Fabienne Roche

Assessing subjective criticality of take-over situations: Validation of two rating scales

Citation details
Roche, F. (2021). Assessing subjective criticality of take-over situations: Validation of two rating scales. In Accident Analysis amp; Prevention (Vol. 159, p. 106216). Elsevier BV.
https://doi.org/10.1016/j.aap.2021.106216.

Title: Assessing subjective criticality of take-over situations: Validation of two rating scales

Author: Fabienne Roche

Corresponding author: Fabienne Roche, fabienne.roche@freenet.de

Address: Technische Universität Berlin

Fachgebiet Kognitionspsychologie und Kognitive Ergonomie

MAR 3-2

Marchstraße 23

10587 Berlin, Germany

Running head: Assessing Subjective Criticality

Manuscript type: Research paper

Declarations of interests: none

Funding sources involved: This work has resulted from the interdisciplinary research project "Analysis and Support of Driver Interventions in Dynamic: Critical Situations during Highly Automated Driving" funded by the German Research Community (Deutsche Forschungsgemeinschaft, DFG, Grant No. 326727090).

Abstract

Assessing subjective criticality of take-over situations is crucial for understanding of take-over behavior and comparing studies. However, no validated rating scales exist that assess subjective criticality of take-over situations. In a driving simulator study, two rating scales, the Scale of Criticality Assessment of driving situations from Neukum et al. (2008) and the Criticality Rating Scale, were tested on their validity to assess the subjective criticality of take-over situations. Besides, the subjective and behavioral changes over the repeated experience of take-over situations were investigated. Twenty-five participants experienced a set of five take-over situations with varying time-to-collisions (TTC) at the moment of the take-over request, twice. After each of the first five take-over situations, participants rated the criticality on one scale, after each of the second five situations on the other scale. Correlation coefficients between TTCs and criticality ratings for each scale were calculated. Also, the changes of subjective and behavioral measures over the trials were investigated. Correlation coefficients indicated a strong correlation between criticality ratings and TTCs. Hence, both scales are equally valid for the assessment of the criticality of take-over situations. The repeated experience of the take-over situations did not affect effort ratings, take-over times, or steering wheel positions. But brake input decreased with increasing practice, indicating a safer take-over behavior. Hence, results of studies with repeated experience of take-over situations are relatively valid as only brake behavior changed with increasing practice.

Keywords: Automated driving, Driver-vehicle interaction, Driver behavior, Criticality, Take-over behavior, Scale validation

1.1 Introduction

In the past years, human factors researchers have accumulated an impressive amount of knowledge, especially on the take-over process (Gold et al., 2016; Jamson et al., 2013; Körber et al., 2016; Murata et al., 2013; Politis et al., 2014; Roche et al., 2018; SAE International, 2018). It was observed that different characteristics of take-over situations may heavily influence the take-over behavior and subjective experience (Damböck et al., 2012; Gold et al., 2016, 2013; Radlmayr et al., 2014; Roche & Brandenburg, 2020, 2018). One of these characteristics is the objective criticality of the take-over situation which is determined by situational parameters. For example, take-over situations with low time budgets are more critical than situations with high time budgets. The objective criticality affects the take-over behavior and subjective criticality, e.g. more extreme behavior and higher subjective criticality when the situation is more critical. There are many options to assess take-over behavior, such as take-over times or steering behavior. It provides insights into how drivers behave depending on different situational parameters. In contrast, to our knowledge, no validated instrument exists to assess the subjective criticality, even though, it supports the interpretation of observed take-over behavior and may enable comparisons between different take-over situations. Therefore, in the present study, two rating scales are validated regarding their suitability to assess the subjective criticality of take-over situations. Besides, subjective and behavioral changes of the repeated experience of take-over situations are investigated.

1.2 Objective Criticality of Driving Situations

The objective criticality of a driving situation is ‘the accident risk’ (Rodemerk et al., 2012, p. 1). Hence, a driving situation, in which a collision is inevitable, constitutes the highest possible objective criticality (Rodemerk et al., 2012). Especially in automated driving, the

objective criticality of a driving situation is crucial since it influences the take-over behavior (Gold et al., 2013; Roche & Brandenburg, 2020, 2018; Zhang et al., 2019).

The objective criticality of a driving situation may be determined by situational parameters such as time budget (Junietz et al., 2017), traffic density, or visibility. Lower time budgets, higher traffic densities, or poor visibility may lead to a higher objective criticality. In this paper, we focus on time budget. It can be quantified by time-to-collision (TTC). TTC describes the available time until a vehicle would collide with a reference object (Vogel, 2003). A reference object may be a preceding vehicle or a system boundary, such as an obstacle on the road. Hence, shorter TTCs indicate a more critical situation. These more critical situations may emerge in case the automated driving system reaches its limits or in the case of driver-initiated take-overs (Roche et al., 2020).

TTC is known to influence take-over behavior. Numerous driving simulator studies varied the TTC in take-over situations. Lower TTCs, hence more critical take-over situations, were associated with lower take-over times (Gold et al., 2013; Roche & Brandenburg, 2020, 2018; Zhang et al., 2019), higher decelerations (Roche et al., 2020; Roche & Brandenburg, 2020, 2018), and larger steering wheel angles (Roche et al., 2020; Roche & Brandenburg, 2020, 2018). While lower take-over times are a desirable behavior, high decelerations and extreme steering are a threat to the drivers' safety for the following reasons: This behavior may result in (a) vehicle instability, (b) rear-end collisions with following vehicles, (c) collisions with vehicles on neighboring lanes or (d) lane departures. Indeed, more critical take-over situations in terms of lower TTCs led to higher error rates, such as collisions or missing lane changes (Damböck et al., 2012), more lane departures (Mok, Johns, Lee, Ive, et al., 2015; Mok, Johns, Lee, Miller, et al.,

2015), and more collisions (Roche & Brandenburg, 2018). This impaired performance points at the threat of take-overs in situations with low TTCs.

1.3 Subjective Criticality of Driving Situations

Analog to the definition of objective criticality, subjective criticality may be defined as the perceived threat or risk of a driving situation (Rodemerk et al., 2012). Hence, situations that are objectively more critical are highly likely to be perceived as more critical. However, next to the objective criticality, further aspects may affect the subjective criticality. These are individual parameters such as the driver's personality (Banet & Bellet, 2008; Mesken et al., 2007), fatigue (Feldhütter et al., 2018), familiarity with the take-over situation (Hergeth et al., 2017), or perceived capability (Fuller, 2011). For instance, it has been demonstrated that fatigued drivers rate the same situations as more critical than alert drivers indicating that they are more stressed (Feldhütter et al., 2018). And Hergeth et al. (2017) observed that criticality ratings decreased with increasing familiarity with the take-over situation.

There are two reasons why it is crucial to assess subjective criticality of take-over situations. First, situational parameters may have diverse effects on take-over behavior and may interact. Assessing the subjective criticality of take-over situations would promote the understanding of the observed behavior. Second, a criticality rating facilitates the comparability of driving situations and the evaluation of take-over behavior. Rodemerk et al. (2012) advocated for the need to compare driving situations and introduced a general criticality criterion to do so. Similarly, Jarosch and Bengler (2018) suggested a holistic view to evaluate the take-over behavior adequately rather than consider the parameters separately. They argue that, for example, a fast reaction cannot generally be evaluated as a good reaction, but has to be looked at in combination with steering and braking behavior. However, both criteria from Rodemerk et al.

(2012) and Jarosch and Bengler (2018) are based on objective parameters, e.g. collision probability. They do not include the subjective aspect of a take-over situation. In contrast, Radlmayr et al. (2018) included a subjective rating parameter to evaluate take-over behavior. Together with two further parameters, it ought to promote the understanding and the comparability of take-over situations (Radlmayr et al., 2018). In line with Radlmayr et al. (2018), we argue that a validated and anchored scale to assess subjective criticality would enable the comparability of different driving situations and the evaluation of take-over behavior in driving simulator studies and real traffic.

In numerous studies, different scales for assessing the subjective criticality of driving situations have been employed. These are, for example, a multi-item Likert-scale (Banet & Bellet, 2008), an eleven-point, single item scale developed by Neukum et al. (2008), or seven-point Likert-scales (Radlmayr et al., 2018; Roche & Brandenburg, 2020, 2018). However, to our knowledge, no publication is available that validates one of them.

1.3.1 Scale of Criticality Assessment of Driving Situations

The *Scale of Criticality Assessment of driving situations* (SCA; Neukum et al., 2018) is a scale that assesses subjective criticality of a driving situation (see figure 1, left in English, right in German). It was already used in various studies (Hergeth et al., 2017; Naujoks et al., 2017; Neukum et al., 2008; Siebert et al., 2014). The scale is a modified version of the judgment *Scale for the Assessment of the Experienced Degree of Disturbance* (Neukum & Krüger, 2003) that was developed based on the Cooper-Harper-Scale (Cooper & Harper, 1969). It is an eleven-point, single item scale and based on a two-step rating procedure. First, participants are asked to rate the criticality of the driving situation by selecting one of the five verbal categories: ‘imperceptible’ (0 pt.), ‘harmless’ (1-3 pts.), ‘unpleasant’ (4-6 pts.), ‘dangerous’ (7-9 pts.), ‘uncontrollable’ (10 pts.,

translation based on Naujoks et al., 2017). Participants are instructed that the driving situation shall be rated concerning the necessary compensatory effort. Second, they are asked to specify their rating by selecting one of the three numerical subcategories of each category (right area in figure 1). It was assumed that 'imperceptible' and 'uncontrollable' are not divisible any further. Hence, these both extreme categories have only one subcategory, 0 respectively 10 pts.

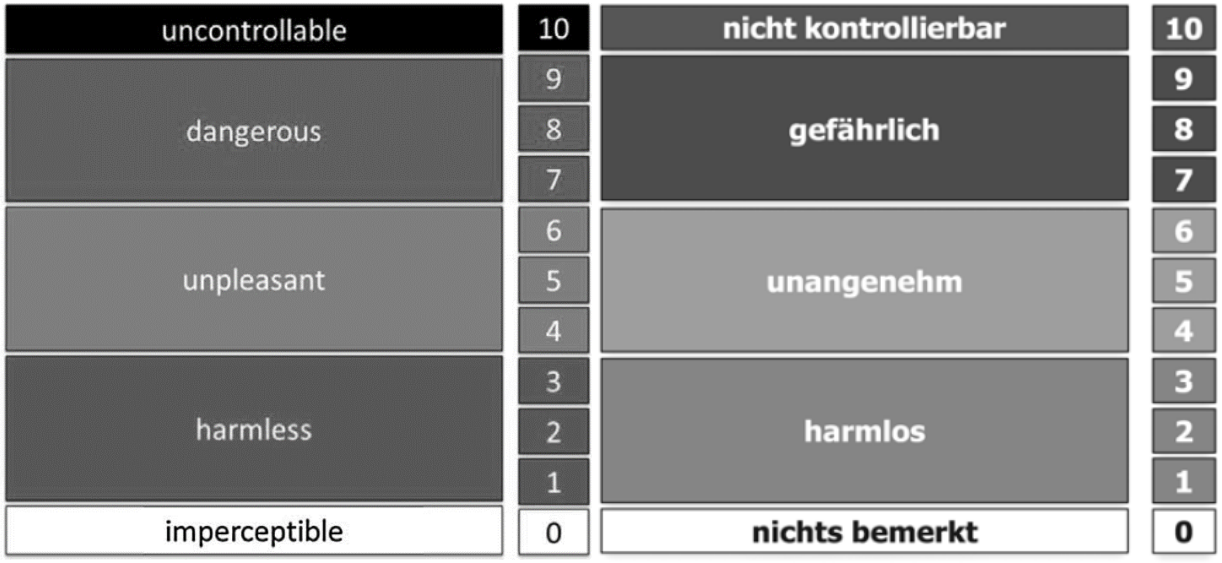


Figure 1: Scale of Criticality Assessment of driving situation (SCA) in English (left, depicted from Naujoks et al., 2017) and in German (right, Neukum et al., 2008). The German version was used in the present study.

This scale holds advantages and disadvantages. On the one hand, Neukum and Krüger (2003) state that an advantage of this scale is the threshold distinguishing between tolerable and intolerable situations (rating above 6). The magnitude of ratings should be, therefore, comparable between participants. Another advantage is its sensitivity that allows for differentiated ratings of subjective criticality across different driving situations (Neukum & Krüger, 2003). Besides, the original scale was highly accepted by naïve and expert participants (Neukum & Krüger, 2003). On the other hand, the scale does not take into account whether the driver's perceived capability

to deal with the driving situation affects the rating, as suggested by Fuller (2011). Hence, drivers that feel very capable might rate a driving situation as tolerable, while others rate it as intolerable. In addition, it is questionable whether the differences between the numerical values are equidistant. For example, it may be assumed that the experienced differences within one category (e.g. two vs. three) differ from the experienced differences between two categories (e.g. three vs. four). Also, the effort for the application and analysis is high because the instruction takes longer and it has to be ensured that the rating of the verbal category corresponds to the rating of the numerical subcategory.

1.3.2 Criticality Rating Scale

The *Criticality Rating Scale* (CRS) is a modification of a criticality scale used in previous studies (Roche & Brandenburg, 2020, 2018). It was modified in the course of this research project aiming at compensating known disadvantages of existing rating scales. In the instruction of the item, it is clearly stated that the criticality should be rated concerning real driving situations. The wording of the question 'How critical did you perceive the experienced driving situation?' avoids complicated syntax, specific terms, and ambiguity as recommended by Moosbrugger and Kelava (2020). The Criticality Rating Scale consists of a single item rating scale. A continuous scale with tick marks is used to visualize the gradation of the rating (see figure 2). This is in contrast to the SCA, but similar to the NASA-TLX scale (Hart & Staveland, 1988). The NASA-TLX is a well-known and widely used tool for the assessment of perceived workload (Hart, 2006). In line with the NASA-TLX, the CRS is designed with 100 points (1-100). The poles are labeled with 'not critical at all' (1 pt., in German 'gar nicht kritisch') and 'very critical' (100 pts., in German 'sehr kritisch'). The verbal labeling is based on the results of Rohrmann's study on rating scales (Rohrmann, 1978). He found that the German versions of 'not

at all' and 'very' were rated as the lowest and highest intensity terms with small scatter among eighteen terms. To allow the perception of an equidistant gradation, no more verbal or numerical anchors had been employed. Moosbrugger and Kelava (2020) recommended to omit a middle answer category because it is often used as a fallback option when the participant does not understand the question, refuses to answer, or does not know the answer. Hence, the number of ticks of the CRS was even to omit a middle answer category (see figure 2). In contrast, the SCA has a middle category (5 pts.). In doing so, the assumed disadvantages of the SCA shall be addressed.

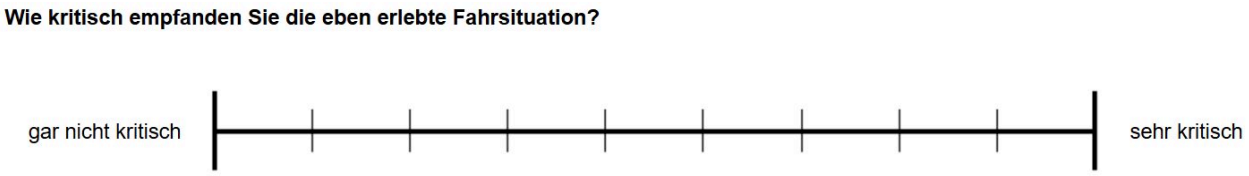


Figure 2: The Criticality Rating Scale (CRS) with poles 'not critical at all' (1 pt.) and 'very critical' (100 pts.).

The advantages of this scale are the following. First, less inter-individual interpretations are assumed since only the poles are labeled as opposed to the SCA. Thereby, a different understanding of the used labels by the raters is less likely. Second, the rating differences between all scale points should be equidistant since the scale points are not additionally labeled. Third, the test efficiency is supposed to be higher than with the SCA due to the short instruction and the familiar rating system. A potential disadvantage of the CRS is the missing threshold that would enable comparability of ratings between raters as assumed with the SCA.

1.4 Test validity

Tests should be valid to ensure they are truly measuring what they are supposed to measure (Hartig et al., 2008). The Standards for Educational and Psychological Testing (2018) state that test validity is the most important quality criteria of a test, next to objectivity and reliability. A test is objective when its result is independent of the experimenter and analyzer (Hartig et al., 2008). A test is reliable when it is precise, i.e. an elastic measuring tape would not be reliable when measuring length (Hartig et al., 2008). Objectivity and reliability are requirements to ensure test validity. It should be noted that none of the quality criteria is binary, hence, a test can be more or less valid, objective, or reliable. For the sake of this study, we assumed that both scales achieve a certain level of reliability due to their wording and scale design. Besides, we aimed at establishing objectivity when applying the scales (more details see Method section).

Different types of test validity exist, among them construct validity. It describes the extent to which a test examines a psychological trait or construct, as defined by theory (Cronbach & Meehl, 1955). Construct validity is considered as the most fundamental type of test validity (Wainer & Braun, 2013) and composes of convergent and discriminant validity. Convergent validity is present when measurements of a construct that are recorded with different methods correlate strongly (Moosbrugger & Kelava, 2020). Usually, a new method is validated by means of another established method. Discriminant validity is high when measurements of different constructs that are recorded using the same or different methods correlate weakly (Moosbrugger & Kelava, 2020). In this study, we focus on investigating the convergent validity of both scales.

1.5 Repeated experience of take-over situations

The repeated experience of similar take-over situations is quite common in driving simulator studies due to the experimental setting and test efficiency. This may lead to subjective and behavioral changes. On the one hand, it would be plausible that take-over behavior deteriorates over the course of an experiment. Reasons for deterioration are increasing fatigue or increasing trust in the system (Hergeth et al., 2016). Indeed, there are indications that with increasing practice take-over behavior becomes riskier: maximal deceleration increases (Brandenburg & Roche, 2020) and observation decreases (Hergeth et al., 2016; Roche et al., 2018). On the other hand, it could be that take-over behavior improves due to the increasing practice. This was demonstrated by Hergeth et al. (2017), Körber et al. (2016), and Payre et al. (2016) concerning decreasing take-over times, larger TTCs, lower maximal lateral accelerations, and lower maximal longitudinal decelerations. This shows that the evidence regarding the effect of experience is not unambiguous.

1.6 Research questions

The two rating scales are supposed to measure the subjective criticality of take-over situations. Hence, testing their validity requires checking whether they properly measure criticality. As a first step, the convergent validity of the scales is tested. Therefore, a variation of objective criticality in the take-over situations should be represented in the criticality ratings. Hence, the following research questions (RQ) are investigated:

- Research question 1: Is the Scale of Criticality Assessment of driving situations (SCA) a valid tool for the assessment of subjective criticality in take-over situations?
- Research question 2: Is the Criticality Rating Scale (CRS) a valid tool for the assessment of subjective criticality in take-over situations?

- Research question 3: Do both scales differ regarding their validity?

For this, objective criticality is varied by the time-to-collision, an established method to vary criticality. In case the ratings and TTC-values correlate strongly, a high convergent validity can be assumed.

The repeated experience of similar take-over situations would indicate whether the rating scales are robust in assessing subjective criticality and whether increasing familiarity affects the ratings. In addition, since our participants experience several monotonous trials, it likely leads to fatigue and lower arousal. De Waard (2002) stated that passive fatigue may be compensated by increasing effort. Hence, increasing perceived effort ratings over the repetition of trials would demonstrate passive fatigue. Furthermore, the available research on the repeated experience of take-over situations indicates that behavioral change may take place. This leads to the fourth research question:

- Research question 4: Do drivers' criticality and effort ratings and take-over behavior change over the repeated experience of take-over situations?

2 Method

The objective criticality of the take-over situation is manipulated by the time-to-collision to a stationary obstacle at the moment of the take-over request. A lane change was chosen as the take-over situation because it is one of the most common maneuvers on the highway, where higher levels of automated driving will be deployed first (Bellem et al., 2017). In a driving simulator, participants experienced five take-over situations twice that varied regarding time-to-collisions. After each of the first five take-over situations, participants rated the criticality on one scale, after each of the second five situations on the other scale.

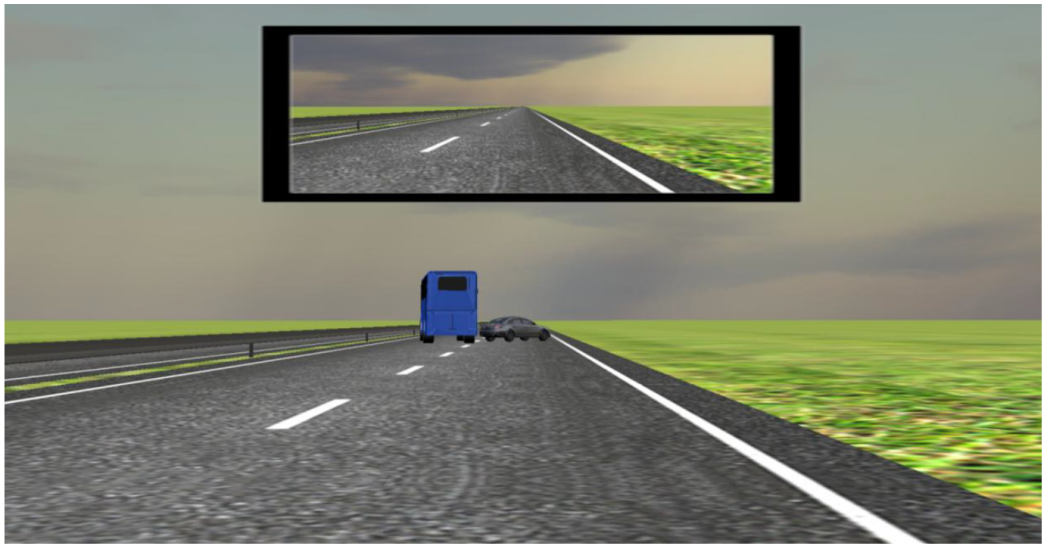
2.1 Participants

Twenty-five persons (13 women, 12 men) between 21 and 37 years of age (M = 27.3 years, SD = 4.8 years) took part in the driving simulator study. An a-priori power analysis with G*Power (Version 3.1.9.2) revealed that a sample size of 23 participants was required to detect a correlation of r = .5 with a given alpha of .05 and a power of .80. With 25 participants, each of the five TTC-values could be presented at each position across the experiment five times, e.g. five participants experienced the shortest TTC-value in the first trial, five in the second. All participants had to be German native or near-native speakers to follow the German instructions. They were required to have been holding a driving license for a minimum of two years (M = 9.6 years, SD = 4.7, Max = 19 years). Twenty participants (80%) were students. Ten participants (40%) reported having experience with advanced driver-assistance systems, such as adaptive cruise control or lane change assistance systems. On average, they used their car at 2.8 days per week. The student participants received course credits as gratification. The experiment was approved by the ethics committee of the Department of Psychology and Ergonomics of Technische Universität Berlin, Germany, and its conditions complied with the tenets of the Declaration of Helsinki. Participants gave their informed consent before the experiment started.

2.2 Materials

The experiment was conducted in a mid-fidelity driving simulator of the Department of Psychology and Ergonomics of Technische Universität Berlin. The same driving simulator was used in Roche and Brandenburg (2020, 2018). It consists of a Volkswagen™ vehicle mock-up including Fanatec pedals, a Fanatec steering wheel, a dashboard, a seat, and a gear shift. Since the simulated vehicle was automatic, the gear shift and clutch pedal were irrelevant for this study. OpenDS 4.5 was utilized to simulate the driving environment: a two-lane rural road including a

crash barrier and other vehicles on both lanes. The driving scene was projected on a screen placed at a distance of 0.80 m to the vehicle mock-up. The size of the projection screen was 3 m x 1.70 m. An image resolution of 1920 x 1080 pixels and a frequency of 60 Hz were used. A rear-view mirror was embedded in the projection (see figure 3). A driving automation corresponding to SAE-level 3 (SAE International, 2018) was active once the experimental trial started. It could be deactivated by a steering wheel or brake pedal input by the driver. A take-over was detected when the steering wheel positions exceeded 0.14 % or the brake pedal position exceeded 0.1 % (Roche & Brandenburg, 2020, 2018). Driving noise and auditory signals were played back via two speakers behind the driver seat. An iPad with standard factory settings was used to administer all questionnaires via the online survey service SoSci-Survey version 3.1 (www.soscisurvey.de). That way, the scales were always presented in the same manner and the ratings could be given without the experimenter being able to see them diminishing the interviewer effect (Bogner & Landrock, 2016). Due to that, the application of the scales was objective to a certain extent.



289 Figure 3: Take-over situation with 2.5 s TTC, the crashed vehicle in grey, and the lead vehicle in
290 blue. The rectangle in the upper-middle represents the rear-view mirror.

291 Subjective criticality was assessed with the Scale of Criticality Assessment of driving
292 situations (SCA; Neukum et al., 2008; see figure 1 and section 1.3.1) or the Criticality Rating
293 Scale (CRS; see figure 2 and section 1.3.2), depending on the block.

294 The effort-subscale of the NASA-TLX was used to assess the perceived effort (Hart &
295 Staveland, 1988). Participants were asked to answer the question 'How hard did you have to
296 work to accomplish your level of performance?'. The scale is a unipolar, single item with tick
297 marks to visualize the gradation of the rating from 'low' (0 pts.) to 'high' (100 pts., see figure 4).
298 We used the German translation by Sepehr (1988). It was applied to assess a further aspect of
299 subjective experience and to investigate possible fatigue over the repetition of trials because
300 driver fatigue may be compensated by higher effort (de Waard, 2002).

301

302 Figure 4: Effort-subscale of the NASA-TLX in German (Hart & Staveland, 1988) ranging from
303 'low' to 'high'.

304 2.3 Procedure and Experimental Design

305 The experiment consisted of an instruction phase, a familiarization phase, a training
306 phase, an experimental phase, and a final interview. In the instruction phase, participants were
307 welcomed and instructed about the procedure of the experiment and the handling of personal
308 data. Then, they were asked to read and sign the informed consent, and answer a demographic

questionnaire, i.e. age and profession. The questionnaire and all following rating scales and questions were presented on the iPad.

In the familiarization phase, participants drove on a two-lane highway for about 3 min in the driving simulator. They practiced accelerating, decelerating, and lane changing to familiarize themselves with the driving simulator.

In the training phase, participants were introduced to the automated system and the driving task. The system was designed to take over longitudinal and lateral control for specific driving tasks, depicting a system at SAE-level 3 (SAE International, 2018). Each of the training trials started with an automatic acceleration of the participant's vehicle to 100 km/h. The vehicle drove on the right lane of a two-lane rural road. Participants were instructed to take hands off the steering wheel and feet off the pedals while driving automated. Upon an acoustic cue, they were instructed to take back control by steering or braking and steer around a construction work on their lane. The acoustic cue consisted of two consecutive sounds with a duration of 0.5 s each, a frequency of 780 Hz, and a volume of approximately 80 dB. Five training trials were driven with varying TTCs at the moment of the acoustic cue: 3.10, 3.35, 3.60, 3.85, and 4.10 s. The TTC-values were on a medium range and the order was balanced across participants. After the last training trial, participants rated the subjective criticality on the SCA and the CRS and their perceived effort on the subscale of the NASA-TLX. The ratings did not enter into the analysis; they were rather applied so that the participants got used to the rating scales.

In the experimental phase, the participants experienced two blocks of five experimental trials each. In each experimental trial, a lane change served as take-over situation with varying TTCs at the moment of the take-over request (TOR). In one block, participants rated the criticality on the SCA, in the other block, on the CRS. The sequence of blocks was balanced

across participants, i.e. 13 participants started rating subjective criticality on the SCA, 12 participants on the CRS. Similar to the training trials, the simulated vehicle started in automated mode executing longitudinal and lateral control. The automation was designed to accelerate to 100 km/h, keep the speed for the course of the trial, and drive at the center of the right lane. In all trials, the participant's vehicle followed a lead vehicle, a blue coach (see figure 3), and maintained a constant distance of 1.8 s time headway, i.e. the time it will take the participant's vehicle to reach the position of the lead vehicle.

After about 1 min in each trial, the take-over situation took place. In this situation, the participants' lane was blocked due to a broken vehicle. As soon as the lead vehicle changed lanes to avoid a collision, the obstacle became visible to the participant. The automation was able to detect the obstacle and requested the driver to take over by an acoustic cue. The same acoustic cue as in the training phase was used. The timing of the lane change maneuver of the lead vehicle, hence the timing of the TOR, was varied within-subjects. This resulted in different TTCs concerning the obstacle at the moment of the TOR. Five equidistant TTC-values were realized: 2.5, 3.0, 3.5, 4.0, 4.5 s. They were presented in a balanced order across participants. This means, over all participants, each TTC-value was presented at each position five times. However, the sequence of TTC-values between the two blocks was held constant for each participant to avoid any sequence effects on criticality ratings. The interval of 0.5 s between the TTC-values was chosen because a pretest showed that 0.5 s was large enough to cover a certain range of TTCs without having too many trials. Participants were instructed to take back control as fast and safely as possible by steering or braking upon the TOR. After steering around the obstacle or braking to a complete stop in front of the obstacle, the simulation was switched off and the simulation was stopped. On the iPad, participants rated the subjective criticality and the perceived effort. An instruction trial was added before each block to avoid surprise effects on the criticality ratings

(see figure 5). For the instruction trials, a TTC from the medium spectrum was used (3.63 s). This resulted in one instruction and five experimental trials per block, hence, participants experienced two instruction trails and ten experimental trials in total (see figure 5).

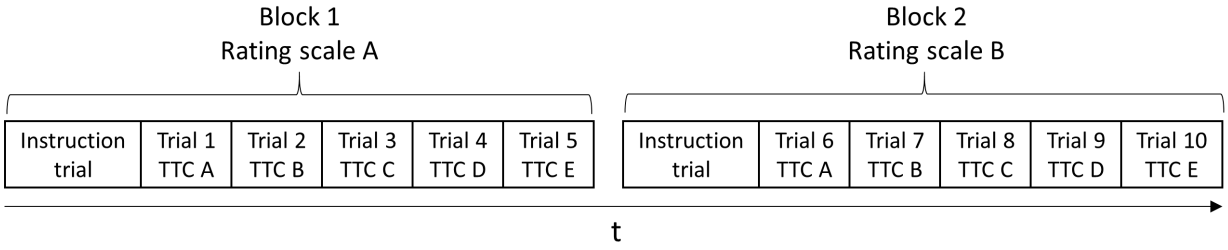


Figure 5: Experimental course. The sequence of TTC-values (A-E) was balanced between-subjects and held constant between blocks (1 and 2). In the first block, criticality ratings were collected on rating scale A, in the second block, on rating scale B. The order of rating scales was balanced between-subjects.

In a final interview, participants were asked for their personal preference regarding the two rating scales. Participants were debriefed of the scope of the experiment. Overall, the experiment lasted about 90 minutes.

2.4 Dependent variables

The subjective criticality rating was the main dependent variable assessed with the SCA [0-10 pts.] or CRS [1-100 pts.] at the end of each trial. Besides, the effort-subscale of the NASA-TLX [0-100 pts.], take-over time [ms], maximal steering wheel position [%], and maximal brake pedal position [%] for each trial served as dependent variables. Take-over time was measured between the onset of the acoustic cue and the driver response in terms of the steering wheel or brake pedal input. For maximal steering wheel position and maximal brake pedal position, the highest values during the take-over were extracted per participant and trial.

2.5 Data Analysis

For the analysis, R version 3.6.1 (R Core Team, 2019) was used. The correlation between TTC and the criticality ratings was calculated based on a method proposed by Bland and Altman (1995). This method accounts for repeated observations as in the present study. It is implemented in the R-package 'rmcorr' (Bakdash & Marusich, 2018). Degrees of freedom were calculated with Bakdash and Marusich's method available in rmcorr (2018). In accordance to Hemphill (2003), a correlation coefficient $r < .21$ indicates a weak correlation, between .21 and .33 a medium correlation, and $r > .33$ a strong correlation. The correlation coefficients were used to answer the question of whether the two scales are valid tools for the assessment of criticality of take-over situations (research questions 1 and 2). Besides, it was tested whether the ratings of each TTC-value differed from the remaining ratings. Since we had paired samples and did not expect a normal distribution of the data, the Friedmann-test is used (Friedman, 1937). In case a significant difference was found, a post-hoc analysis was calculated using the Nemeyi-test (Nemenyi, 1962) of the R-package 'PMCMR' (Pohlert, 2014). The Bonferroni-method was used to adjust p-values.

To compare the two rating scales against each other (research question 3), it was tested whether the correlation coefficients differed significantly based on a method suggested by Eid, Gollwitzer, and Schmitt (2017). The method determines a z-value of the two fisher-Z transformed correlation coefficients that can be tested on significance.

For the analysis of the change of ratings and behavior over trials (research question 4), mixed-effects models for each dependent variable were calculated with the 'lme4'-package (Bates et al., 2015). The independent variable 'trial' served as a linear predictor. We accounted for inter-individual differences mentioned in the introduction and for the repeated measurement

by adding a random intercept for each participant. Degrees of freedom were estimated with Satterthwaite's method available in the 'lmerTest'-package (Kuznetsova et al., 2017). The goodness-of-fit of each model is characterized by the marginal and conditional coefficient of determination (R^2, Nakagawa & Schielzeth, 2013).

3 Results

Since participants could rate subjective criticality either on the SCA or CRS, 125 trials (25 participants x 5 trials) were available for the analysis of the SCA- and 125 trials for the CRS-ratings. For the analysis of the perceived effort ratings and the behavioral data, data from ten trials per participant were available, resulting in 250 trials.

3.1 Correlation of the Scale of Criticality Assessment of driving situations with time-to-collision (RQ 1)

Figure 6 visualizes the mean SCA-ratings for all five TTC-values. The mean SCA-rating across all TTC-values was 4.62 pts. (*SD* = 1.9). This mean value corresponds to the verbal category 'unpleasant' and is located at the threshold between tolerable and intolerable situations defined by Neukum et al. (2008). The ratings decrease with increasing TTC-values. The minimum ('imperceptible', 0 pts.) and maximum ('uncontrollable', 10 pts.) were not chosen by any participant.

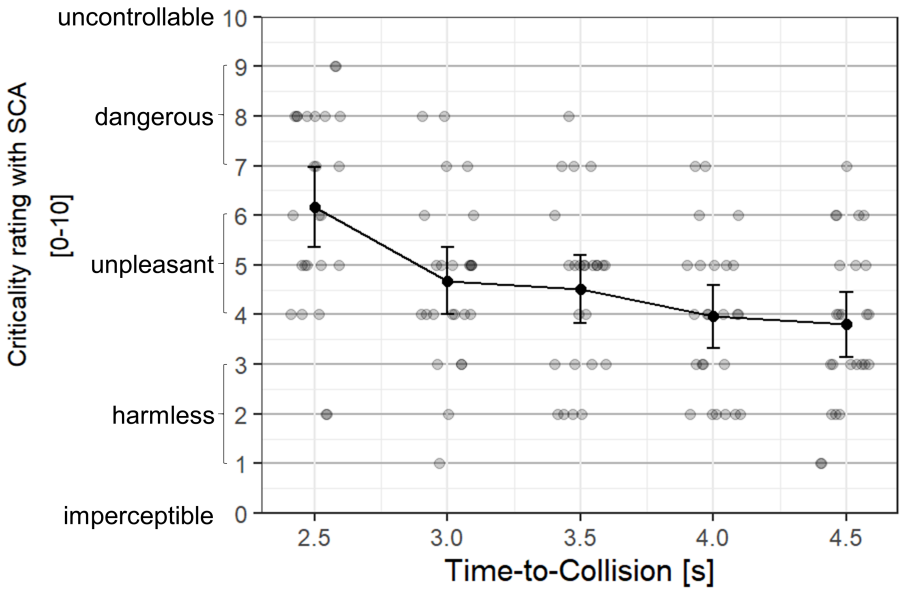


Figure 6: Means, standard errors, and raw values of the Scale of Criticality Assessment of driving situations (SCA) for each time-to-collision-value from 2.5 to 4.5 s.

The SCA-ratings correlated significantly with the TTC-values ($r_{TTC\_SCA}(99) = -.59$, $p < .001$). Based on Hemphill (2003), the magnitude of the correlation coefficient indicates a strong correlation. The Friedman-test revealed that the SCA-ratings from at least two TTC-values differed significantly from each other ($\chi^2(4) = 35.36$, $p < .001$). The post-hoc test showed a significant difference between the most critical TTC-value (2.5 s) and the two least critical ones (4.0 s resp. 4.5 s, see the adjusted p-values for all comparisons in table 1).

Time-to-collision	2.5 s	3.0 s	3.5 s	4.0 s
3.0 s	.715			
3.5 s	.229	1		
4.0 s	< .001 ***	1	1	
4.5 s	< .001 ***	1	1	1

Table 1: Friedman-test results for the SCA-ratings. Adjusted p-values for post-hoc comparisons of SCA-ratings between all TTC-values. Significance symbols: * $p < .05$, ** $p < .01$, *** $p < .001$

3.2 Correlation of the Criticality Rating Scale with time-to-collision (RQ 2)

The mean CRS-rating across all TTC-values was 45.98 pts. (SD = 25.2). Figure 7 visualizes the means, standard errors, and raw values of the CRS. As expected, with increasing TTC, the CRS-ratings decreased (see figure 7). The minimum value ('not critical at all', 1 pt.) was selected 16 times, while the maximum ('very critical', 100 pts.) was never selected.

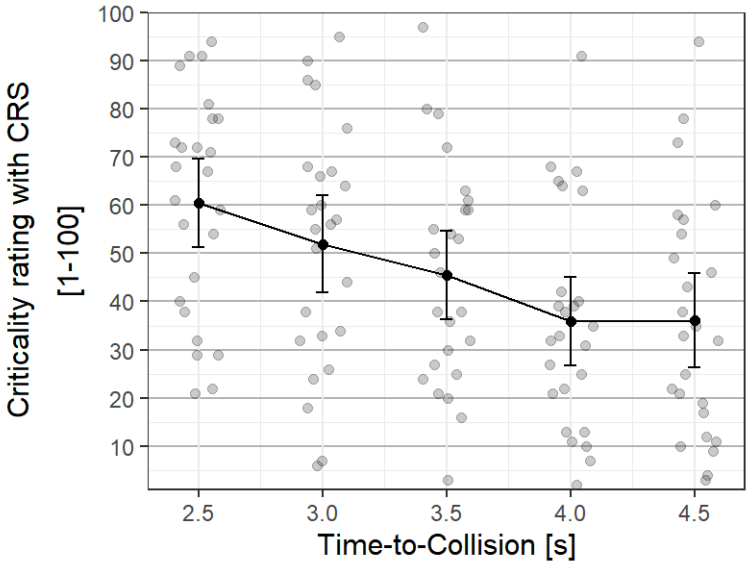


Figure 7: Means, standard errors, and raw values of the Criticality Rating Scale (CRS) for each time-to-collision-value from 2.5 to 4.5 s.

The correlation between CRS-ratings and TTC-values was highly significant ($r_{TTC\_CRS}(99) = -.66$, $p < .001$). Again, this represents a strong correlation. The Friedman-test revealed that at least two CRS-ratings differed significantly from each other ($\chi^2(4) = 46.53$, $p < .001$). The post-hoc comparisons showed that in four cases the CRS-ratings differed from each other (see

adjusted p-values in table 2). These significant differences were between the most critical TTC-value (2.5 s) and the two less critical ones (4.0 s and 4.5 s), similar to the SCA-ratings. Also, the CRS-ratings of the TTC-value 3.0 s differed significantly from 4.0 s and 4.5 s.

Time-to-collision	2.5 s	3.0 s	3.5 s	4.0 s
3.0 s	1			
3.5 s	.229	1		
4.0 s	< .001 ***	.009 **	.636	
4.5 s	< .001 ***	.014 *	1	1

Table 2: Friedman-test results for the CRS-ratings. Adjusted p-values for post-hoc comparisons of CRS-ratings between all TTC-values. Significance symbols: * $p < .05$, ** $p < .01$, *** $p < .001$

3.3 Comparison of both scales (RQ 3)

For the comparison, the two correlation coefficients of the SCA and the CRS were used. The method suggested by Eid et al. (2017) revealed that the coefficients of the scales did not differ significantly ($z = 0.52$, $p = .603$).

The final interview showed that 84 % of the participants ($N = 21$) preferred the SCA for assessing subjective criticality of a take-over situation. 42.3 % of the participants ($N = 9$) reasoned their voting with the subdivision of the SCA into verbal and numerical categories. 38.1 % of the participants ($N = 8$) preferred the SCA due to the better description of the take-over situation by the verbal categories. Two of the participants preferring the CRS stated that the labeling of the poles were better suited for the take-over situation.

3.4 Repeated experience of take-over situations (RQ 4)

Mixed-effect models were calculated to investigate the change of ratings and take-over behavior over the repeated experience of the experimental trials. Significant estimates of the factor 'trial' would indicate a change of ratings or take-over behavior over the repeated experience. The statistical results are presented in table 3 and the mean-values and standard errors per trial are plotted in figure 8. Only the maximal brake pedal position decreased significantly over trials. Descriptively, the SCA-, CRS-, and perceived effort ratings decreased slightly over the trials (see negative estimates for trial in table 3 and figure 8). However, the models showed that none of these changes reached significance (all $t < 2$, $p > .05$). Hence, the ratings, take-over times, and steering wheel positions were not affected by the repeated experience of take-over situations. For all models, the marginal coefficient of determination was very small, hence, the variance explained by the fixed factor 'trial' was very low (below 1 %).

Criticality rating on SCA [0-10]	Estimate	Std. Error	df	t-value	p-value
Intercept	4.98	0.39	79.65	12.86	< .001 ***
Trial	-0.12	0.09	100	-1.32	0.191
Variance explained: $R^2_{marginal}$ = 0.8 %, $R^2_{conditional}$ = 41.9 %					N_{trials} = 125
Criticality rating on CRS [1-100]	Estimate	Std. Error	df	t-value	p-value
Intercept	49.02	5.09	51.97	9.62	< .001 ***
Trial	-1.02	0.97	100	-1.05	.297
Variance explained: $R^2_{marginal}$ = 0.3 %, $R^2_{conditional}$ = 62.6 %					N_{trials} = 125
Perceived effort rating [0-100]	Estimate	Std. Error	df	t-value	p-value
Intercept	41.41	4.05	48.88	10.21	< .001 ***
Trial	-0.41	0.40	225	-1.04	.299
Variance explained: $R^2_{marginal}$ = 0.2 %, $R^2_{conditional}$ = 44.3 %					N_{trials} = 250

Take-over time [ms]	Estimate	Std. Error	df	t-value	p-value
Intercept	764.27	32.34	43.13	23.63	< .001 ***
Trial	-3.37	2.90	225	-1.16	.246
Variance explained: $R^2_{marstginal}$ = 0.3 %, $R^2_{conditional}$ = 51.2 %					N_{trials} = 250
Maximal steering wheel position [%]	Estimate	Std. Error	df	t-value	p-value
Intercept	0.94	0.06	146.07	16.46	< .001 ***
Trial	0.00	0.01	225	0.24	.809
Variance explained: $R^2_{marginal}$ = 0.0 %, $R^2_{conditional}$ = 9.1 %					N_{trials} = 250
Maximal brake pedal position [%]	Estimate	Std. Error	df	t-value	p-value
Intercept	57.43	4.88	39.79	11.76	< .001 ***
Trial	-0.81	0.41	225	-1.99	.048 *
Variance explained: $R^2_{marginal}$ = 0.7 %, $R^2_{conditional}$ = 56.4 %					N_{trials} = 250

Table 3: Summary of statistics for the repeated experience of trials of all dependent variables. Significance symbols: * $p < .05$, ** $p < .01$, *** $p < .001$.

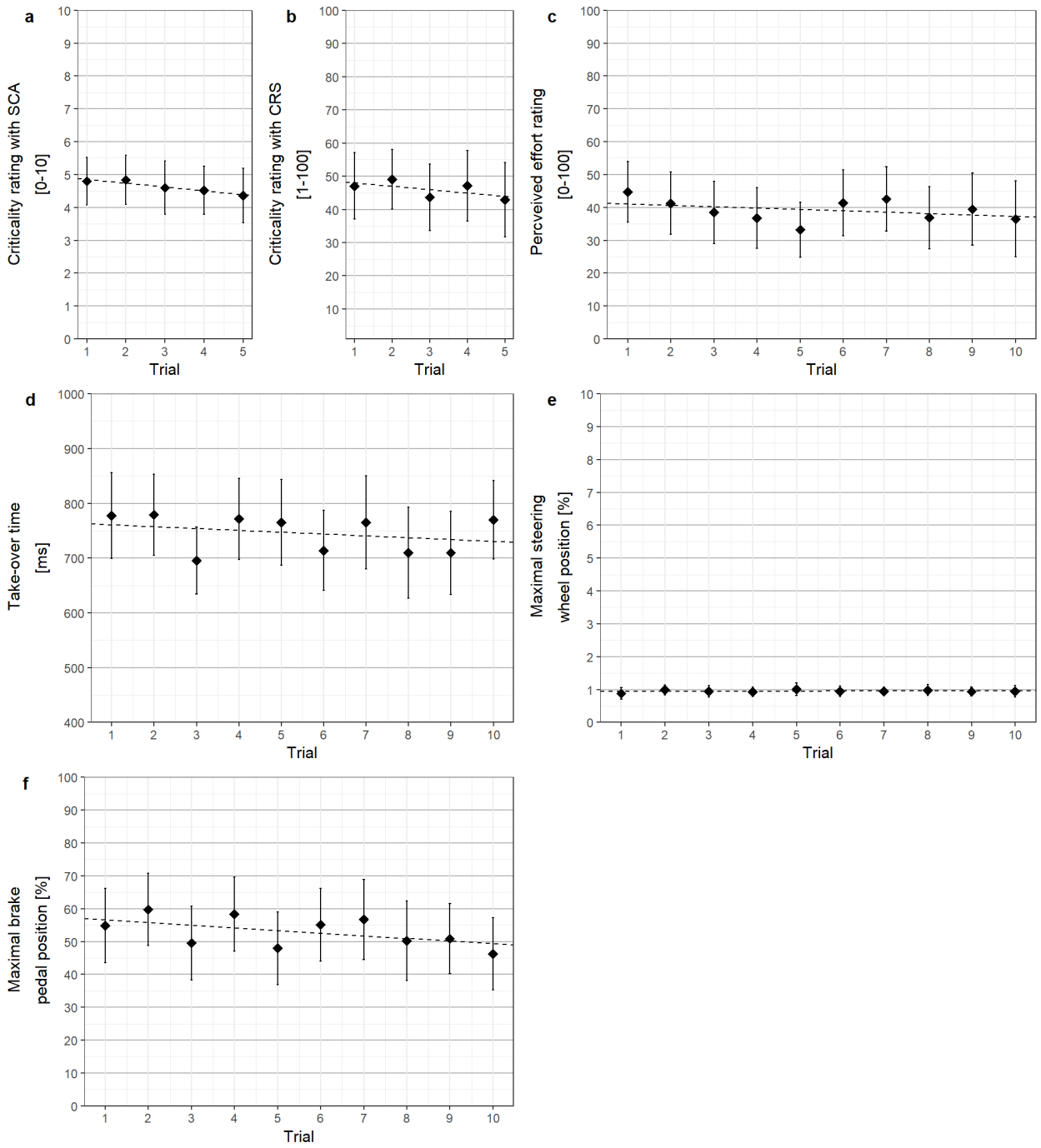


Figure 8: Means, standard errors, and regression lines of the dependent variables per trial over the repeated experience of take-over situations.

Note. Five trials were evaluated on the SCA, five with the CRS. Concerning the remaining dependent variables, values are available for all ten experimental trials.

4 Discussion

The present study investigated whether two rating scales, the Scale of Criticality Assessment of driving situations (SCA) and the Criticality Rating Scale (CRS), are valid tools for the assessment of subjective criticality of take-over situations (RQ 1 and 2) and whether one is superior to the other one (RQ 3). Besides, the effects of the repeated experience of take-over situations on ratings and take-over behavior (RQ 4) were investigated. Participants experienced five experimental take-over situations twice that differed regarding time-to-collision. They provided their criticality rating either on the SCA or the CRS. Perceived effort ratings and take-over behavior were recorded in the ten experimental trials.

Before discussing the research questions, it should be noted that the take-over times were very small. This could be due to several reasons. First and in contrast to most other studies, our participants did not perform a non-driving related task. Hence, they could focus on the driving situations. Second, it could be that participants were highly trained to take over very fast after the training trials. Third, we assume that they were highly alert to expect a take-over by its frequent occurrence. Forth, the time budget used in this study was smaller than in most other studies on take-over time (2.5 s – 4.5 s in our study vs. 5 s and 7 s in Gold et al. (2013) or 8.6 s in Roche et al. (2018)). Previous research showed that shorter time budgets lead to shorter take-over times.

4.1 Research question 1 and 2: Are the Scale of Criticality Assessment of driving situations and the Criticality Rating Scale valid tools for the assessment of subjective criticality in take-over situations?

The study showed that both scales correlate strongly with the TTC-values that were varied to manipulate objective criticality of the take-over situations. This indicates that the SCA and the

CRS are valid tools to assess the subjective criticality. The study paved the way of validating criticality rating scales in driving studies. However, convergent validity was tested, while different types such as discriminant or criterion validity were not investigated. This should be addressed in future studies. Furthermore, as mentioned in the introduction, validity is continuous and cut-off values for correlations for validity testing do not exist. Hence, one could argue that higher correlations are requested to infer validity. Besides, the conclusion is limited to a lane change and take-over situations in which objective criticality is varied by TTC. It is questionable whether the correlations between objective criticality and criticality ratings would be equally high in other maneuvers or when objective criticality is varied by different variables. For example, it could be that increasing traffic density from low to medium traffic would have a different effect on criticality ratings than an increase from medium to high traffic. Also, the rating scales are only validated for time-to-collisions in take-over situations. A transferability to driving situations in general is not given. Finally, it should be noted that the scales measure a general perception of criticality of take-over situations. Specific aspects such as collision risk or vehicle stability cannot be extracted. For this purpose, more comprehensive questionnaires would be necessary. Hence, future studies should validate the rating scales in different maneuvers and with other situational parameters, e.g. traffic density, to manipulate objective criticality and test different types of validity.

4.2 Research question 3: Do both scales differ regarding their validity?

Even though the two rating scales use different scale designs, the comparison of the correlation coefficients demonstrated that they do not differ. Hence, the two scales are equally well suited for the assessment of subjective criticality in this specific take-over situation with this manipulation of objective criticality.

The results are noteworthy. On the one hand, the SCA is more time-consuming regarding instruction and processing than the CRS. More effort has to be put in explaining this scale since it is an unusual design. When processing the SCA, it has to be checked whether the rating of the first step (verbal category) corresponds to the rating of the second step (numerical subcategory). Besides, the correlation coefficient of the CRS was slightly higher and the CRS could better discriminate between different TTC-values as indicated by the higher amount of significant differences of the post-hoc comparisons. On the other hand, more participants preferred the SCA when rating subjective criticality of a take-over situation. Furthermore, as stated by Neukum and Krüger (2003), an advantage of the SCA is the threshold between tolerable and intolerable situations that is supposed to make ratings more comparable between raters. However, this reason was not yet proven. These aspects should make an impact on the researchers' decision on which scale to use in the future.

Apart from the research question, two additional insights concerning the two rating scales should be mentioned: First, the criticality ratings of the SCA and the CRS showed that differences of objective criticality are rather experienced with the more critical TTCs than with the less critical TTCs (see figures 6 and 7). This is in line with Siebert et al. (2014), who found rating differences between more critical THWs and no differences between less critical THWs to a lead vehicle. Siebert et al. (2014) interpreted this result as a threshold effect for the relation between objective criticality and subjective variables. While they used a car-following scenario, we likely observed the same effect in a different driving situation.

Second, participants neither used the minimum category of the SCA ('imperceptible') nor the maximum categories of the SCA ('uncontrollable') or the CRS ('very critical'). It seems as the lane change could not be ignored because no participant selected the minimum category.

Concerning the maximum categories, the impression arises as none of the TTC-values was small enough to not be coped with because none of them was rated as maximal critical. It could be that the realization of the take-over situation did not achieve to cover a wide range of objective criticality. Hence, future studies may seek to cover a broader range of criticality.

4.3 Research question 4: Do drivers' criticality and effort ratings and take-over behavior change over the repeated experience of take-over situations?

Neither the ratings nor the take-over behavior changed over the repeated experience of the ten take-over situations, except maximal brake pedal position (RQ 4). This might be due to two reasons. First, the behavior and subjective experience likely changed within the five training trials. Hence, participants were already highly trained and habituated when the experiment started. Second, it could also be that subjective experience and behavior change within the first experimental trials and does not change in the following. Forster et al. (2019) found stabilized reaction times after three trials for transitions between SAE-level 2 and 3. Our analysis across all ten trials might have overruled a potential effect. In the present study, participants experienced twelve take-over situations, while participants of other studies experienced fewer situations, for example two in Hergeth et al. (2016) or six in Roche et al. (2018).

The criticality ratings of the take-over situations seem to be robust to a certain extent with the exception that the respondents possibly were already habituated to the take-over situations. This is a promising finding, as it suggests that even after the repeated experience of a take-over situation, the criticality ratings on the SCA and the CRS are still valid and comparable to the first rating.

Regarding the perceived effort ratings, the results showed that participants did not experience increased or decreased effort over trials, even though, the setting was monotonous with ten similar experimental and two instruction trials. Based on de Waard (2002), increasing fatigue due to the monotonous experimental setting may become apparent by increasing effort to cope with the situation. A reason why no change of perceived effort over trials was observed is that participants' increasing practice had compensated for the increasing passive fatigue. In consequence, participants might not have experienced increasing effort.

In line with Brandenburg and Roche (2020), we neither observed a change of take-over times nor of steering wheel positions over trials. The missing effects might be due to three reasons. First, similar to the perceived effort ratings, participants' practice likely increased due to the repeated experience of the ten take-over situations. This would allow drivers to anticipate future states and, usually, improve their performance (Endsley, 1995). Passive fatigue possibly increased at the same time. Hence, increasing practice and increasing fatigue might have compensated each other and led to a constant level of behavior. Second, as indicated earlier, take-over times and steering behavior might have changed within the training (and first experimental) trials and stabilized in the following. Such way, a significant change during the experimental trials was not detectable. Third, it could be that we observed a floor effect concerning take-over times because they were very small, making a faster reaction nearly impossible. Similarly, Brandenburg and Roche (2020) argued that a reason for the missing effect of repeated experience on take-over times might be a floor effect due to the very fast take-overs.

A decrease of brake pedal position was observed but no other behavioral change over the repeated experience. In contrast to our results, Brandenburg and Roche (2020) showed an

increase in deceleration. It seems as brake behavior does not adapt as fast as other behavioral or subjective measures.

To conclude, these results indicate that studies with repeated experience of take-over situations are relatively valid as only brake behavior changed with increasing practice. However, it could be that subjective experience and behavior already adopted within the training or first experimental trials. Besides, it should be noted that the marginal coefficients of determination of all mixed-effect models were very small while the conditional coefficients of determination were quite high. This means that the fixed factor 'trial' did not explain much variance but the random intercept 'participant' did. Hence, the ratings and take-over behavior were mainly affected by inter-individual differences to rate or react rather than by the repeated experience of take-over situations.

4.4 Limitations

The study has some limitations that should be kept in mind when interpreting the results. First, the investigated take-over scenario was limited to one scenario: a lane change due to an obstacle in the participant's lane. Hence, the two rating scales have only been validated for this scenario. Besides, objective criticality was varied by manipulating TTC at the moment of the take-over request. Other characteristics of a take-over situation may also affect objective criticality. It should be tested whether similar correlations would be found if another take-over situation was used or if objective criticality would have been varied by different parameters.

Second, it must be noted that the driving simulator was mid-fidelity. The degree of immersion of the presented scenarios may be low compared to a high-fidelity simulator and the effect on perception and take-over behavior might differ from the one in real traffic. However,

driving simulators allow low-cost and low-risk experiments in a controlled environment, especially for preliminary research (van Nes et al., 2010).

Third, due to the restrictions of the driving simulator, the lowest feasible TTC was 2.5 s and, due to the experimental design, the highest TTC was 4.5 s. However, the ratings show that almost the whole ranges of both scales were used, except the maximum and minimum categories. Future studies may aim at covering the whole range of the scales by presenting more and less objectively critical driving situations.

Forth, participants experienced many take-over situations in a row. This is an unrealistically high occurrence. Future studies should have a lower portion of take-over situations per session or more filler trials.

Fifth, the NASA TLX was used to assess the development of passive fatigue over the course of the experiment. Rating scales on fatigue, e.g. Karolinska sleepiness scale (Shahid et al., 2011), would have been more appropriate.

Sixth, with a mean age of 27 years, the participants of the present study can be assigned to the younger population. Potential effects that come along with aging are impaired information processing (Salthouse, 1991) and increased hazard perception times (Horswill et al., 2008) which may result in slower take-overs. However, Körber et al. (2016) observed that take-over times did not differ between younger (≤ 28 years) and older drivers (≥ 60 years). But the older participants showed different take-over behavior than the younger ones, i.e. more and stronger braking and higher TTCs (Körber et al., 2016). Hence, it may be assumed that the study observed the best possible take-over behavior because mainly young drivers participated. Older participants might have shown different take-over behavior, i.e. larger take-over times, stronger braking, stronger steering.

4.5 Conclusion

The Scale of Criticality Assessment of driving situations (SCA) and the Criticality Rating Scale (CRS) are equally valid tools for the assessment of the subjective criticality of take-over situations. The ratings are robust over time. However, it should be noted that the two scales were only tested on convergent validity in this specific take-over situation of a lane change with this specific variation of objective criticality. Validation tests in other take-over situations and with different variations of objective criticality are pending. Besides, different types of validity should be investigated.

A behavioral change over the repeated experience of experimental take-over situations was only observed regarding braking. Possibly, subjective experience and take-over behavior adopted within the training trials, hence a change was not quantifiable. Effort ratings, take-over times, and steering wheel positions did not change.

5 Acknowledgements

We thank Oliver Blum for the cooperation regarding study planning and his careful data collection. We also thank Mario Lasch for his technical support and Stefan Brandenburg for proof reading.

6 References

Bakdash, J. Z., & Marusich, L. R. (2018). *rmcorr: Repeated Measures Correlation*. https://CRAN.R-project.org/package=rmcorr

Banet, A., & Bellet, T. (2008). Risk awareness and criticality assessment of driving situations: A comparative study between motorcyclists and car drivers. *IET Intelligent Transport Systems*, *2*(4), 241. https://doi.org/10.1049/iet-its:20080037

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bellem, H., Klüver, M., Schrauf, M., Schöner, H.-P., Hecht, H., & Krems, J. F. (2017). Can We Study Autonomous Driving Comfort in Moving-Base Driving Simulators? A Validation Study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(3), 442–456. https://doi.org/10.1177/0018720816682647

Bland, J. M., & Altman, D. G. (1995). Statistics notes: Calculating correlation coefficients with repeated observations: Part 1—correlation within subjects. *BMJ*, *310*(6977), 446. https://doi.org/10.1136/bmj.310.6977.446

Bogner, K., & Landrock, U. (2016). Response Biases in Standardised SurveysResponse Biases in Standardised Surveys. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_016

Brandenburg, S., & Roche, F. (2020). Behavioral Changes to Repeated Takeovers in Automated Driving: The Drivers' Ability to Transfer Knowledge and the Effects of Takeover Request Process. *Transportation Research Part F: Psychology and Behaviour*. https://doi.org/10.1016/j.trf.2020.06.002

Cooper, G. E., & Harper, P. (1969). *The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities*. 60.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Damböck, D., Farid, M., Tönert, L., & Bengler, K. (2012). Übernahmezeiten beim hochautomatisierten Fahren. *Tagung Fahrerassistenz*, 16–28.

de Waard, D. (2002). Mental Workload. In *Human Factors for Highway Engineers* (1st ed., pp. 161–175). Bingley. https://trid.trb.org/view/708725

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5.). Beltz.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64.

Feldhütter, A., Kroll, D., & Bengler, K. (2018). Wake Up and Take Over! The Effect of Fatigue on the Take-over Performance in Conditionally Automated Driving. *Proceedings of 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2080–2085. https://doi.org/10.1109/ITSC.2018.8569545

Forster, Y., Hergeth, S., Naujoks, F., Beggiato, M., Krems, J. F., & Keinath, A. (2019). Learning to use automation: Behavioral changes in interaction with automated driving systems. *Transportation Research Part F: Traffic Psychology and Behaviour*, *62*, 599–614. https://doi.org/10.1016/j.trf.2019.02.013

Frey, B. B. (2018). Standards for Educational and Psychological Testing. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139.n662

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Fuller, R. (2011). Driver Control Theory. From Task Difficulty Homeostasis to Risk Allostasis. In *Handbook of Traffic Psychology*. Academic Press. https://doi.org/10.1016/B978-0-12-381984-0.10002-5

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*, 1938–1942. https://doi.org/10.1177/1541931213571433

699 Gold, C., Körber, M., Lechner, D., & Bengler, K. (2016). Taking over control from highly
700 automated vehicles in complex traffic situations: The role of traffic density. *Human*
701 *Factors*, *58*(4), 642–652. https://doi.org/10.1177/0018720816634226
702 Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Th ANNUAL*
703 *MEETING*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909
704 Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results
705 of Empirical and Theoretical Research. *Advances in Psychology*, *52*, 139–183.
706 https://doi.org/10.1016/S0166-4115(08)62386-9
707 Hartig, J., Frey, A., & Jude, N. (2008). Validität. In H. Moosbrugger & A. Kelava (Eds.),
708 *Testtheorie und Fragebogenkonstruktion*. Springer.
709 Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American*
710 *Psychologist*, *58*(1), 78–79. https://doi.org/10.1037/0003-066X.58.1.78
711 Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior Familiarization With Takeover Requests
712 Affects Drivers' Takeover Performance and Automation Trust. *Human Factors: The*
713 *Journal of the Human Factors and Ergonomics Society*, *59*(3), 457–470.
714 https://doi.org/10.1177/0018720816678714
715 Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep Your Scanners Peeled: Gaze
716 Behavior as a Measure of Automation Trust During Highly Automated Driving. *Human*
717 *Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519.
718 https://doi.org/10.1177/0018720815625744
719 Horswill, M. S., Marrington, S. A., McCullough, C. M., Wood, J., Pachana, N. A., McWilliam,
720 J., & Raikos, M. K. (2008). The Hazard Perception Ability of Older Drivers. *The Journals*
721 *of Gerontology Series B: Psychological Sciences and Social Sciences*, *63*(4), P212–P218.
722 https://doi.org/10.1093/geronb/63.4.P212

723 Jamson, A. H., Merat, N., Carsten, O. M. J., & Lai, F. C. H. (2013). Behavioural changes in
724 drivers experiencing highly-automated vehicle control in varying traffic conditions.
725 *Transportation Research Part C: Emerging Technologies*, *30*, 116–125.
726 https://doi.org/10.1016/j.trc.2013.02.008
727 Jarosch, O., & Bengler, K. (2018). Rating of Take-Over Performance in Conditionally
728 Automated Driving Using an Expert-Rating System. In N. Stanton (Ed.), *Advances in*
729 *Human Aspects of Transportation* (Vol. 786, pp. 283–294). Springer International
730 Publishing. https://doi.org/10.1007/978-3-319-93885-1_26
731 Junietz, P., Schneider, J., & Winner, H. (2017). *Metrik zur Bewertung der Kritikalität von*
732 *Verkehrssituationen und -szenarien*. 12.
733 Körber, M., Gold, C., Lechner, D., & Bengler, K. (2016). The influence of age on the take-over
734 of vehicle control in highly automated driving. *Transportation Research Part F: Traffic*
735 *Psychology and Behaviour*, *39*, 19–32. https://doi.org/10.1016/j.trf.2016.03.002
736 Kuznetsova, A., Bruun Brockhoff, P., & Haub Bojesen Christensen, R. (2017). *lmerTest: Tests in*
737 *Linear Mixed Effects Models*. https://CRAN.R-project.org/package=lmerTest
738 Mesken, J., Hagenzieker, M. P., Rothengatter, T., & de Waard, D. (2007). Frequency,
739 determinants, and consequences of different drivers' emotions: An on-the-road study
740 using self-reports, (observed) behaviour, and physiology. *Transportation Research Part*
741 *F: Traffic Psychology and Behaviour*, *10*(6), 458–475.
742 https://doi.org/10.1016/j.trf.2007.05.001
743 Mok, B. K.-J., Johns, M., Lee, K. J., Ive, H. P., Miller, D., & Ju, W. (2015). Timing of
744 unstructured transitions of control in automated driving. *Intelligent Vehicles Symposium*
745 *(IV)*, 1167–1172. https://doi.org/10.1109/IVS.2015.7225841

Mok, B. K.-J., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015). Emergency, Automation Off: Unstructured Transition Timing for Distracted Drivers of Automated Vehicles. *Proceedings of the IEEE 18th Intelligent Transportation Systems (ITSC)*, 2458–2464. https://doi.org/10.1109/ITSC.2015.396

Moosbrugger, & Kelava, A. (Eds.). (2020). *Testtheorie und Fragebogenkonstruktion*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4

Murata, A., Kanbayashi, M., & Hayami, T. (2013). Effectiveness of Automotive Warning System Presented with Multiple Sensory Modalities. In V. G. Duffy (Ed.), *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management. Healthcare and Safety of the Environment and Transport* (Vol. 8025, pp. 88–97). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39173-6_11

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Naujoks, F., Purucker, C., Wiedemann, K., Neukum, A., Wolter, S., & Steiger, R. (2017). Driving performance at lateral system limits during partially automated driving. *Accident Analysis & Prevention*, *108*, 147–162. https://doi.org/10.1016/j.aap.2017.08.027

Nemenyi, P. (1962). Distribution-free Multiple Comparisons. *Journal of the International Biometric Society*, *18*(2), 263.

Neukum, A., & Krüger, H.-P. (2003). Fahrerreaktionen bei Lenksystemstörungen – Untersuchungsmethodik und Bewertungskriterien. *Reifen - Fahrwerk - Fahrbahn.*, *1791*, 297–318.

768 Neukum, A., Lübbeke, T., Krüger, H. P., Mayser, C., & Steinle, J. (2008). ACC-Stop&Go:
769 Fahrerverhalten an funktionalen Systemgrenzen. *5. Workshop Fahrerassistenzsysteme-*
770 *FAS*, 141–150.
771 Payre, W., Cestac, J., & Delhomme, P. (2016). Fully Automated Driving: Impact of Trust and
772 Practice on Manual Control Recovery. *Human Factors: The Journal of the Human*
773 *Factors and Ergonomics Society*, *58*(2), 229–241.
774 https://doi.org/10.1177/0018720815612319
775 Pohlert, T. (2014). *The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)* [R].
776 https://CRAN.R-project.org/package=PMCMR
777 Politis, I., Brewster, S., & Pollick, F. (2014). Evaluating multimodal driver displays under
778 varying situational urgency. *Proceedings of the SIGCHI Conference on Human Factors in*
779 *Computing Systems*, 4067–4076. https://doi.org/10.1145/2556288.2556988
780 R Core Team. (2019). *R: A language and environment for statistical computing* (3.6.1)
781 [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
782 Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How Traffic Situations and
783 Non-Driving Related Tasks Affect the Take-Over Quality in Highly Automated Driving.
784 *Proceedings of the 58th Human Factors and Ergonomics Society Annual Meeting*, *58*,
785 2063–2067. https://doi.org/10.1177/1541931214581434
786 Radlmayr, J., Ratter, M., Feldhütter, A., Körber, M., Prasch, L., Schmidtler, J., Yang, Y., &
787 Bengler, K. (2018). Take-Overs in Level 3 Automated Driving – Proposal of the Take-
788 Over Performance Score (TOPS). In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander,
789 & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics*
790 *Association (IEA 2018)* (Vol. 823, pp. 436–446). Springer International Publishing.
791 https://doi.org/10.1007/978-3-319-96074-6_46

Roche, F., & Brandenburg, S. (2020). Should the urgency of visual-tactile takeover requests match the criticality of takeover situations? *IEEE Transactions on Intelligent Vehicles*, *5*(2), 306–313. https://doi.org/10.1109/TIV.2019.2955906

Roche, F., & Brandenburg, S. (2018). Should the urgency of an auditory-tactile takeover request match the situational criticality? *Proceedings 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 1035–1040. https://doi.org/10.1109/ITSC.2018.8569650

Roche, F., Somieski, A., & Brandenburg, S. (2018). Behavioral Changes to Repeated Takeovers in Highly Automated Driving: Effects of the Takeover Request-Design and the Non-Driving Related Task-Modality. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(5), 839–849. https://doi.org/10.1177/0018720818814963

Roche, F., Thüring, M., & Trukenbrod, A. K. (2020). What happens when drivers of a highly-automated vehicle take over control in critical brake situations? *Accident Analysis & Prevention*, *144*. https://doi.org/10.1016/j.aap.2020.105588

Rodemerk, C., Habenicht, S., Weitzel, A., Winner, H., & Schmitt, T. (2012). Development of a general criticality criterion for the risk estimation of driving situations and its application to a maneuver-based lane change assistance system. *2012 IEEE Intelligent Vehicles Symposium*, 264–269. https://doi.org/10.1109/IVS.2012.6232129

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift Für Sozialpsychologie*, *9*, 222–245.

SAE International. (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (J3016)*. https://www.sae.org/standards/content/j3016_201806/

Salthouse, T. A. (1991). Mediation of Adult Age Differences in Cognition by Reductions in Working Memory and Speed of Processing. *Psychological Science*, *2*(3), 179–183. https://doi.org/10.1111/j.1467-9280.1991.tb00127.x

Sepehr, M. (1988). *NASA Task Load Index: Deutsche Version* (No. 284; pp. 1–16). Technische Universität Berlin. https://www.tib.eu/de/suchen/id/TIBKAT%3A019690789/NASA-Task-Load-Index-deutsche-Version/#

Shahid, A., Wilkinson, K., Marcu, S., & Shapiro, C. M. (2011). Karolinska Sleepiness Scale (KSS). In A. Shahid, K. Wilkinson, S. Marcu, & C. M. Shapiro (Eds.), *STOP, THAT and One Hundred Other Sleep Scales* (pp. 209–210). Springer New York. https://doi.org/10.1007/978-1-4419-9893-4_47

Siebert, F. W., Oehl, M., & Pfister, H.-R. (2014). The influence of time headway on subjective driver states in adaptive cruise control. *Transportation Research Part F: Traffic Psychology and Behaviour*, *25*, 65–73. https://doi.org/10.1016/j.trf.2014.05.005

van Nes, N., Brandenburg, S., & Twisk, D. (2010). Improving homogeneity by dynamic speed limit systems. *Accident Analysis & Prevention*, *42*(3), 944–952. https://doi.org/10.1016/j.aap.2009.05.002

Vogel, K. (2003). A comparison of headway and time to collision as safety indicators. *Accident Analysis & Prevention*, *35*(3), 427–433. https://doi.org/10.1016/S0001-4575(02)00022-2

Wainer, H., & Braun, H. I. (Eds.). (2013). *Test validity*. Routledge. https://books.google.de/books?hl=de&lr=&id=1i98Bl6EEZ0C&oi=fnd&pg=PP2&dq=wainer+braun+test+validity&ots=YnqaWKUnn8&sig=8kzSFoe81xaIVFBA83_L6KZPTYg#v=onepage&q=wainer%20braun%20test%20validity&f=false

Zhang, B., de Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation*

Research Part F: Traffic Psychology and Behaviour, *64*, 285–307. https://doi.org/10.1016/j.trf.2019.04.020