ and ATS showed positive results (i.e., high α values) of the reliability analysis. Reliability of the two-factorial structure of the SUS (Bangor et al. 2009) turned out to be sufficient for the usability subscale but insufficient for the learnability subscale. Reliability results for the SUS suggest following the SUS score calculation instructions in Brooke (1996) and rather rely on a one-factorial solution. Subscales that revealed unreliable results were the UTAUT Social Influence subscale and UTA subscales Process and Purpose. These results discourage future administration of these scales in the ADS context because the UTAUT Social Influence subscale was highly unreliable and the UTA showed two out of three subscales that could not reach the criterion of .7. Furthermore, there were also instances in the meCUE (usefulness, intention) and UEQ (Dependability) with insufficient Cronbach's alpha values. These are also considered as problematic to use. However, the UEQ subscale was just close to the threshold value. Regarding the meCUE, this does not mean that the entire scale might not be used since the modules can be applied separately. Still, one might consider the van-der-Laan scale for usefulness and UTAUT for intention as superior when it comes to reliability.

4.2 Validity

4.2.1 Content validity

Content validity as indicated by face validity was high for the SUS, van-der-Laan scale and UEQ (see Table 6). Furthermore, participants considered only one item each of the PSSUQ and ATS as problematic to answer. Thus, we consider content validity for these two measures as given. Face validity investigation revealed that people struggled with scales and items that relate to the opinion to other people such as the UTAUT Social Influence and the meCUE Status subscale. A possible explanation for this finding might be that L3 ADS are not yet commercially available. Thus, peer-related questions require a lot of imagination and do not lead to valid results. Items that cover the HMIs suitability for communication such as three AttrakDiff items were considered as problematic to answer, as communication with other people was not a design purpose of the present HMIs for automated driving. The present results suggest omitting questions on other peoples' opinions as long as there is no commercial availability or distribution on the consumer market. There are two different reasons for low validity of UTA items. First, complexity of an L3 ADS is high and people can hardly judge, how the system makes decisions and comes to conclusions. Furthermore, the term performance within the L3 ADS context remains unclear. Generally, the interaction success with a certain technology is considered as performance (e.g., driving a vehicle), while in the L3 ADS context, the performance per se (i.e., driving) is executed by the system function and the driver's performance is rather reflected in NDRT engagement or reaction to a TOR. Therefore, the performance term remains obscure for many participants and should be applied with caution in this context.

4.2.2 Construct validity

Construct validity examination led to a three-factor solution for self-report measures (see Table 7). The first factor includes mostly graphical design-related measures. The second factor is composed of the usability and instrumental scales and is therefore interaction-oriented. The third factor combines scales that assess usage intention and therefore we consider the factor acceptance-oriented. Two separate factors for instrumental and non-instrumental qualities as suggested by Hassenzahl et al. (2003) were apparent in the present solution. Acceptance-related scales are separated from instrumental and non-instrumental qualities. Minge et al. (2016) have suggested this additional dimension but point towards the fact that there are correlations between measures of acceptance and usability. Support for this assumption comes from the results of discriminant validity. Subscales from the intention factor also revealed remarkable loadings on both the design and usability factor. The present analysis found that the SUS and meCUE showed best results of construct validity due to high discriminant and convergent validity on the respective factors. SUS subscales loaded on the interaction-oriented factor and meCUE subscales loaded across all factors in the way that was expected according to Minge et al. (2016). It is unclear whether the construct of trust relates to these measures. The ATS subscales did not align with the present solution while the UTA subscales aligned with the interaction and intention factor. Considering that reliability and face validity results were more positive for the ATS, we argue against the application of trust measures when investigating HMI preference. The PSSUQ as a proposed usability measure was located not only on the expected interaction factor but also on the unexpected design factor. Therefore, validity is on a medium level. UTAUT's intention and effort subscales loaded as expected. The performance subscale was expected to also align with the interaction factor but eventually showed more alignment with the intention factor leading to a medium validity. Validity on the van-der-Laan scale was low due to the observation of loadings on multiple factors (e.g., usefulness on design and interaction). Factor loadings of the AttrakDiff revealed that they all assess design-oriented criteria in this context. This is not in accordance to the originally proposed structure of pragmatic and hedonic product qualities (Hassenzahl et al. 2003). Hence, low validity was assigned. With efficiency as a clearly interaction-oriented factor loading on design, validity of the UEQ was impaired. The other subscales accumulated on the factor that could be expected from their original proposition (Laugwitz et al. 2008). Table 9 summarizes the results of construct validity examination.

4.2.3 Criterion-related validity

Examining criterion validity, main effects of the inferential tests showed that all scales could discriminate between the two experimental conditions (see Table 8). Even though there are scales with reliability and validity concerns they can detect a difference if HMI design guidelines are violated (Naujoks et al. 2019b). Significant interaction effects indicate that differences between the high compliance HMI and low compliance HMI are not reflected in the same way across all subscales. These questionnaires incorporate subscales with a varying degree of sensitivity to the experimental HMI compliance manipulation. The SUS due to the single percent measure for sensitivity, van-der-Laan, AttrakDiff and ATS showed continuously strong differences between the two HMI variations. When evaluating L3 ADS HMIs with any of the other scales, one has to be aware of differences in sensitivity of the subscales within the questionnaire. The evidence from the present study indicates that all the

Table 9 Result overview of construct validity analysis

Construct	Questionnaire	Result	Interpretation
Usability	SUS	Both subscales located on interaction factor	High validity
	PSSUQ	Located on both design and interaction factor	Medium validity
Acceptance	UTAUT	Effort and intention loading on respective factors Performance on acceptance dimension	Medium validity
	van-der-Laan scale	Usefulness loading on design and acceptance factor Satisfaction loading on design and interaction factor	Low validity
User experience	AttrakDiff	All subscales on design-factor	Low validity
	UEQ	High loadings on design and interaction factor Efficiency loading on design factor	Medium validity
	meCUE	Loadings on all three factors Module 1 and Module 3 factors loading as expected	High validity
Trust	ATS	Poor alignment with present factor solution Suitability for HMI evaluation questionable	Low validity
	UTA	Congruency with acceptance and usability measures Reflection of trust component in relation to ATS unclear	Medium validity

present self-report measures adhere to the external criterion of adherence to HMI design guidelines.

4.3 Limitations and future research

Analysis of reliability through the calculation of internal consistency by means of Cronbach's alpha is considered problematic. Sijtsma (2009) outlines that there are better measures for reliability such as the lower bounds (Guttman 1945) or omega (Revelle and Zinbarg 2009). For the sake of comparability with the originally reported values, we chose the Cronbach's alpha approach to reliability.

One drive in the present experiment lasted 15 min. This short amount of time restricts the possible amount of interactions with the HMI. Especially for attitudes that require long-term experience (i.e., trust, acceptance) this might represent a limitation. However, the use-cases (see Table 2) chosen for the present experiment already represent a good portion of interactions that are possible with an HMI for L3 ADS (for more possible use cases see Naujoks et al. 2018). Hence, participants could derive a good impression of HMI functionality and interaction possibilities from the 15-min driver. Still, for evaluation of the ADS itself regarding lane or distance keeping and maneuvering, a longer experience might be required to provide information about a user's acceptance and trust. Since the study thus only allows short-term evaluation of the automated vehicle HMI, only initial knowledge and interactions are the basis for user preference ratings. However, there is evidence that interaction performance (Forster et al. 2019b) and mental models that discriminate between L2 and L3 automation (Forster et al. 2019a) change with rising experience. Therefore, additional knowledge gained through experience might generate the occurrence of dissonances (Vanderhaegen and Carsten 2017). Such dissonances are characterized by an inconsistency between initial and additional knowledge and brings the potential of influencing long-term perception of the automated vehicle HMI. In this sense, future research should also consider conflicting information for different levels of automation over prolonged time periods.

As outlined in Sect. 1.2, particular circumstances apply to experimental settings in driving simulation research. The present study included $N=57$ participants. Concerning the reliability analysis, the lowest subject-to-item ratio for the entire scale is observed for the meCUE equaling 1.73 and this approach can be considered uncritical. To carry out factor analyses in psychometric evaluation, Anthoine et al. (2014) found that a large number of studies reported minimum subject-to-item ratios of close to two. Due to the present sample size, an EFA on item level was not possible here. Therefore, the EFA approach was conducted using the subscales instead of single items. The sample size of $N=57$ cases for the $n=34$ subscales refers to a subject-to-case ratio of 1.68 and thus we conclude that this approach is reasonable. Still, the sample size in relation to conducting an EFA with proposed requirements of up to 300 cases (Tabachnick and Fidell 2007) is a drawback of this study. Conclusions of the factor structure are drawn from an aggregated level. The possibility that a different factor structure for self-report measures might have emerged on item level might have emerged cannot be categorically ruled out.

The examination of criterion validity used a criterion on system level (i.e., HMI guideline compliance). Future research also needs to bring forth evidence of criterion validity on subject level. Especially for the usability-related factor of user preference, it remains to be seen whether and how well self-reports are reflected in interaction measures such as accuracy, speed or attentional demand (Wickens et al. 2015).

In that vein, Forster et al. (2018a) outlined the importance of a multi-method approach (Hornbæk 2006; Nielsen and Levy 1994) when evaluating ADS. The present work contributes to this call for methodological development as it provides empirical evidence of the suitability of different self-report measures for L3 ADS. However, future research efforts are necessary to find out about the suitability of different observational measures and their relationship with self-report measures. First empirical results on this issue are reported by Forster et al. (2019b).

The present sample was drawn from BMW employees and supplier companies. This might be a critique for the study outcome concerning self-report measures. Although BMW employees might differ from the general population in some aspects, this does not necessarily limit the external validity of the current findings. The main focus of the current study was not participants' general attitude towards automated driving and cars, but specific aspects of the HMI as targeted by the scale items. The sample consisted of people with diverse backgrounds. Among others, the sample included participants working for suppliers, business partners and interns, who differed in demographic variables (see for example age) as well as educational background (e.g., economists, psychologists, computer scientists). In this regard, it might be argued that the sample could have been even more representative of the population in question than, for example a sample drawn from college students, who have been shown to differ substantially from the population at large and "are among the least representative populations one could find for generalizing about humans" (Henrich et al. 2010). Taken together, inferences drawn from the sample investigated in the current study should also generalize to the population of drivers evaluating an ADS in the future.

The open question of how to proceed with scales that revealed limited psychometric properties remains. For example, one might still apply a scale but discard certain items that were problematic in this context (see Table 6). This could also improve reliability of the respective subscale but at the same time one could debate whether the scale still covers the initially proposed construct comprehensively. If a scale, however, brings specific instructions on how to calculate overall scores such as the SUS, deleting items is not an option. The adaptation of wording of single items is also a possible approach. Especially in acceptance research using items of the UTAUT framework, subtle differences were apparent in the works on different levels of automation (Adell 2010; Rahman et al. 2017). When adapting existing items or even adding new items due to certain peculiarities of HMI functionality, one has to consider that this influences reliability of the scale. Moreover, adding items to an already reduced item pool can affect construct validity by adding to the correlational matrix between items and potentially leading to new dimensions in factor analyses. Also it can change content validity through shifting the focus of the subscale. Some tools such as the meCUE bear the advantage of their modular nature. This means that not all subscales need to be applied but can be administered independently from each other. Therefore, if the present work for example found limited reliability of one pragmatic subscale, the other scales can still be used regardless in terms of reliability. Concluding, there are several possibilities on how to improve certain questionnaires in new contexts but one has to keep in mind that interventions might affect psychometrics not only to the better.

5 Conclusion

To conclude, SUS, UTAUT, UEQ and meCUE revealed the most positive results concerning psychometrics in L3 ADS evaluation. For an overall L3 ADS HMI evaluation, results from this study suggest to apply the meCUE when all dimensions of preference are of interest. Depending on the specific aspect of a particular study (i.e., design, interaction or acceptance evaluation), we recommend to apply scales or subscales that suit the respective purpose (see Sects. 3.4.2 and 4.2.2). The present work points towards the importance of psychometric scale evaluation in a new context. Since the L3 ADS circumstances are fundamentally different from conventional human–computer interaction, self-report measures do not necessarily work as proposed in their original context. When setting up an experiment for automated driving and HMI research, one has to face challenge of choosing between available self-report measures. The present study provides researchers and practitioners with a recommendation of self-report measures and their suitability for evaluating L3 ADS.

References

AdaptIVe Consortium (2017) Final functional human factors recommendations (Deliverable D3.3)

Adell E (2010) Acceptance of driver support systems. Proc Eur Conf Hum Centered Design Intell Transp Syst 2:475–486

Adell E, Nilsson L, Várhelyi A (2014) How is acceptance measured? Overview of measurement issues, methods and tools. In: Horberry T, Regan MA, Stevens A (eds) Driver acceptance of new technology theory measurement and optimisation. CRC Press, London, UK, pp 73–89

Anthoine E, Moret L, Regnault A, Sébille V, Hardouin J-B (2014) Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. Health Qual Life Outcomes 12:176. https://doi.org/10.1186/s12955-014-0176-2

Auricht M, Stark R, Blume C (2014) Integrating user experience validation into a new engineering development process for advanced driver assistance systems. In: Boyle LN (ed) The 6th international conference of automotive user interfaces and interactive vehicular applications. Seattle, WA, USA

Bangor A, Kortum P, Miller J (2009) Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 4(3):114–123

Bazilinskyy P, Eriksson A, Petermeijer B, de Winter J (2017) Usefulness and satisfaction of take-over requests for highly automated driving. In: Road safety and simulation international conference (RSS 2017), The Hague, Netherlands

Beggiato M, Pereira M, Petzoldt T, Krems JF (2015) Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. Transp Res Part F Traffic Psychol Behav 35:75–84. https://doi.org/10.1016/j.trf.2015.10.005

Brooke J (1996) SUS—a quick and dirty usability scale. Usability Eval Ind 194(189):4–7

Bubb H, Bengler K, Grünen RE, Vollrath M (2015) Automobilergonomie. Springer, Berlin

Bühner M (2011) Einführung in die Test-und Fragebogenkonstruktion. Pearson Deutschland GmbH, München

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 56(2):81

Chien S-Y, Semnani-Azad Z, Lewis M, Sycara K (2014) Towards the development of an inter-cultural scale to measure trust in automation. In: International conference on cross-cultural design

Crash Avoidance Metrics Partnership (2016) Automated vehicles research for enhanced safety. NHTSA, Department of Transportation, Washington, DC

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334. https://doi.org/10.1007/BF02310555

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281–302. https://doi.org/10.1037/h0040957

Davis FD (1985) A technology acceptance model for empirically testing new end-user information systems: theory and results. Massachusetts Institute of Technology, Boston

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quart 13(3):319. https://doi.org/10.2307/249008

Eichinger A (2011) Untersuchungskonzepte für die Evaluation von Systemen zur Erkennung des Fahrerzustands: BASt-Forschungsbericht: FE 82.369/2009. Berichte der Bundesanstalt für Straßenwesen 80:45–94

Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. J Market Res 18:39–50

Forster Y, Naujoks F, Neukum A (2016) Your turn or my turn? Design of a human–machine interface for conditional automation. In: Green P (ed) Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications. Ann Arbor, MI, USA

Forster Y, Naujoks F, Neukum A (2017) Increasing anthropomorphism and trust in automated driving functions by adding speech output. Intelligent Vehicles Symposium (IV), 2017 IEEE, Redondo Beach, California, USA

Forster Y, Kraus J, Feinauer S, Baumann M (2018a) Calibration of trust expectancies in conditionally automated driving by brand, reliability information and introductionary videos: an online study. In: Donmez B, Walker BN, Fröhlich K (Chairs) Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications, Toronto, CN

Forster Y, Hergeth S, Naujoks F, Krems, JF (2018b) How usability can save the day: methodological considerations for making automated driving a success story. In: Donmez B, Walker BN, Fröhlich K (Chairs) Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications, Toronto, CN

Forster Y, Hergeth S, Naujoks F, Beggiato M, Krems JF, Keinath A (2019a) Learning and development of mental models in interaction with driving automation: a simulator study. Driving Assessment Conference, Santa Fe, NM, USA

Forster Y, Hergeth S, Naujoks F, Beggiato M, Krems JF, Keinath A (2019b) Learning to use automation: behavioral changes in interaction with automated driving systems. Transp Res Part F Traffic Psychol Behav 62:599–614

Forster Y, Hergeth S, Naujoks F, Krems JF, Keinath A (2019) Empirical validation of a checklist for heuristic evaluation of automated vehicle HMIs. In: 10th international conference on applied human factors and ergonomics, Washington D.C., USA

François M, Osiurak F, Fort A, Crave P, Navarro J (2016) Automotive HMI design and participatory user involvement: review and perspectives. Ergonomics 60(4):541–552

Frison A-K, Wintersberger P, Riener A, Schartmüller C (2017) Driving hotzenplotz: a hybrid interface for vehicle control aiming to maximize pleasure in highway driving. In: Boll (ed) Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications, Oldenburg, Germany

Frison A-K, Wintersberger P, Riener A, Schartmüller C, Boyle LN, Miller E, Weigl K (2019) UX We Trust: investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. The 2019 CHI Conference, Glasgow, UK

Ghazizadeh M, Lee JD, Boyle LN (2012) Extending the technology acceptance model to assess automation. Cognit Technol Work 14(1):39–49

Gold C, Körber M, Hohenberger C, Lechner D, Bengler K (2015) Trust in automation—before and after the experience of take-over scenarios in a highly automated vehicle. Procedia Manufact 3:3025–3032

Gold C, Naujoks F, Radlmayr J, Bellem H, Jarosch O (2017) Testing scenarios for human factors research in level 3 automated vehicles. In: International conference on applied human factors and ergonomics, Los Angeles, CA, USA

Green P, Levison W, Paelke G, Serafin C (1994) Suggested human factors design guidelines for driver information systems. UMTRI, Michigan

Guttman L (1945) A basis for analyzing test–restest reliability. Psychometrika 10:255–282

Hassenzahl M (2001) The effect of perceived hedonic quality on product appealingness. Int J Hum Comput Interact 13(4):481–499. https://doi.org/10.1207/S15327590IJHC1304_07

Hassenzahl M, Burmester M, Koller F (2003) AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In: Ziegler J, Szwillus G (eds) Mensch & Computer 2003. Interaktion in Bewegung. B. G. Teubner, Stuttgart, Leipzig, pp 187–196

Häuslschmid R, von Bülow M, Pfleging B, Butz A (2017) SupportingTrust in autonomous driving. In: Papadopoulos GA, Kuflik T, Chen F, Duarte C, Fu W-T (Chairs) The 22nd international conference on intelligent user interfaces, Limassol, Cyprus

Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? Behav Brain Sci 33(2–3):61–83. https://doi.org/10.1017/S0140525X0999152X **(discussion 83–135)**

Hergeth S (2016) Automation trust in conditional automated driving systems: approaches to operationalization and design. PhD Thesis. Chemnitz University of Technology. Chemnitz, Germany

Hergeth S, Lorenz L, Krems JF (2017) Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. Hum Factors 59(3):457–470

Hoff K, Bashir M (2016) Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434

Hornbæk K (2006) Current practice in measuring usability: challenges to usability studies and research. Int J Hum Comput Stud 64(2):79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

Hornbæk K, Law EL-C (2007) Meta-analysis of correlations among usability measures. The 2007 CHI conference, San Jose, California, USA. https://doi.org/10.1145/1240624.1240722

International Organization for Standardization (2018) Ergonomics of Human–System Interaction—Part 11: Usability: Definitions and Concepts. Geneva, Switzerland: ISO 9241-11

ISO (2012) Road vehicles—ergonomic aspects of transport information and control systems—calibration tasks for methods which assess driver demand due to the use of in-vehicle systems. (ISO, 14198). Geneva, Switzerland

Jarosch O, Kuhnt M, Paradies S, Bengler K (2017) It's out of our hands now! Effects of non-driving related tasks during highly automated driving on drivers' fatigue. In: Driving Assessment Conference, Manchester Village, Vermont, USA

Jian J-Y, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cognit Ergon 4(1):53–71

Jones PS, Lee JW, Phillips LR, Zhang XE, Jaceldo KB (2001) An adaptation of Brislin's translation model for cross-cultural research. Nursing Res 50(5):300–304

Kline TJB (1999) The team player inventory: reliability and validity of a measure of predisposition toward organizational team-working environments. J Specialists Group Work 24(1):102–112

König M, Neumayr L (2017) Users' resistance towards radical innovations: the case of the self-driving car. Transp Res Part F Traffic Psychol Behav 44:42–52

Kyriakidis M, Happee R, de Winter JCF (2015) Public opinion on automated driving: results of an international questionnaire among 5000 respondents. Transp Res Part F Traffic Psychol Behav 32:127–140

Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. Symp Aust HCI Usability Eng Group. https://doi.org/10.1002/9783527617272.ch1

Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80

Lewis JR (2002) Psychometric evaluation of the PSSUQ using data from 5 years of usability studies. Int J Hum Comput Interact 14(3–4):463–488. https://doi.org/10.1515/9783110887242.3

Lewis JR, Sauro J (2009) The factor structure of the system usability scale. International conference on human centered design

Lüdtke O, Robitzsch A, Trautwein U, Köller O (2007) Umgang mit fehlenden Werten in der psychologischen Forschung. Psychologische Rundschau 58(2):103–117. https://doi.org/10.1026/0033-3042.58.2.103

Minge M, Thüring M, Wagner I, Kuhr CV (2016) The meCUE questionnaire: a modular tool for measuring user experience. In: Advances in ergonomics modeling 2016, pp 115–128

Naujoks F, Purucker C, Neukum A, Wolter S, Steiger R (2015) Controllability of partially automated driving functions—does it matter whether drivers are allowed to take their hands off the steering wheel? Transp Res Part F Traffic Psychol Behav 35:185–198

Naujoks F, Forster Y, Wiedemann K, Neukum A (2016) Speech improves human-automation cooperation in automated driving. In: Workshopband Mensch und Computer 2016. Aachen, Germany

Naujoks F, Forster Y, Wiedemann K, Neukum A (2017) Improving usefulness of automated driving by lowering primary task interference through HMI design. J Adv Transp. https://doi.org/10.1155/2017/6105087

Naujoks F, Hergeth S, Keinath A, Wiedemann K, Schömig N (2018) Use cases for assessing, testing, and validating the human–machine interface of automated driving systems. Human Factors and Ergonomics Society Annual Meeting, Philadelphia

Naujoks F, Hergeth S, Wiedemann K, Schömig N, Forster Y, Keinath A (2019a) Test procedure for evaluating the human–machine-interface of vehicles with automated driving. Traffic Injury Prevent 20:146–151

Naujoks F, Wiedemann K, Schömig N, Hergeth S, Keinath A (2019b) Towards guidelines and verification methods for automated vehicle HMIs. Transp Res Part F Traffic Psychol Behav 60:121–136

Nees M (2016) Acceptance of self-driving cars: an examination of idealized versus realistic portrayals with a self-driving cars acceptance scale. In: Human Factors and Ergonomics Society 60th Annual Meeting, Washington, D.C

Nielsen J, Levy J (1994) Measuring usability: preference vs. performance. Commun ACM 37(4):66–75

Nordhoff S, van Arem B, Happee R (2016) Conceptual model to explain, predict, and improve user acceptance of driverless podlike vehicles. Transp Res Record J Transp Res Board 2602:60–67. https://doi.org/10.3141/2602-08

Norman D, Miller J, Henderson A (1995) What you see, some of what's in the future, and how we go about doing it: HI at apple computer. In: Conference companion on human factors in computing systems

Nunes A, Reimer B, Coughlin JF (2018) People must retain control of autonomous vehicles. Nature 556(7700):169–171. https://doi.org/10.1038/d41586-018-04158-5

Nunnally JC (1978) Psychometric theory. McGraw-Hill, New York

O'connor BP (2000) SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. Behav Res Methods Instrum Comput 32(3):396–402. https://doi.org/10.3758/BF03200807

Osswald S, Wurhofer D, Trösterer S, Beck E, Tscheligi M (2012) Predicting information technology usage in the car: towards a car technology acceptance model. In: Kun AL, Boyle LN, Reimer B, Riener A (Chairs) Proceedings of the 4th international conference on automotive user interfaces and interactive vehicular applications. symposium conducted at the meeting of ACM, Portsmouth, New Hampshire, US

Payre W, Cestac J, Delhomme P (2014) Intention to use a fully automated car: attitudes and a priori acceptability. Transp Res Part F Traffic Psychol Behav 27:252–263. https://doi.org/10.1016/j.trf.2014.04.009

Rahman MM, Lesch MF, Horrey WJ, Strawderman L (2017) Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. Accid Anal Prevent 108:361–373. https://doi.org/10.1016/j.aap.2017.09.011

Revelle W, Zinbarg RE (2009) Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. Psychometrika 74(1):145

Society of Automotive Engineers International J3016 (2018) Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. SAE International, Warrendale, PA

Sijtsma K (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74(1):107–120. https://doi.org/10.1007/s11336-008-9101-0

Spain RD, Bustamante EA, Bliss JP (2008) Towards an empirically Developed Scale for System Trust: take Two. Proc Hum Factors Ergon Soc Ann Meeting 52(19):1335–1339. https://doi.org/10.1177/154193120805201907

Stevens A, Quimby A, Board A, Kersloot T, Burns P (2002) Design guidelines for safety in-vehicle information systems. TRL Limited, Crowthorne

Tabachnick BG, Fidell LS (2007) Using multivariate statistics. Allyn & Bacon/Pearson Education, Boston

Tullis TS, Boynton TL, Hersh H (1995) Readability of fonts in the windows environment. Conference companion on Human factors in computing systems

Van der Laan JD, Heino A, de Waard D (1997) A simple procedure for the assessment of acceptance of advanced transport telematics. Transp Res Part C Emerg Technol 5(1):1–10

Vanderhaegen F, Carsten O (2017) Can dissonance engineering improve risk analysis of human–machine systems? Cogn Technol Work 19(1):1–12. https://doi.org/10.1007/s10111-017-0405-7

Velicer WF (1976) Determining the number of components from the matrix of partial correlations. Psychometrika 41(3):321–327. https://doi.org/10.1007/BF02293557

Venkatesh V, Morris MG, Davis GB, Davis FD (2003) User acceptance of information technology: toward a unified view. MIS Quart (27:3):425–478

Verberne FMF, Ham J, Midden CJH (2012) Trust in smart systems: sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Hum Factors 54(5):799–810

Walch M, Baumann M, Jaschke L, Weber M, Hock P (2017) Touch screen maneuver approval mechanisms for highly automated vehicles: a first evaluation. Adjunct proceedings of the 9th international ACM conference on automotive user interfaces and interactive vehicular applications, Oldenburg, Germany

Waytz A, Heafner J, Epley N (2014) The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol 52:113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wickens CD, Hollands JG, Banbury S, Parasuraman R (2015) Engineering psychology and human performance. Psychology Press, Boca Raton

Zoellick JC, Kuhlmey A, Schenk L, Schindel D, Blüher S (2019) Assessing acceptance of electric automated vehicles after exposure in a realistic traffic environment. PLOS One. https://doi.org/10.1371/journal.pone.0215969