Driving Simulators in User-Centered Research: Validity, Stress Responses, and Implications for the Interpretation of Physiological Data

Dissertation

zur Erlangung eines Doktors der Wirtschaftswissenschaft

der Rechts- und Wirtschaftswissenschaftlichen

Fakultät der Universität Bayreuth

vorgelegt

von

Marcin Adam Czaban

aus

Hof

Dekan:	Herr Prof. Dr. Claas Christian Germelmann
Erstberichterstatter:	Herr Prof. Dr. Daniel Baier
Zweitberichterstatter:	Herr Prof. Dr. Joachim Riedl
Tag der mündlichen Prüfung:	14.10.2025

„Alles, was deine Hand zu tun findet, das tue mit deiner ganzen Kraft.“

Prediger 9,10

Danksagung

In Anlehnung an den von mir zitierten Bibelvers möchte ich nicht verheimlichen, dass sehr viel Kraft in diese Arbeit geflossen ist. Kraft, die ich neben meiner Motivation vor allem aus meiner Familie, meinen Freunden und meinem sozialen Umfeld geschöpft habe. Deshalb widme ich diese Seite den großartigen Menschen, ohne die es nahezu unmöglich gewesen wäre, diese Reise durchzustehen.

In erster Linie möchte ich mich bei Herrn Professor Daniel Baier bedanken, der es mir ermöglicht hat, eine externe Promotion an seinem Lehrstuhl durchzuführen und mir zu jeder Zeit mit Rat und Tat als Doktorvater zur Seite stand sowie mich stets inspiriert und unterstützt hat.

Im höchsten Maße danke ich meiner „akademischen Familie“ in Form von Professor Joachim Riedl und Professor Stefan Wengler. Danke, dass Ihr mich zu diesem Schritt ermutigt und es mir ermöglicht habt, diesen Weg mit Eurer Betreuung zu gehen. Danke, dass Ihr immer Zeit für mich hattet (auch für die ein oder andere Nachtschicht). Danke für Eure Offenheit und Ehrlichkeit und das daraus entstandene freundschaftlich-familiäre Verhältnis. Egal ob in Bezug auf die Dissertation oder private Themen: Der Austausch mit Euch hat mich immer vorangebracht und ich durfte viel lernen und mitnehmen. Ich könnte mich noch für unzählige andere Dinge bei Euch bedanken - nur leider wäre diese Seite dann ganz schnell überfüllt. Deshalb: DANKE.

Des Weiteren möchte ich mich bei meinen Eltern bedanken, die damals mit nichts nach Deutschland kamen, während ihres Lebens vor vielen Herausforderungen standen und auf vieles verzichtet haben, um meinen Geschwistern und mir ein besseres Leben zu ermöglichen und mir damit den Weg geebnet haben, der erste Akademiker in der Familie zu werden.

Darüber hinaus gilt mein größter Dank meiner Partnerin Sabine, die stetig für mich da ist, mir in sehr herausfordernden Zeiten permanent zur Seite stand und mir Mut zugesprochen hat. Danke, dass ich mich immer auf dich verlassen kann.

Weiterhin danke ich meinen engsten Vertrauten Tobias, Sebastian, Ulrich, Ahmad und Michael, die ein Teil meiner Familie geworden sind. Ich konnte mich immer auf euch verlassen und meine Sorgen mit euch teilen. Immer habt ihr es stets geschafft, mich in kritischeren Phasen meiner Dissertation auf andere Gedanken zu bringen und mich von meinen Sorgen wegzulenken.

Abschließend möchte ich noch meinem Freund Waldemar gedenken, der vor 15 Jahren leider viel zu früh von uns gehen musste. Danke, dass du während meiner Schulzeit an mich geglaubt hast, während andere von meinem Scheitern überzeugt gewesen sind.

Abstract

(Driving) simulators offer a cost-effective and time-efficient way to test new systems and human-machine interfaces under controlled, safe, and standardized conditions. To be considered a valid research tool, however, simulator-based results must be comparable to those obtained in real-world environments, a premise that requires empirical confirmation. This dissertation addresses three central research gaps: **(1)** Most existing simulator validation studies primarily focus on driving performance parameters, while intraindividual processes such as physiological and cognitive stress responses remain largely underexplored. **(2)** While the acceptance of autonomous shuttle systems has been studied extensively, most investigations rely solely on self-report measures without offering participants direct experience with the technology. **(3)** Although multimodal stress indicators promise a more holistic understanding, the integration of heterogeneous single indicators often leads to inconsistent results and limited interpretability.

This dissertation is structured into two main parts. **Part A** investigates simulator validity with a particular focus on cognitive and physiological stress reactions to provide a more comprehensive understanding of user responses. **Part B** examines whether the acceptance of autonomous shuttles can be validly assessed in a simulator using an extended UTAUT2 model, and to what extent the inclusion of cognitive and physiological stress indicators enhances the model's explanatory power. Furthermore, it develops two robust composite indicators (Physiological Reaction and Cognitive Reaction) based on multiple individual measures, aiming to improve the reliability and validity of stress assessments in user experience research.

The findings show that **1)** while the simulator is subjectively perceived as more stressful than real driving, physiological responses demonstrate a high degree of similarity across both environments. **2)** Moreover, the use of a shuttle bus simulator proved to be a valid approach for acceptance research, and the integration of cognitive and physiological stress indicators significantly increased the explanatory strength of the acceptance model. **3)** The developed composite indicators showed consistent loading patterns across a variety of conditions and offer a practical, scalable approach for analyzing stress-related responses in mobility contexts.

Despite these positive results, the study also highlights certain limitations. The general application of findings to other user groups or mobility formats remains limited and requires further empirical investigation. Future research should apply more fine-grained, segment-specific data collection methods to better capture situational stress responses and explore the applicability of the composite indicators in real-world mobility scenarios.

Table of contents

List of tables

1 Motivation

Vision Zero is an internationally recognized guiding principle in transport policy with a clear goal: zero traffic fatalities. While many countries have seen a decline in the road death rate, the numbers remain high. In Germany alone, around 2,780 people died in road traffic in 2024 (Statistisches Bundesamt, 2025), while in the United States the figure was 39,345 (Sheperdson, 2025). The reported fatalities don`t include accidents causing severe injuries.

Approximately 90% of all car accidents are attributable to human error (Singh, 2018). Autonomous driving aims to counteract this issue. Autonomous vehicles (AVs) promise to significantly reduce this risk through precise sensors, vehicle communication, and the extensive elimination of human error (Abdel-Aty & Ding, 2024; Stoma et al., 2021)

On the path towards AVs, the automotive industry is undergoing major change due to technological advances in the field of Driver Assistance Systems (DAS) (Stoma et al., 2021). Nevertheless, these systems are spreading more slowly than expected. A key reason is limited user acceptance, often caused by poor usability or inadequate user guidance (Riedl et al., 2024). Increasing acceptance requires a customer-centered development approach, making customer centricity a key factor (Gummesson, 2008; Kleinaltenkamp et al., 2022). By placing users at the center of the development process, technical innovations can better align with user expectations, resulting in improved usability, safety, and satisfaction, which ultimately supports the adoption of both DAS and new mobility concepts such as autonomous shuttles.

A major part of the development of DAS is covered by driving tests. However, traditional driving tests face limitations. They are expensive, usually take place in late development stages, and are typically conducted with professional test drivers (Mohajer et al., 2015), which means the perspective of typical end users is often excluded.

Driving simulators offer a promising alternative. They enable standardized, controllable test scenarios in a safe environment (Caird & Horrey, 2016; Winter et al., 2012), including with laypeople and in very early development phases (Xue et al., 2023). Moreover, they allow for scenarios that would be ethically or practically difficult to test in real life (Caird & Horrey, 2016). In addition to the benefits in terms of cost and time, this also creates scientific value.

For simulators to be used as a valid research tool, behavior in the simulator must be comparable to that in real driving situations (Czaban & Himmels, 2025). The comparability between a driving simulator and a real vehicle is referred to as simulator validity. This generally refers to the extent to which the behavior and reactions of participants in the simulator correspond to those in real driving situations on a behavioral, physiological, or cognitive level (Himmels, 2025).

Previous validation studies have focused heavily on driving dynamics such as e.g. braking, often without sufficiently considering the human perspective (Wynne et al., 2019). In a similar vein, traditional technology acceptance models have increasingly been criticized for their limitations in explaining user behavior (Blut et al., 2022). This work goes beyond the original approaches by systematically integrating physiological

and cognitive (stress-related) indicators. Including these physiological and cognitive parameters provides a more comprehensive picture of user responses, as it captures unconscious, emotional, and stress-related reactions that cannot be adequately assessed through traditional methods. The integration of these indicators is intended to enhance the explanatory power of the model.

Across all studies in this work, a consistent approach is applied, incorporating cognitive and physiological indicators into both the validation studies and the acceptance study. Moreover, this dissertation develops aggregated indicators that integrate individual measurements into interpretable metrics.

This thesis addresses three interrelated problem areas: the validity of simulators, the measurement and explanation of acceptance of autonomous shuttle buses using a simulator, and the question of how complex stress responses can be reliably captured. This gives rise to six central research questions, which are structured along these three thematic areas (see Chapter 3).

Each part of this work is based on the central approach of integrating cognitive and physiological indicators. In Part A, their relevance is demonstrated within validation studies, while in Part B they are employed as an extension to existing acceptance models to enhance explanatory power and beyond this these indicators are combined into a composite measure, which is intended to increase both usability and interpretability.

The thesis is structured as follows: Chapter 2 provides the theoretical framework necessary for this work. Part A (Appendix A) addresses the issue of simulator validity in three papers (Papers 1–3), focusing on cognitive and physiological responses. Part B (Appendix B) uses a shuttle bus simulator in Paper 4 to investigate user acceptance under realistic conditions. Finally, it deals with the development of an overall indicator that meaningfully integrates individual cognitive and physiological response indicators (Paper 5 and 6).

2 Theoretical background

2.1 Driving simulators as a research tool

The first driving simulators were already used in the 1930s (Lauer, 1960). Their goal remains to this day to develop vehicles more safely and to better understand user driving behavior (Carroll et al., 2023). Driving simulators have been a fixed component of automotive research for decades and offer a number of clear advantages.

They provide a safe, standardized, and therefore controllable test environment (Galante et al., 2018; Hussain et al., 2019; Pawar & Velaga, 2020; Winter et al., 2012), an aspect that is especially crucial in critical driving situations. They also enable tests with laypersons, while real-world prototype tests may only be conducted with trained drivers (Brookhuis & Waard, 2010). This results in high potential for time and cost efficiency in research and development (Drosdol & Panik, 1985; Pawar et al., 2022). At the same time, it becomes possible to involve end users at an early stage of the development process (Xue et al., 2023).

Moreover, driving simulators allow for targeted comparisons of different scenarios while controlling external variables that could influence driving behavior (Hussain et al., 2019). All these factors make the simulator a versatile and powerful research tool.

In order for findings from simulator studies to be transferable to real-world driving environments, the simulated environment must reflect relevant aspects of reality (Himmels et al., 2024; Pawar et al., 2022). Such transferability is only ensured if behavior in the simulator is at least similar to that in real driving situations. Empirical validation is essential for this.

Simulator validity, in the sense of the transfer-of-training theory (Blume et al., 2010; Liu et al., 2023) is achieved when behavior in the simulator is comparable to that in reality or at least elicits comparable reactions (Donkor et al., 2014; Himmels et al., 2024; Y. Wang et al., 2010; Wynne et al., 2019). Only then can reliable conclusions be drawn about user behavior in the real world.

Several validity concepts exist in the literature to evaluate this transferability. The most common are physical and behavioral validity (Bella et al., 2014). In addition, psychological validity also plays a role in this work (Vienne et al., 2014). In these concepts, relevant outcome variables are compared between real and simulated driving (Klüver, 2016; Zöller, 2015).

Physical validity (also called “fidelity”) refers to the degree to which the simulator technically and visually corresponds to a real vehicle (Klüver, 2016). However, high fidelity does not automatically mean higher simulator validity. Valid results can also be obtained with so-called low-fidelity simulators and vice versa. The choice of simulator type should therefore be based on the specific research question. The most cost-effective suitable configuration should be preferred (Himmels et al., 2024).

In contrast to physical validity, psychological validity focuses on ensuring not only that the external environment appears realistic, but also that internal cognitive processes (e.g., hazard assessment, decision-making) occur in the simulator as they would in real situations (Vienne et al., 2014). If thought and reaction patterns are comparable, psychological validity can be assumed. It is closely related to behavioral validity.

Behavioral validity is considered the central aspect of simulator validity overall (Godley et al., 2002; Terumitsu et al., 2007). Within behavioral validity, a distinction is made between absolute and relative validity (Blaauw, 1982). Absolute validity exists when numerical values (e.g., speed, reaction time) are identical between reality and simulation (Blaauw, 1982; Kaptein et al., 1996). Relative validity, on the other hand, means that effects go in the same direction in both conditions (e.g., speed increases in the real scenario and increases in the simulated scenario). Due to practical limitations (e.g., time, costs), absolute validity is often not achievable (Branzi et al., 2017). Relative validity is therefore considered sufficient for drawing valid conclusions (Pawar et al., 2022; Törnros, 1998). The required degree of validity ultimately depends on the specific research objective (Himmels et al., 2024; Mullen et al., 2011).

The outcome variables used for validation can be roughly divided into three categories: psychological, physiological, and objective measurements.

Psychological variables (e.g., NASA-TLX, Perceived Stress State Questionnaire) are usually easy to collect via questionnaire. However, a limitation is that such measurements are typically only possible before or after the driving scenario and are prone to biases such as social desirability bias (Nederhof, 1985).

Physiological measurements such as galvanic skin response or the use of electrocardiograms provide a more objective supplement. They can be continuously recorded during the drive and offer insights into unconscious responses, as they bypass cognitive filters (Healey & Picard, 2005; Lohani et al., 2019). However, their use requires special equipment and expertise. They may also be difficult to interpret due to competing influencing factors (e.g., stress vs. simulator sickness) (Dużmańska et al., 2018). Simulator sickness is a particularly relevant confounding factor. It involves a dissonance in the human vestibular system, resulting in symptoms such as nausea, dizziness, or discomfort, which may influence physiological responses.

Finally, objective parameters relate directly to observable driving behavior, such as lane keeping or braking behavior (Wynne et al., 2019). They provide indications of whether the simulator elicits realistic driving behavior (Blaauw, 1982; Blana, 1996).

The validity of simulator studies varies depending on the use case (Ahlström et al., 2012; Bella, 2008; Engen, 2008; Parduzi, 2021; Wynne et al., 2019), the outcome variables used (Himmels et al., 2024; Wynne et al., 2019) and the simulator configuration (Fischer et al., 2015; Himmels, 2025).

Since application scenarios and target groups vary greatly depending on the research objective, each simulator must be empirically validated for its specific use case (Blana, 1996).

While most previous validation studies have focused on vehicle dynamics parameters (e.g., acceleration, braking; see Wynne et al., 2019), relatively few studies have incorporated physiological and cognitive (stress-related) measures. Table 1 provides an overview of studies that extend the classical stimulus-response approach by including physiological indicators, thereby adopting a stimulus-organism-response framework (Davis & Granić, 2024). In some cases, these physiological measurements are complemented or even fully replaced by cognitive assessments. The use of physiological and cognitive indicators shifts the focus more strongly toward the individual.

At the physiological level, most studies primarily employ electrocardiogram (ECG) parameters and galvanic skin response (GSR). At the cognitive level, workload and stress questionnaires dominate, particularly the NASA-TLX and the Short Stress State Questionnaire.

Data analysis is predominantly conducted using frequentist methods, especially null hypothesis significance testing (NHST), an approach that has been increasingly questioned. NHST allows only for the detection of effects, a non-significant result, however, does not provide evidence for equivalence nor for sufficient statistical power (Himmels et al., 2024). This limitation is particularly problematic in validation studies with small sample sizes, which are common in driving simulator research. Himmels et al. (2024) therefore recommend the use of Bayes factors, which allow for inferences

about both differences and equivalence. Unlike p-values, Bayes factors provide relative probabilities for competing hypotheses (H_0 vs. H_1), enabling interpretable conclusions even in the presence of ambiguous findings.

Table 1. Key studies on (cognitive and physiological) driving simulator validation

Source	Research Question	Method	n	Results
Mueller, 2015	To what extent can driver behavior, performance measures, and physiological responses under high mental workload in the driving simulator be compared with those in real road traffic, and which variables are best suited for a valid transfer?	Test persons drive high- and low-complex sections in a simulator and in a real car. NASA-TLX is used for cognitive measurement. Heart rate, heart rate variability, galvanic skin response, and pupil diameter are used for physiological measurement. Measures are compared using MANOVA and ANOVA	34	Heart rate, heart rate variability, and gaze-related variables provide valid results across sections. Skin conductance level and pupil diameter do not provide valid results. Most NASA-TLX subscales show relative validity.
Reimer & Mehler, 2011	How well can physiological measures detect changes in cognitive workload, and to what extent can findings from driving simulators be transferred to real driving situations?	Test persons drive on a highway in a simulator and in a real car. Driving tasks include single task driving and a secondary task with three levels of difficulty. Heart rate and skin conductance level are measured. Data are analyzed using General Linear Models and pairwise t-tests.	26	Heart rate shows absolute and relative validity. Skin conductance level shows relative validity
Carter & Laya, 1998	How do driving experience, task type (straight driving vs. overtaking), and environment (road traffic vs. simulator) affect drivers' visual search strategies?	Test persons drive in a simulator and in a real car. Tasks include straight driving and overtaking maneuvers. Eye-tracking measurement is used. Data are analyzed using ANOVA.	16	Scan paths show relative validity. More fixations are observed in the simulator compared to real driving.
Milleville-Pennel & Charron, 2015	How do cognitive workload, affective experience (e.g., stress, enjoyment), and sense of presence differ when driving in a simulator compared to real driving conditions (driving school vehicle vs. personal car)?	Test persons drive in a simulator and in a real car for 30–50 minutes under four different conditions. NASA-TLX and the Questionnaire of Psychological Feeling are used for cognitive and affective measurement. Data are analyzed using t-tests, Multiple Factorial Analysis, RV-Index, and squared correlations.	14	Stress levels are higher in the simulator. Heart rate is also higher in the simulator. NASA-TLX shows relative validity, though not for all items. Questionnaire of Psychological Feeling shows relative validity, though not for all items.
Galante et al., 2018	To what extent is a driving simulator suitable for investigating mental workload compared to real road conditions?	Test persons drive a 78 km loop in a simulator and in a real car, including car-following, controlled approaching maneuvers, and rural single-carriageway driving. A	100	The sum score of NASA-TLX shows relative validity. All dimensions of the Short Stress State Questionnaire show relative validity. The distress subscale of the

	rotated figures task is used as a secondary task. NASA-TLX and the Short Stress State Questionnaire are used for cognitive measurement. Data are analyzed using ANOVA.	Short Stress State Questionnaire shows absolute validity.

2.2 Cognitive and physiological (stress-)reactions

To comprehensively capture stress responses, the literature recommends combining cognitive and physiological measurements (Dimoka et al., 2012). This multimodal approach is now considered standard in stress research, as it allows for better explanation of variance and more reliable prediction of stress (Becker et al., 2023; Tams et al., 2014).

Based on Selye‘s (1950) original definition, stress describes a nonspecific physiological response to demands for change. Later work shows that stress is a complex response pattern with psychological, cognitive, and behavioral components (Crosswell & Lockwood, 2020; Feuerstein et al., 2013). These responses indicate a disruption of physical or psychological balance, known as homeostasis (Cannon & Rosenberg, 1932; Chrousos, 1992; Robinson, 2018). A situation is perceived as stressful when the required resources for coping are judged as insufficient (Lazarus, 1990), making the situation feel overwhelming or threatening (Lee & See, 2004). The greater the discrepancy between demands and available resources, the more intense the measurable stress response becomes (Cohen et al., 2016).

Stress can be experienced as both positive (“eustress”) and negative (“distress”) (Lazarus, 1966; Selye, 1976). Eustress can enhance performance, whereas distress is perceived as burdensome. The transition between the two is individually variable. In relation to performance, the Hebbs curve is often cited, which describes a U-shaped relationship between stress and performance (Hebb, 1955). In research, however, the term “stress” is most often used synonymously with “distress.”

In addition to direction and intensity, duration also plays a key role. Acute stress occurs within seconds to days, while chronic stress can last for months or years (Baum, 1990; Crosswell & Lockwood, 2020). Stressors, internal or external stimuli, trigger a stress response when their intensity or duration exceeds a critical threshold. The resulting reaction involves the activation of the so-called "fight-or-flight" response (Cannon, 1939). The body attempts to restore disrupted homeostasis through physiological and psychological adaptation mechanisms (Boucsein, 2012; Giannakakis et al., 2022; Sapolsky, 2004). Affected functions include heart rate, blood pressure, respiration, and body temperature (Giannakakis et al., 2022).

Biologically, two primary stress systems are involved: the hypothalamic-pituitary-adrenal (HPA) axis and the sympathoadrenal medullary (SAM) system, the latter being part of the sympathetic branch of the autonomic nervous system (Cacioppo et al., 2017; Chrousos, 1992). The HPA axis responds to stressors with the release of corticotropin from the hypothalamus, leading to the secretion of adrenocorticotropin. This in turn stimulates the adrenal gland to produce cortisol, adrenaline, and

noradrenaline. These hormones raise blood glucose levels and temporarily supply muscles and the brain with more energy (Chrousos, 2009; Giannakakis et al., 2022).

The SAM system is responsible for the unconscious activation of the body. It increases sympathetic nervous system (SNS) activity while inhibiting the parasympathetic nervous system (PNS). Typical reactions include elevated heart rate, increased blood pressure, bronchodilation, and reduced activity of less acute functions such as digestion. While the SNS has an activating function, the PNS promotes relaxation and recovery through opposing mechanisms (Hall & Hall, 2020).

Central method for measuring physiological stress responses is skin conductance (Galvanic Skin Response (GSR)), particularly due to its ease and low cost of use (Caruelle et al., 2019). GSR measures changes in the electrical conductivity of the skin caused by activity in the eccrine sweat glands, which are solely controlled by the sympathetic nervous system and activated during emotional arousal (Boucsein, 2012; Setz et al., 2010). An increase in skin conductance is associated with emotional arousal, regardless of whether it is perceived as positive (e.g., eustress) or negative (e.g., anxiety) (Lang et al., 1993). Either the tonic level (skin conductance level (SCL)) or the phasic response to individual stimuli (skin conductance response (SCR)) is measured. Research shows that both values increase with rising stress levels (Ren et al., 2013; Setz et al., 2010).

One advantage of GSR is that it can be recorded continuously without interrupting task flow (Healey & Picard, 2005). For this reason, it is often used in combination with self-reports to obtain a more complete picture of the stress response.

Another common method for stress measurement is the electrocardiogram (ECG), which records the heart's electrical activity. The heart beats autonomously via electrical impulses generated in the sinus node and transmitted through a specialized conduction system (Hall & Hall, 2020). Both the sympathetic and parasympathetic nervous systems influence heart rate: parasympathetic activity reduces it via acetylcholine, while sympathetic activation increases frequency and contractility via noradrenaline (Giannakakis et al., 2022). In stress situations, sympathetic activity dominates: heart rate increases, contractile force rises, blood pressure and oxygen supply improve, typical characteristics of the fight-or-flight response (Engert et al., 2014; Hall & Hall, 2020).

To holistically assess a stress response, physiological measurements should be complemented by psychological assessments (Dimoka et al., 2012). The psychological component is usually measured via questionnaires and focuses on emotional or cognitive evaluation criteria, such as perceived mental or physical workload (Kabilmiharbi et al., 2022; Kelly et al., 2009), arousal level (Roos et al., 2021), subjective experience of stress (Qu et al., 2016; Rowden et al., 2011; Zhong et al., 2022) or acceptance of the deployed technology (Albers et al., 2020).

These cognitive assessments complement physiological data by adding a perception-based layer (Cohen et al., 1983) thereby aiding the interpretation of measured responses (Witte et al., 2021).

Common questionnaires for assessing subjective workload or stress include the NASA Task Load Index (NASA-TLX), the Short Stress State Questionnaire (SSSQ), and single-item scales. The NASA-TLX (Hart, 2006; Hart & Staveland, 1988) measures perceived workload across six dimensions: mental, physical, and temporal demands, perceived performance, effort, and frustration. The resulting workload is closely related to perceived stress (Alsuraykh et al., 2019; Rubio et al., 2004). The SSSQ is a 24-item questionnaire that captures short-term changes in stress perception (Helton, 2004; Ringgold et al., 2024). The items can be grouped into three subdimensions: worry, distress, and engagement. In addition to validated multi-item questionnaires, single-item measures can also be used to assess perceived workload or stress. These often correlate strongly with more extensive scales (Barré et al., 2017). A widely used instrument is the Visual Analogue Scale for Stress (VASS), where perceived stress is indicated on a continuum, typically ranging from “no stress” to “maximum stress” (Arza et al., 2019; Kabilmiharbi et al., 2022).

Table 2 provides an overview of studies investigating stress while driving in either a simulator or a real vehicle. Both frequentist methods (e.g., correlation analyses, ANOVA) and machine learning approaches are employed for analysis. These studies utilize both physiological and cognitive indicators.

At the physiological level, common measures include electrocardiogram (ECG) parameters (heart rate, heart rate variability), galvanic skin response (GSR), and eye-tracking metrics. In some studies, additional physiological indicators such as respiratory rate, muscle activity, or salivary amylase are also recorded. At the cognitive level, workload and stress questionnaires, particularly the NASA-TLX and the Short Stress State Questionnaire, are predominantly used, sometimes complemented by single-item measures, for example in the form of visual analog scales.

Table 2. Key studies on (physiological and cognitive) stress measurement during driving

Source	Research Question	Method	n	Results
Rendon-Velez et al., 2016	How does time pressure affect driving behavior, physiology, and drivers’ adaptation strategies?	Test persons drive in a simulator under conditions with and without time pressure. Cognitive measurements include the Mini Driver Behavior Questionnaire, the Multidimensional Driving Style Inventory, NASA-TLX, a confidence questionnaire, and a perceived time pressure questionnaire. Physiological measurements include eye tracking, electrocardiogram, respiration rate, and limb movement. Data are analyzed using correlations.	56	Physiological activity increases under time pressure, with higher heart rate, respiration rate, and pupil diameter. Blink rate decreases under time pressure.
Foy & Chapman, 2018	How do different road types affect drivers’ mental workload, and to what extent can these changes	Test persons drive in a simulator on different road types. Cognitive measurements include NASA-TLX and an inhibitory control task. Physiological	30	Galvanic skin response increases with rising workload. Heart rate and respiration rate do not show significant changes. Fixation duration decreases and

	be detected through behavior, subjective assessments, physiological measurements, eye movements, and prefrontal cortex activity?	measurements include functional near-infrared spectroscopy, eye tracking, heart rate, galvanic skin response, and respiration rate. Data are analyzed using repeated measures ANOVA.		horizontal scanning increases with higher workload. Subjective perception is significantly related to physiological responses.
Yamaguchi & Sakakima, 2007	Is salivary amylase activity (sAA) a reliable and rapid biomarker for detecting acute psychological stress responses during simulator driving, and how does sAA compare to subjective questionnaires and oculomotor measurements?	Test persons drive in a simulator. Cognitive measurement is conducted with a self-developed questionnaire containing seven adjectives (relaxed, fun, anxious, refreshed, stressed, uplifted, tired). Physiological measurements include salivary amylase and electrooculography. Data are analyzed using paired t-tests.	20	Questionnaire results do not show significant changes. Salivary amylase increases significantly during driving, indicating stress. Stress during driving can be detected through salivary amylase.
Healey & Picard, 2005	Can driver's workload and stress during real driving situations (urban, highway, resting periods) be reliably detected and classified using physiological measurements, and are these signals suitable for continuously and automatically recognizing driver states so that adaptive vehicle systems can respond?	Test persons drive in real traffic on urban and highway routes with resting periods. Physiological measurements include galvanic skin response, electrocardiogram, respiration rate, skin temperature, and electromyography. Data are analyzed using video coding, pattern recognition, and correlation analysis.	16	Low, medium, and high stress levels can be distinguished with 97.4% accuracy. Galvanic skin response is the most reliable stress marker. Heart rate and heart rate variability are reliable markers for stress detection. Electromyography and mean respiratory rate are less suitable.
Daviaux et al., 2020	Can acute stress in realistic driving situations be objectively quantified using phasic components of electrodermal activity, particularly during	Test persons drive in a simulator and are exposed to unexpected, stress-inducing traffic events. Mechanical driving data such as braking force are collected. Cognitive measurements include the Visual Analogue Scale, the Arousal Predisposition Scale, and the Edinburgh Handedness Inventory. Physiological measurement is	12	Stressful scenarios lead to increased galvanic skin response values. Galvanic skin response correlates with subjective stress ratings. Phasic galvanic skin response components are better suited than tonic components to quantify acute stress.

unexpected stress-inducing events in traffic?	conducted with galvanic skin response. Data are analyzed using repeated measures ANOVA, paired samples t-tests, and correlation analysis.

2.3 User acceptance of technical systems

For novel technical systems such as autonomous vehicles to successfully establish themselves in the transport sector, user acceptance is just as crucial as technological advancement. Various theoretical models and constructs exist in the literature to study this acceptance. In general, technology acceptance is understood as the willingness to use new technologies and integrate them into everyday life (Davis, 1989; Venkatesh et al., 2003).

The theoretical foundations of technology acceptance are based on behavioral psychology models, particularly the Theory of Reasoned Action (TRA) by Fishbein et al. (1975), and the subsequent Theory of Planned Behavior (TPB) by Ajzen (1991). The TRA posits that behavior is determined by behavioral intention, which in turn is influenced by subjective norms and personal attitudes toward the behavior. The TPB extends this model by adding the concept of perceived behavioral control, i.e., a person's assessment of whether they are actually capable of performing a given behavior.

A model specifically tailored to technology acceptance is the Technology Acceptance Model (TAM) by Davis (1985, 1989). It is based on the constructs of Perceived Usefulness, the user's subjective perception of the technology's benefit, and Perceived Ease of Use, the perceived effortlessness of using it. The model assumes that technologies perceived as easy to use are also regarded as more useful, which in turn increases acceptance.

Building on TAM, Venkatesh et al. (2003) developed the Unified Theory of Acceptance and Use of Technology (UTAUT), which identifies four key influencing factors: Performance Expectancy, i.e., the belief that using the technology will provide personal benefit; Effort Expectancy, the perceived ease of use; Social Influence, the effect of others' opinions; and Facilitating Conditions, the perceived availability of supporting resources such as technical infrastructure or assistance. These relationships are moderated by factors such as gender, age, experience, and voluntariness of use.

The UTAUT2 model (Venkatesh et al., 2012) expands the original UTAUT by adding three constructs: Hedonic Motivation, i.e., enjoyment and user experience; Price Value, the perceived trade-off between cost and benefit; and Habit, the extent to which using the technology has become habitual. In contrast to the original UTAUT, the moderator voluntariness of use is no longer included in UTAUT2.

For technologies such as autonomous driving, which are still in early stages of adoption, traditional acceptance models are often insufficient. Particularly relevant in this context are the constructs of trust and perceived risk. Trust refers to the belief that a technology is reliable, safe, and functional, which reduces uncertainty and strengthens the intention to use it (Gefen et al., 2003). Perceived risk describes the subjective evaluation of potentially negative consequences associated with using a

technology. In the case of autonomous vehicles, it is considered a common barrier to acceptance (Featherman & Pavlou, 2003; Menon, 2017). Both constructs, trust and perceived risk, are closely interconnected (Featherman & Pavlou, 2003) and form important extensions to existing acceptance models, particularly in evaluating innovative technologies within the mobility context.

Table 3 presents key studies in which acceptance models have been applied to investigate the use of autonomous shuttles. Nearly all of these studies are based on (extended) UTAUT models, with analyses predominantly conducted using structural equation modeling. Notably, data collection in most cases relies exclusively on questionnaires. Since autonomous shuttles are still only available in limited quantity on the mass market, respondents' answers often reflect expectations and perceptions rather than actual experiences, a point that should be considered critically.

Table 3. Key studies on technology acceptance models regarding autonomous shuttle buses

Source	Research Question	Method	n	Results
Korkmaz et al., 2022	Which factors influence individuals' behavioral intention to use autonomous public transport systems, and how can an extended acceptance model (based on UTAUT2) explain this behavior?	Participants complete online surveys and paper-pencil interviews. Used constructs include performance expectancy, perceived usefulness, perceived value, facilitating conditions, hedonic motivation, effort expectancy, trust and safety, habit, perceived risk, and behavioral intention. Data are analyzed using exploratory factor analysis, confirmatory factor analysis, and structural equation modeling.	303	Trust and safety show the strongest influence on behavioral intention, followed by social influence, performance expectancy, and habit. The model explains 72% of the variance in behavioral intention.
Rejali et al., 2024	Which factors influence the public's willingness to use Autonomous Modular Transit in the future?	Participants complete an online survey. Used constructs include performance expectancy, effort expectancy, social influence, facilitating conditions, hedonic motivation, perceived value, habit, trust, green perceived usefulness, and behavioral intention. Data are analyzed using structural equation modeling.	1662	Performance expectancy has the strongest influence on behavioral intention to use autonomous modular transit. Additional significant influences are found for social influence, hedonic motivation, and trust.
Nordhoff et al., 2018	How do potential users accept automated shuttles – both with regard to the shuttle itself and its role as a feeder in public transport – and which factors determine usage intention and willingness to pay?	Participants complete a questionnaire after a ride in a real automated shuttle. Used constructs include perceived enjoyment, performance expectancy, perceived safety, control, social influence, environmental attitudes, intention to use, and willingness to pay. Data are analyzed using principal component analysis and Pearson correlations.	274	Perceived enjoyment, service quality, environmental attitudes, and intention to use have the strongest influence on acceptance. Performance expectancy and perceived safety show moderate

				influence. Control, social influence, and willingness to pay show little or no influence.
Nordhoff et al., 2021	How do various factors from the UTAUT and DIT models, as well as trust and automated shuttle sharing, affect the behavioral intention to use automated shuttles in public transport?	Participants complete a questionnaire after a shuttle ride. Used constructs include performance expectancy, facilitating conditions, social influence, trust, behavioral intention, trialability, compatibility, and automated shuttle sharing. Data are analyzed using confirmatory factor analysis and structural equation modeling.	340	In the standard model, performance expectancy is the only predictor of behavioral intention, with an explanatory contribution of 39.7%. In the extended model, compatibility, trust, and automated shuttle sharing have significant influence, increasing the explanatory contribution to 48.5%.

The table also illustrates that acceptance models are widely used in this field of research. Their application, however, is not limited to autonomous shuttles, but extends to many other domains. With their extensions, these models rank among the most cited theories in technology acceptance research and are also broadly applied in disciplines such as service management and marketing (Baier et al., 2025).

At the same time, there is criticism that models like UTAUT are increasingly reaching their limits and no longer generate truly novel insights. While the constructs employed are helpful for explaining acceptance, they only partially cover central influencing factors (Blut et al., 2022). Since a large portion of studies still relies on TAM surveys, which evaluate products or services exclusively through self-reports (Baier et al., 2025), a comprehensive understanding is lacking. Accordingly, it is recommended to integrate new predictors into the theory and to adopt methodologically broader research designs, for example, by including observational data, qualitative analyses, or longitudinal studies (Blut et al., 2022)

An early alternative approach was developed by Rese et al. (2014) with so-called TAM dictionaries. In this approach, terms from online reviews were automatically classified as positive or negative to computationally analyze customer opinions. The results were comparable to classical TAM surveys, although data processing was labor-intensive and the statistical power was limited. Subsequent studies (Rese et al., 2017; Schreiber, 2020) were able to replicate the findings and confirm the usefulness of this approach.

Another innovative avenue involves extending the models beyond consciously controlled processes. By incorporating physiological measurements, automatic and unconscious responses can also be captured (Dimoka et al., 2012). This opens the “black box” of the stimulus-response approach, allowing cognitive, emotional, and attention-related processes to be considered, processes that often remain undetected in classical surveys (Davis & Granić, 2024). The goal is to integrate neurophysiological methods and theories into the explanation of technology acceptance, thereby

developing individualized TAM models that account for differences in cognition, emotion, and neurobiology, enabling more precise predictions.

2.4 Challenges of using stressindicators in ux-research

As described in Chapter 2.2, the stress response involves a complex interplay of physiological and cognitive processes (Pinel & Barnes, 2021). Consequently, no single stress marker is capable of capturing the human experience of stress fully and validly (Arza et al., 2019).

The exclusive use of subjective methods is influenced by cognitive appraisal processes, which can lead to distortions, for example, through conscious or unconscious self-regulation or misjudgment (Lin et al., 2005). Moreover, many of these tools are not designed for continuous measurement during an experiment but instead provide only momentary assessments.

Established biochemical markers such as cortisol do allow for the detection of actual physiological stress responses, but they require invasive methods (e.g., saliva samples) and do not support real-time continuous measurement (Arza et al., 2019).

In contrast, physiological methods such as heart rate or skin conductance measurement allow for continuous data collection. However, their interpretation is often ambiguous. For instance, an increase in skin conductance can indicate either elevated stress or positive excitement (Boucsein, 2012; Cacioppo et al., 2017). Additionally, studies have shown that physiological parameters do not always respond consistently, some stressors may be clearly reflected in one signal, while others remain unchanged (Leis & Lautenbach, 2020).

Combining multiple indicators can help generate a more comprehensive picture (Arza et al., 2019; L. Chen et al., 2017), even though discrepancies may arise between physiological and subjective results (Lin et al., 2005). A multivariate approach can support a more nuanced understanding of the breadth of the stress response and enable valid conclusions about users' experience and behavior (Arza et al., 2019; L. Chen et al., 2017). In this context, composite indicators play a key role: they simplify complex data analyses by aggregating multiple measures into a single, interpretable index (Nurdianto et al., 2024).

Given the complexity of human stress responses, it is therefore necessary to go beyond isolated single measurements. Multimodal approaches that combine various physiological and cognitive methods are essential for developing robust and reliable overall indicators (Apraiz Iriarte et al., 2021; Arza et al., 2019; L. Chen et al., 2017; Lin et al., 2005; Mauri et al., 2010).

3 Research questions and methodologies

Already in the 1970s, the first validation studies on driving simulators were conducted (e.g., Barker et al., 1978), and their relevance has steadily increased since then. However, most research has so far focused on objective driving data such as speed or braking behavior. Studies that examine physiological parameters (e.g. eye-tracking; Carter & Laya, 1998) or cognitive aspects (e.g. self-reports; Reimer et al., 2006) are comparatively rare. A systematic literature review by Wynne et al. (2019) supports this

pattern: out of 44 identified comparison studies, the majority focused on objective indicators.

While objective driving data is technically easy to collect, its interpretation often follows a classical stimulus–response (S-R) model. In this model, the driving situation is considered the stimulus, and the resulting behavior the direct response. However, this approach neglects intra-individual cognitive and physiological processes that significantly influence behavior. For example, test subjects may behave similarly in simulated and real driving scenarios in terms of observable actions, while their underlying physiological responses differ fundamentally. These processes are crucial to reliably assess the validity of simulators compared to real-world driving.

Measuring and interpreting cognitive and physiological stress indicators is far more complex than analyzing objective driving data (Czaban & Himmels, 2025). Nevertheless, such an approach is essential to expand the classical S-R model into a stimulus–organism–response (S-O-R) model. Multiple studies therefore emphasize the need to systematically assess emotional and cognitive responses (Blana, 2000; Boer, 2000). This leads to the following research questions:

RQ1: To what extent do physiological stress indicators correlate between simulated and real-world driving? How valid are these indicators in the simulation context?

RQ2: To what extent do cognitive stress indicators correlate between simulated and real-world driving? How valid are they?

RQ3: To what extent does gaze behavior correlate between simulator and real-world driving? Can gaze behavior serve as a valid comparison indicator?

Beyond the issue of simulator validity, another central challenge arises: to what extent are simulation environments suitable for studying the acceptance of new mobility technologies, especially autonomous shuttle buses? As a disruptive innovation, autonomous shuttles offer numerous potential advantages, including improved traffic safety (Dehghani et al., 2025), more efficient resource use (Bansal et al., 2016; Othman, 2023), better traffic flow (Mira Bonnardel et al., 2020) and demand-responsive public transport (Golbabaei et al., 2022; Mahmud et al., 2022). However, public acceptance is a critical factor for the successful diffusion of such technologies (Riedl et al., 2024).

Although a substantial number of studies on the acceptance of autonomous shuttles exist (e.g. Cai et al., 2023; Madigan et al., 2017; Nordhoff et al., 2017), typically based on established models like UTAUT or UTAUT2, three core weaknesses can be identified: First, most studies are purely hypothetical, participants have not actually experienced the technology. Since experience has been shown to significantly increase acceptance (e.g. Eden et al., 2017; Salonen & Haavisto, 2019), the external validity of such studies remains limited.

Second, the few empirical studies that include real (e.g. Herrenkind et al., 2019; Madigan et al., 2017) or simulated driving situations typically focus on everyday scenarios. Critical driving situations, which are particularly relevant for perceived safety, have rarely been considered, although these scenarios are likely key in shaping acceptance.

Third, traditional acceptance models based on self-reports have methodological limitations: they capture only conscious evaluations, neglecting unconscious and emotional responses. Integrating physiological and cognitive indicators, in the sense of a NeuroIS approach, may offer a deeper understanding (Davis & Granić, 2024).

This leads to two additional research questions:

RQ4: Can an autonomous shuttle simulator serve as a valid tool for measuring real-life acceptance of autonomous shuttles?

RQ5: Does integrating physiological and cognitive stress indicators increase the explanatory power of acceptance models?

To capture stress responses, this dissertation employs a multi-method approach combining subjective (e.g., NASA-TLX) and physiological indicators (e.g., skin conductance). While this method mix allows for a more holistic view of stress responses, previous studies have shown that the results are often heterogeneous or even contradictory (Li et al., 2013). In particular, discrepancies between subjective and physiological indicators frequently complicate interpretation. One potential solution lies in the development of composite indicators, which integrate diverse data sources into a valid overall metric. Several approaches of this kind have been proposed (e.g. Apraiz Iriarte et al., 2021; Lin et al., 2005) and partially tested (e.g. L. Chen et al., 2017; Mauri et al., 2010), but are still not widely used, mainly due to methodological, technological, and financial barriers.

So far, literature lacks a systematic examination of which individual indicators, especially physiological and cognitive, are most suitable for creating valid composite indicators. This leads to a final research question:

RQ6: Which physiological and cognitive indicators are suitable for forming a composite indicator that offers higher validity and explanatory power for assessing stress responses?

This cumulative dissertation is structured as follows: Part A (Appendix A) presents three studies focusing on simulator validity, addressing research questions RQ1 through RQ3. Paper 1 examines various cognitive and physiological stress indicators across different driving scenarios. Paper 2 analyzes the similarity of physiological stress patterns over the course of a drive. Paper 3 compares gaze behavior in real-world and simulated environments.

Part B (Appendix B) includes a fourth paper that investigates the acceptance of autonomous shuttles using an extended UTAUT2 model. A shuttle simulator is employed to enable realistic user experiences and to deliberately test critical driving situations. The integration of physiological and cognitive indicators serves to deepen and expand the model (addressing RQ4 and RQ5).

Beyond that Part B is dedicated to the development and application of valid composite indicators. Paper 5 identifies suitable individual indicators and combines them into composite measures. In Paper 6, these are tested within an intervention context, specifically addressing whether the use of lavender scent in high-stress driving situations can contribute to a reduction in physiological stress responses (RQ6).

Table 1 provides an overview of the publications included in this dissertation and their current status.

Table 4. Submissions and publication status of the research papers

#	Title	Authors (CRediT authorship contribution statement)	Journal (VHB Jourqual 4 Rating)	Status
1	Investigating simulator validity by using physiological and cognitive stress indicators	**Marcin Czaban** (Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization) **Chantal Himmels** (Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis)	Transportation Research Part F: Traffic Psychology and Behaviour (VHB B)	Published
2	Comparison of gaze behavior in real and simulated driving	**Marcin Czaban** (Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization) **Christian Purucker** (Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis)	Proceedings of the NeuroIS Retreat 2025 (VHB C)	Published
3	User interaction with digital twins: how comparable are simulation and reality	**Marcin Czaban** (Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization) **Eldar Sultanow** (Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation) **Alina Chircu** (Writing – review & editing, Writing – original draft) **Christian Czarnecki** (Writing – review & editing, Writing – original draft) **Joachim Riedl** (Writing – review & editing, Writing – original draft) **Stefan Wengler** (Writing – review & editing, Writing – original draft)	Business & Information Systems Engineering (VHB B)	Under Review (1st round)
4	User Acceptance of Autonomous Shuttle Systems: A UTAUT2-Based Analysis	**Marcin Czaban** (Writing – review & editing, Writing – original draft, Visualization, Project	Journal of Public Transportation	Under Review (1st round)

	with simulated driving tests and physiological measurement	administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization) **Daniel Baier** (Writing – review & editing, Writing – original draft, Methodology, Formal analysis)	(Not in VHB; Q1 rated)	
5	Single measurement vs. composite indicators for user experience research	**Marcin Czaban** (Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization), **Joachim Riedl** (Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis) **Stefan Wengler** (Writing – review & editing, Writing – original draft)	Behavior Research Methods (VHB B)	Under Review (1st round)
6	Scent and stress: The role of lavender and perception in simulated driving scenarios	**Marcin Czaban** (Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization) **Sarah Victoria Mohr** (Writing – review & editing, Writing – original draft, Methodology, Formal analysis) **Joachim Riedl** (Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis) **Stefan Wengler** (Writing – review & editing, Writing – original draft)	Proceedings of the NeuroIS Retreat 2025 (VHB C)	Published

Table 5 provides a comprehensive overview of the studies listed in Table 4 of this dissertation, focusing on the research question, methodology, data basis, and key findings of each work.

It becomes evident that this cumulative dissertation extends existing research approaches by systematically integrating physiological and cognitive (stress) indicators into both data collection and modeling.

This extension is based on the assumption that intrapersonal processes should be systematically considered to obtain a holistic understanding of user experience and thereby increase the explanatory power of existing models.

Particularly in driving simulator validation studies, the focus has so far been on vehicle dynamics, largely overlooking the perspective and subjective experiences of users. Only a few studies explicitly examine the individual and their physiological and/or cognitive responses (e.g., Wynne et al., 2019; see table 1).

In the field of technology acceptance research, classical survey-based models have also been criticized for providing an incomplete explanation of acceptance (Blut et al., 2022). This dissertation addresses this gap by following the recommendation of Davis and Granić (2024) to extend traditional acceptance models with physiological variables, thereby transforming the classic stimulus-response approach into a stimulus-organism-response framework.

The use of physiological and cognitive stress indicators can, however, lead to inconsistent results. The methodological innovation of this work lies in developing stable and valid composite indicators from these individual measures, which can be applied universally.

Overall, the dissertation contributes on three levels:

1. Expanding driving simulator validation studies to include the user perspective.
2. Strengthening technology acceptance models through the integration of a NeuroIS approach.
3. Developing a new methodological tool in the form of combined composite indicators for user research.

Table 5. In-depth summary of included studies: objectives, methodology, and results

#	Research Question	Method	n	Results
1	To what extent can physiological stress indicators obtained in a medium-fidelity driving simulator be transferred to real-world driving (absolute and relative validity)? To what extent can cognitive stress indicators obtained in a medium-fidelity driving simulator be transferred to real-world driving?	Participants complete both a 23 km real-world and a simulator drive in a within-subject design. Physiological measurements include electrocardiogram-based variables, galvanic skin response-based variables, and salivary cortisol. Cognitive measurements include NASA-TLX, the Short Stress State Questionnaire (SSSQ), and a single-item stress measure. Data are analyzed using Bayesian ANOVA and Bayesian paired t-tests	68	Skin conductance response, RMSSD, SDNN and skin conductance tonic level show absolute validity. Skin conductance response, skin conductance level, RMSSD, and SDNN. Peak Amplitude, heart rate, and RR-Interval show no validity. For cognitive measures, only the SSSQ worry dimension shows absolute validity. NASA-TLX, the SSSQ dimensions distress and engagement, and single-item stress measures show no validity. Subjective stress is perceived as higher in the simulator.
2	Does gaze behavior (fixation patterns) systematically differ between real driving and simulator driving across different road types (urban, rural, highway)?	Participants complete a 23 km real-world and simulator drive in a within-subject design. Eye-tracking is used for physiological measurement. Data are analyzed with gaze-point-plot analysis and expert ratings.	12	Overall fixation patterns are moderately similar between real-world and simulator driving. The highest similarity is found for the route as a whole, while the lowest similarity occurs in urban sections. Fixation density is highest in the center, with greater peripheral dispersion in the real vehicle. The simulator shows a slightly shifted field of view. Patterns are similar across segments, with low intrapersonal differences.
3	What are the limitations of mean analysis across different segments?	Participants complete a 23 km real-world and simulator drive in a within-subject design. Physiological measurements include electrocardiogram-based and galvanic skin	68	Mean values obscure dynamics, peak values, and contextual differences. Variability, learning effects, and complex interrelations are lost in mean analysis.

	What insights can be gained from time-series analysis?	response-based variables. Data are analyzed using t-tests for segments and time-series analysis with correlation for the whole drive.		Time-series analysis shows a moderate correlation of skin conductance level between simulator and real driving. Mean comparisons indicate larger differences than time-series analysis.
4	Is a shuttle bus simulator a suitable tool for acceptance studies? Which factors explain the intention behavior to use an autonomous shuttle bus? Is the explanatory power increased by integrating physiological and cognitive stress responses?	Participants complete a simulated shuttle bus ride with critical situations. Physiological measurements include electrocardiogram-based and galvanic skin response-based variables. Cognitive measurements include the Perceived Stress Scale (PSS-10), NASA-TLX, single-item stress measurement, and the extended UTAUT2 model. Data are analyzed using partial least squares structural equation modeling.	104	Social influence, facilitating conditions, trust, perceived risk, and perceived usefulness have a significant positive effect on behavioral intention. Cognitive reaction and hedonic motivation show a significant negative effect. Performance expectancy and effort expectancy show no significant effect. Cardiac activation has a significant positive effect on cognitive response, while electrodermal activation has a significant negative effect.
5	How strongly do physiological and cognitive indicators of stress correlate with each other, and can stable combined stress indicators be derived from them? To what extent does a combined stress indicator increase with rising situational demands or the requirements of a stress condition? Does the composition of a combined stress indicator remain stable across different situational conditions, ensuring reliable measurement?	Participants complete a 23 km real-world and simulator drive in a within-subject design. Physiological measurements include electrocardiogram-based variables, galvanic skin response-based variables, and salivary cortisol. Cognitive measurements include NASA-TLX, the Short Stress State Questionnaire (SSSQ), and a single-item stress measure. Data are analyzed using reliability analysis, correlation analysis, and principal component analysis.	68	Higher demands during driving lead to stronger stress responses. Single physiological and cognitive indicators provide an inconsistent picture. Aggregated indicators form two dimensions, physiological response and cognitive response. This two-dimensional structure remains stable across segments of both real and simulated driving.

	Do individual physiological measures and aggregated indicators truly reflect participants' subjective perception of stress, for example through correlation with objective biological stress markers such as cortisol?			
6	Does the controlled use of lavender scent during a driving task reduce measurable stress levels at the cognitive and physiological level compared to no scent exposure? Does the conscious perception of lavender scent influence its effect on stress levels?	Participants complete a simulator ride with critical situations in a between-subject design with two groups. During the ride, lavender scent is applied. Physiological measurements include galvanic skin response-based and electrocardiogram-based variables. Cognitive measurements include NASA-TLX and a single-item stress measure. Data are analyzed using reliability analysis, principal component analysis, and independent t-tests.	26	Exposure to lavender scent does not generally reduce stress. When participants consciously perceive the scent, cognitive stress levels are significantly lower, while physiological stress levels do not differ. Conscious perception may positively influence stress processing through cognitive mechanisms such as placebo effects or Hawthorne effects.

4 Results

4.1 Part A: Driving simulator validity

4.1.1 Summary of Research Paper No.1

Driving simulators have become indispensable in automotive research and development, offering cost and time efficiency as well as standardized testing procedures. However, for a simulator to serve as a meaningful tool, it must produce results that are comparable to those obtained in real vehicles, a concept referred to as simulator validity.

Previous validation studies have primarily focused on driving dynamics such as acceleration and braking, while underlying behavioral responses such as stress have often been neglected. This paper investigates simulator validity in the context of stress research by comparing physiological and cognitive stress indicators between real-world driving and a digital replication of the same route in a driving simulator. Since classical null hypothesis significance testing (NHST) faces limitations in this context, a Bayesian analytical approach was applied, which can provide evidence both for differences and for equivalence between conditions.

A total of 68 participants took part in the study. Each participant completed a 23 km route (divided into 7 sections consisting of rural roads, urban roads and highway) both in a real vehicle and in a medium-fidelity simulator. On the physiological level, galvanic skin responses (skin conductance response(SCR), skin conductance level (SCL), peak amplitude(PA)), cardiac activity (heart Rate (HR) and heart Rate Variability (HRV): RR-interval, RMSSD, SDNN), and salivary cortisol were recorded. On the cognitive level, data were collected using the NASA-TLX, the Short Stress State Questionnaire (SSSQ), as well as single-item measures on perceived stress, vehicle operation, and well-being.

The results revealed a mixed picture: among physiological measures, SCR, RMSSD, and SDNN demonstrated both absolute and relative validity. Salivary cortisol showed absolute validity, while SCL demonstrated only relative validity. PA, HR, and RR-interval failed to reach validity criteria. On the cognitive level, only the “Worry” dimension of the SSSQ showed absolute validity. All other cognitive measures scored higher in the simulator, suggesting that it was subjectively experienced as more stressful.

Overall, the findings suggest that driving simulators are well-suited for analyzing intraindividual physiological stress responses, whereas cognitive stress indicators should be interpreted with greater caution. Limitations of the study include the absence of driving dynamics data, high interindividual variability, and the fixed order of drives (with the real drive always preceding the simulated one).

4.1.2 Summary of Research Paper No. 2

Driving simulators provide safe, efficient, and standardized testing environments, making them highly relevant for vehicle development. For simulator-based tests to

yield meaningful insights, their outcomes should be comparable to those obtained from real-world driving. Studies addressing such comparisons are referred to as simulator validation.

Most existing validation studies have primarily focused on driving performance metrics such as lane keeping or speed. In contrast, gaze behavior, which is central to situational awareness, has received little attention to date, and the few available studies report inconsistent findings. This reveals a research gap regarding the validity of eye-tracking data in driving simulators. The present study therefore aimed to systematically examine whether gaze behavior differs between real and simulated driving conditions. Specifically, we investigated potential differences both across road types (urban, rural, highway) and between driving conditions (real vs. simulator).

To address this research question, twelve participants completed a 23 km drive (comprising urban, rural, and highway sections) in both a real vehicle and a medium-fidelity simulator using a digital replication of the route. Gaze behavior was recorded with eye-tracking glasses. Data analysis was conducted through gaze-point plots, supplemented by expert ratings assessing the similarity of gaze patterns.

The results indicate a moderate similarity of gaze patterns between real and simulated drives. At the same time, fixation patterns differed systematically across road types. Compared to real driving, simulator drives showed reduced peripheral dispersion and a systematic downward shift of gaze points. Patterns were most similar in urban and rural sections, whereas highway driving yielded the highest similarity values but also the greatest interindividual variability.

In summary, gaze patterns between real and simulated conditions can be considered largely comparable. The observed differences appear to be driven primarily by environmental and interface-related factors. Study limitations include the small sample size, potential learning effects due to the fixed order of drives, and the subjectivity of expert ratings. Future research should employ more advanced statistical approaches and compare simulators of different fidelity levels.

4.1.3 Summary of Research Paper No. 3

Digital twins are digital representations of physical entities designed to replicate their dynamics as accurately as possible. In this context, test tracks realistically reproduced in a driving simulator can be considered a digital twin, serving to virtually replicate real driving situations. Accordingly, the output of the digital twin should closely mirror that of real drives. The aim of the present study was to examine whether a driving simulator, in its role as a digital twin, elicits comparable physiological stress responses to those observed in real driving, and whether it may thus serve as a partial substitute. The research questions specifically addressed the limitations of segment-level mean analyses and the additional insights that can be gained from time-series analysis.

To answer these questions, 68 participants completed a 23 km test route consisting of urban, rural, and highway sections, both in a real vehicle and in a medium-fidelity simulator. During the drives, indicators of electrocardiography (heart rate (HR), RR-interval, RMSSD, SDNN) and galvanic skin response (skin conductance response

(SCR), skin conductance level (SCL), peak amplitude (PA)) were recorded. Data were analyzed both segment-wise using mean comparisons and across the entire driving duration in the form of time-series analysis. For the time-series analysis, SCL (tonic signal) was used.

The mean comparisons revealed no significant differences between real and simulated driving for SCR, RMSSD, and SDNN. By contrast, SCL and PA were significantly higher in the simulator, while HR was higher in the real vehicle. The RR-interval was longer in the simulator than in the real drive.

Because mean analyses smooth out fluctuations and thereby obscure temporal dynamics or extreme values, time-series analysis allows for a more fine-grained examination. Using SCL as an example, the time-series analysis revealed moderate correlations between conditions for urban and rural driving. Visual inspection of the trajectories also suggested a high similarity between curves.

In summary, a driving simulator can, as a digital twin, reproduce fundamental patterns of physiological stress responses. It appears particularly suitable as a substitute for real driving in controlled and less complex scenarios, though it does not capture the full range of responses observed in reality. The observed differences are primarily attributable to environmental and interface-related factors.

The study further highlights the limitations of mean analyses, as they smooth relevant dynamics of physiological responses. Time-series analysis thus represents a valuable complement. Another limitation lies in the absence of real-time bidirectionality between real and simulated driving environments: the digital twin currently functions only as a static replica without direct feedback. Future studies should therefore implement adaptive algorithms that allow for flexible real-time adjustments.

4.2 Part B: Acceptance and stress measurement using simulators

4.2.1 Summary of Research Paper No. 4

To promote the diffusion of autonomous shuttle buses, a priori acceptance of the technology is required in order to identify potential influencing factors. The present study therefore investigates which factors affect the behavioral intention to use autonomous shuttles. Since these vehicles are not yet widely available in the market, most previous studies have relied on survey data from individuals without practical experience. However, it is well established that experience with such technologies can substantially influence acceptance.

To address this research gap, a shuttle bus simulator was employed to present potentially critical driving scenarios and examine the impact of physiological and cognitive stress responses on technology acceptance. A total of 104 participants completed an approximately eight-minute simulated ride that included five potentially critical situations. Physiological measures included heart rate and RMSSD (cardiac activation) as well as skin conductance response and skin conductance level (electrodermal activation). Cognitive measures included an extended UTAUT2

questionnaire, the Perceived Stress Scale (PSS-10), NASA-TLX, and self-developed single-item measures. Data were analyzed using structural equation modeling (SEM).

Results indicate significant positive effects of social influence, facilitating conditions, trust & perceived risk, and perceived usefulness on behavioral intention. Contrary to expectations, hedonic motivation had a significant negative effect. Cognitively perceived stress reduced behavioral intention and was primarily explained by cardiac activation. Overall, the model accounted for 61.1 % of the variance in behavioral intention.

These findings suggest that acceptance of autonomous shuttles is influenced not only by technical and usability factors but also by users' perceived safety and stress levels.

Limitations of the study include a homogeneous participant group and a relatively small sample size. Additionally, participants were aware of the safety of a simulated environment, which may have affected physiological responses. Future research should apply these methods in real driving scenarios to further validate the findings.

4.2.2 Summary of Research Paper No. 5

In line with a customer-centered marketing approach, understanding users' opinions and their usage experience is crucial for product development. A wide range of questionnaires and physiological measurements are available for use in product testing to capture users' reactions. However, interpreting single measures often poses a challenge, and comparisons across individual indicators may even yield contradictory results.

This study therefore investigates whether established (physiological and cognitive) single indicators for measuring stress and user reactions in technology interactions are suitable for providing reliable insights. The aim is to derive combined and more stable measures from individual indicators that offer higher explanatory power and reliability than single metrics.

To address this question, 68 people participated in a test drive boasting a within-subject design. They completed a 23 km route divided into seven segments (14 "situations" in total) both in a real vehicle and in an identically modeled driving simulator. During the drives, physiological stress indicators (Galvanic Skin Response: skin conductance level (SCL), skin conductance response (SCR), peak amplitude (PA); electrocardiogram: heart rate (HR), RR-interval, RMSSD, SDNN) as well as cognitive stress indicators (NASA-TLX, Short Stress State Questionnaire, single items: stress, physical well-being, vehicle operation) were collected. The data were analyzed using correlation analyses as well as reliability and factor analyses.

The results indicate that the simulator ride was perceived as more stressful than the real drive. Analyses of single indicators yielded partly inconsistent findings. However, the factor analysis revealed two stable composite factors: Physiological Reaction (comprising SCR and HR) and Cognitive Reaction (comprising NASA-TLX, single-item stress, and single-item physical well-being). These combined indicators remained stable across all situations and showed more consistent associations with situational coping and cortisol levels than the single indicators.

It is recommended that future user studies rely on these combined indicators, with cognitive measures in particular offering an efficient alternative to more elaborate physiological procedures due to their simplicity and validity. Limitations of the study include the homogeneity of the sample, the limited variance in driving difficulty, and missing controls in cortisol assessment.

4.2.3 Summary of Research Paper No. 6

Over 90% of traffic accidents are attributable to human error. Stress is a central risk factor, as it can impair driving performance and increase accident risk. While moderate stress in terms of eustress may support performance, excessive stress (distress) clearly has a negative impact on cognitive and motor abilities. It is well established that scents such as lavender can have calming and stress-reducing effects and are therefore discussed as potential interventions in critical driving situations. Against this background, the present study investigates whether the use of lavender scent in critical situations in a driving simulator can reduce participants' stress levels.

A total of 26 participants completed a simulator drive that included five stress-inducing events. In the experimental group, lavender scent was released during the drive, whereas the control group drove without scent. Physiological (skin conductance response (SCR); heart rate (HR)) and cognitive stress indicators (NASA-TLX, self-reports) were measured. Based on a principal component analysis, the measures were aggregated into two factors: Physiological Reaction and Cognitive Reaction.

For analysis, participants were divided into three groups: no scent, scent without perception, and scent with perception. Data were analyzed using t-tests. The results suggest that lavender scent does not automatically lead to stress reduction. Physiological stress values were even lower in the control group. A significant reduction in cognitive stress was observed only when participants consciously perceived the scent, whereas unconscious exposure was associated with higher stress values.

In summary, scent interventions do not appear to be effective per se but require conscious perception, possibly due to a placebo mechanism. For practical applications, this implies that scents should be administered in a way that ensures participants are aware of them.

The study's limitations lie in the small sample size and the absence of baseline controls for physiological measures. The results should therefore be regarded as exploratory and require further validation in future research.

5 Conclusion

The present study pursued the goal of examining the suitability of a driving simulator as a valid substitute for real-world driving in terms of simulator validity. The focus was particularly on physiological and cognitive stress responses. Furthermore, it was examined whether an autonomous shuttle bus simulator is suitable for conducting acceptance studies and whether the inclusion of physiological and cognitive indicators can provide an additional explanatory contribution to acceptance measurement. Finally, an approach was developed for how complex stress responses can be validly represented by bundled indicators (so-called composite indicators). Based on six

research questions, three thematic blocks were addressed: simulator validity, acceptance research using a simulator, and indicator construction. Thec key findings are summarized and contextualized below along these research questions.

RQ1: To what extent do physiological stress indicators correlate between simulated and real-world driving? How valid are these indicators in the simulation context?

To answer the first research question, the results from Paper No. 1 (Appendix A.1) and Paper No. 3 (Appendix A.3) were used. While Paper No. 1 follows a segment-based analysis using a Bayesian approach, Paper No. 3 uses a time series analysis over the entire driving process. The combination of both methodological approaches enables a differentiated view: while the mean comparisons from Paper No. 1 allow conclusions to be drawn about section-specific differences between real and simulated driving, the time series analysis from Paper No. 3 allows for a dynamic evaluation of the physiological response course, thus avoiding potential distortions caused by mean formation.

Overall, a heterogeneous picture emerges regarding the validity of the physiological indicators examined. Some metrics show both absolute and relative validity between real and simulated driving, while others show no or only limited alignment.

In Paper No. 1, various indicators were considered: skin conductance-based measures such as skin conductance response, skin conductance level, and peak amplitude; ECG-based metrics such as heart rate, RR-interval, RMSSD, and SDNN; and salivary cortisol as an endocrine stress marker. For all parameters except cortisol, section comparisons were conducted using a Bayesian ANOVA; cortisol was compared using a Bayesian paired t-test due to the single measurement per drive.

The results in the field of skin conductance show that both absolute and relative validity could be established for the skin conductance response. As this indicator has hardly been considered in comparable simulator studies so far, no direct comparative data is available. Relative but not absolute validity could be demonstrated for the skin conductance level, a finding consistent with the results of Reimer and Mehler (2011). In contrast, Mueller (2015) reports no validity for the same indicator. The Peak Amplitude, in turn, showed neither absolute nor relative validity; here, too, comparative studies that would enable classification are lacking.

With regard to the ECG-based parameters, only RMSSD and SDNN, both measures of heart rate variability, show both absolute and relative validity. This finding is based on continuously recorded data that reveal consistent patterns between real and simulated driving. The RR-interval, on the other hand, proved not to be valid. For heart rate, no reliable evidence of validity was found in the present study. Thus, the results are consistent with studies by Johnson et al. (2011) and Milleville-Pennel and Charron (2015), which also report no validity for this parameter. Other studies, such as those by Reimer und Mehler (2011) or Mueller (2015), report partly contradictory results and indicate relative or absolute validity, suggesting a possible context or participant dependence.

Regarding salivary cortisol, absolute validity could be established: the values after real and simulated driving do not differ significantly. However, since only one measurement was taken per driving time point, relative validity in terms of section-related courses could not be examined. Comparable studies that use this parameter in a similar context are not yet known.

The time series analysis conducted in Paper No. 3 focused on the skin conductance level within urban and rural road segments. The analysis revealed moderate linear correlations in the response course between simulator and reality. This result suggests that the physiological reactions over time are similar in their dynamics, an aspect that could not be captured by mean comparisons alone.

In summary, with regard to RQ1, it can be stated that some continuously recorded physiological parameters, particularly SCR, RMSSD, and SDNN, show validity and are therefore suitable for use in simulator studies. These indicators offer a high degree of informative value, especially in intra-individual analyses. The salivary cortisol level also proves to be a robust parameter that shows no difference between real and simulated driving. Other indicators such as heart rate, peak amplitude, or RR-interval show less consistency and seem to be more influenced by contextual or individual factors. The time series analysis usefully complements the results by making parallels in the course of stress reactions visible. Overall, it appears that a driving simulator can represent a valid instrument, particularly in less complex environments or when examining general physiological reactions.

RQ2: To what extent do cognitive stress indicators correlate between simulated and real-world driving? How valid are they?

To answer the second research question, various cognitive stress indicators were analyzed in Paper No. 1. These included the NASA Task Load Index (NASA-TLX), the Short Stress State Questionnaire (SSSQ), as well as three single-item self-reports that captured perceived stress level (STRESS), vehicle operation, and physical wellbeing. Since all cognitive indicators were collected only retrospectively, i.e., after completion of the respective drives, only statements about absolute validity could be made. Section-related or time-dynamic analyses were not possible in this case.

The results show that the NASA-TLX does not exhibit absolute validity, meaning that subjective workload was not rated equally in simulator and real driving. This result contradicts several earlier studies: Diels et al. (2011), Galante et al. (2018), Milleville-Pennel and Charron (2015), Mueller (2015) and Lobjois et al. (2021) report relative validity for NASA-TLX, partly also for specific workload dimensions. One explanation for the deviation could lie in the retrospective collection, which may have led to distorted or context-dependent judgments.

The SSSQ was evaluated in the three subscales distress, engagement, and worry. Of these, only the dimension worry showed absolute validity, while distress and engagement did not deliver consistent results between the two driving situations. Again, there is some deviation from previous findings: Galante et al. (2018) reported absolute validity for the distress dimension and also found relative validity for all three dimensions. In the present work, such a pattern was not observed, which could also be due to methodological or contextual differences.

For the self-developed single-item scales STRESS, vehicle operation, and physical wellbeing, no absolute validity could be demonstrated. In all three cases, the responses differed significantly between simulator and real driving, indicating different perceptions of the driving situations.

Overall, a clear picture emerges: the drive in the simulator was subjectively perceived as more stressful and burdensome than the real drive. This result runs consistently through the various indicators and suggests that the simulator represented an unfamiliar, possibly even irritating situation for many participants. The increased cognitive stress in the simulator could therefore be explained less by the driving task itself and more by the unfamiliar environment, the lack of motion impressions, or other simulator-related factors. The overall inconsistent validity pattern could also be due to the fact that subjective assessments were given retrospectively for the entire drive, which may have blurred subtle differences between individual road segments.

In summary, it can be stated that cognitive stress indicators correspond only to a limited extent validly between real and simulated driving. The only dimension with reliable agreement is worry from the SSSQ. All other indicators suggest a higher subjective burden in the simulator. For future studies, it may therefore be useful to collect cognitive indicators in a more differentiated way, e.g., section-wise or in real time, to better capture validity in higher resolution and control contextual influences.

RQ3: To what extent does gaze behavior correlate between simulator and real-world driving? Can gaze behavior serve as a valid comparison indicator?

To answer the third research question, the results from Paper No. 2 (Appendix A.2) were used. In this study, gaze behavior in three route sections (urban, rural road, highway) as well as for the entire drive was compared between driving simulator and real-world driving, based on gazepoint plots and expert assessments of the similarity of fixation patterns.

For the entire drive, there is overall a moderate visual similarity between the two experimental environments. Across all routes, typical gaze patterns were recognizable, which appeared in similar form in both settings. This supports the assumption of relative validity of gaze behavior. At the same time, however, systematic differences in the absolute gaze distribution occurred: gaze dispersion was generally more limited in the simulator, and fixations were positioned slightly lower in the image frame. These deviations can be explained by reduced environmental stimuli, the design of the interface, or a potentially lower degree of realism of the simulator, an effect also observed in previous studies (e.g. Fors et al., 2013).

The differentiated analysis of individual route segments suggests that especially in the urban drive, high visual similarities were present, as the gazepoint plots with comparable fixation patterns in the central field of view and wide peripheral dispersion indicate. However, the expert ratings showed the highest average similarity value for the highway drive, with simultaneously the greatest interindividual variation. In both environments, fixations were strongly concentrated on the central field of view, supplemented by broad peripheral dispersion. On the rural road, a similar focus on the roadway was observed, with the difference that in the simulator, horizontal dispersion was lower, presumably due to fewer peripheral stimuli such as oncoming traffic. On the

highway, gaze was most centralized in both conditions, which can be attributed to the lower complexity and stimulus density of this driving situation.

The expert assessments support these observations: the highest average similarity was awarded for the entire route, followed by the highway, whereby the highest interindividual variation occurred precisely in the latter. This result suggests that some participants showed very similar gaze behavior in both environments, while others showed greater differences. The variance could be due to individual differences in the perception of realism or a changed sense of risk during real-world driving, especially on the highway with real traffic. At the same time, the graphical design of the simulated highway was less complex, therefore potentially reducing visual exploration.

Overall, it can be concluded that gaze behavior is fundamentally suitable as an indicator of comparability between simulated and real driving, particularly for the investigation of visual attention. Relative validity is supported by parallel gaze patterns across different environments. At the same time, systematic distortions such as reduced peripheral dispersion, vertical shifts, and context-dependent differences must be taken into account. These findings support using gaze behavior as a complementary indicator that provides valuable insights but also has limitations regarding its absolute validity.

RQ4: Can an autonomous shuttle simulator serve as a valid tool for measuring real-life acceptance of autonomous shuttles?

To answer RQ4, whether an autonomous shuttle bus simulator is suitable as a tool for realistically measuring the acceptance of autonomous shuttles, findings from Paper No. 4 (Appendix B.1) were used. Unlike many previous studies, which rely on surveys or experiences from rather everyday driving situations, the use of a simulator allows the targeted experience of critical driving situations. This enables the evaluation of acceptance based on actual, albeit simulated, experiences. To model acceptance, an extended UTAUT2 model was used, which examined the constructs Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Hedonic Motivation, Trust & Perceived Risk, and Perceived Usefulness with regard to their influence on Behavioral Intention.

The results show, deviating from Korkmaz et al. (2022) and Rejali et al. (2024), no significant effect of Performance Expectancy on the intention to use. A possible explanation lies in a content-related overlap with the construct of Facilitating Conditions. Effort Expectancy also showed no effect, unlike in Madigan et al. (2016), where a positive correlation was found. It is conceivable that in the present study, the subjectively perceived effort was rated higher due to the critical situations.

The strongest predictor in the model was Social Influence: the social context, i.e., norms and group influences, played a central role for the intention to use, similar to Kapser and Abdelrahman (2020). Facilitating Conditions showed, in accordance with Madigan et al. (2017), a positive effect, although this construct was captured with only one item due to model fit. Unexpected was the significant negative effect of Hedonic Motivation. A possible explanation is that the experience of critical driving situations reduced the feeling of fun or pleasure, or that the simulation itself was perceived as unrealistic or emotionally detached.

The construct Trust & Perceived Risk, based on Korkmaz et al. (2022) showed a positive effect on Behavioral Intention. As also in Choi and Ji (2015) the results confirm that trust and a sense of safety are key factors for acceptance. The second strongest predictor in the model was Perceived Usefulness, which, analogous to C.-F. Chen (2019), was positively associated with the intention to use. Thus, perceived usefulness proves to be an important driver of acceptance.

In summary, it can be said that Social Influence had the strongest impact on the intention to use, followed by Perceived Usefulness. Facilitating Conditions and Trust & Perceived Risk also contributed to the explanatory value, while Performance Expectancy and Effort Expectancy showed no significant effects. Hedonic Motivation had a negative effect, contrary to the original hypothesis. The results suggest that the use of a shuttle bus simulator, especially when including critical situations, is a valid and practice-oriented tool for capturing acceptance. It should be noted that participants were always aware that they were not in an actually life-threatening situation, which may have influenced their reactions.

While traditional UTAUT studies are based on questionnaires and thus measure conscious, cognitively filtered response behavior, this study additionally followed the suggestion of Davis and Granić (2024), to integrate unconscious physiological stress responses. The aim was to examine whether these could provide an additional explanatory contribution to the acceptance model. Since stress is experienced not only physiologically but also cognitively, the model was expanded to include cognitive stress indicators to depict a more comprehensive picture of the stress response.

RQ5: Does integrating physiological and cognitive stress indicators increase the explanatory power of acceptance models?

RQ5, based on Paper No. 4, aimed to clarify whether the integration of physiological and cognitive stress indicators increases the explanatory contribution to acceptance. It is assumed that physiological activation influences the subjective stress experience (Cacioppo et al., 2017). A distinction was made between Cardiac Activation (heart rate and heart rate variability, controlled by the autonomic nervous system) and Electrodermal Activation (skin conductance, controlled by the sympathetic nervous system). Both systems together form the physiological stress response, which in turn influences cognitive stress experience, which acts as a predictor for Behavioral Intention.

As expected, a significant positive effect of Cardiac Activation on the cognitive response was shown, which is in line with psychophysiological theories. Contrary to the assumption, however, a negative effect of Electrodermal Activation was found. This could be related to the fact that GSR responses in this study did not specifically reflect stress but were also influenced by unspecific activation such as attention, curiosity, or positive arousal.

Overall, the model suggests that subjective stress is an inhibiting factor for the acceptance of autonomous shuttle buses. The physiological measurements provide objective evidence of individual response patterns, but their interpretation must always be made in context. The results support psychophysiological models that assume bodily responses influence cognition and behavior. The combination of physiological

(cardiac, electrodermal) and cognitive stress indicators increases the explanatory contribution of the acceptance model. The physiological reactions influence cognitive stress processing, which in turn influences the intention to use.

The results suggest that cognitive load and stress, especially in critical driving situations, play an important role in the acceptance of autonomous shuttle buses. The integration of physiological and cognitive stress indicators not only provides additional insights into their direct effect on Behavioral Intention but also enables a deeper understanding of the underlying psychophysiological mechanisms.

RQ6: Which physiological and cognitive indicators are suitable for forming a composite indicator that offers higher validity and explanatory power for assessing stress responses?

To answer RQ6, which physiological and cognitive single indicators are suitable for forming a composite indicator that offers higher validity and explanatory power in capturing stress responses, established indicators were systematically examined within the context of real and simulated driving situations in Paper No. 5 (Appendix B.2). The starting point was the problem that single indicators often yield inconsistent or difficult-to-compare results when measuring situational activation and stress. The aim was therefore to identify valid and robust indicators suitable for constructing overarching, stable stress composites.

On the physiological level, parameters of galvanic skin response (skin conductance level and skin conductance response), electrocardiogram measures (including heart rate), and salivary cortisol values were used. The cognitive level was represented by the NASA-TLX, the Short Stress State Questionnaire (SSSQ), and two self-developed single-item scales on physical wellbeing" and self-reported stress.

The analysis of correlations and an exploratory factor analysis showed that the skin conductance response (SCR) is the most promising physiological single indicator: it correlates best with cognitive stress perceptions. While there was a stronger correlation between skin conductance level and cortisol level, SCR was overall more consistent. Among cardiovascular parameters, heart rate proved to be the most robust indicator and was therefore selected for further modeling. Other physiological measures such as respiratory rate or invasive procedures were not pursued further due to limited validity or high practicality requirements.

Although salivary cortisol is a very reliable biological stress marker, it was not continued as part of a continuously measurable indicator due to its limited temporal resolution but served for external validation of the identified stress composites.

On the cognitive level, NASA-TLX, the assessment of physical wellbeing, and self-reported stress emerged as suitable single indicators. The subscales of the SSSQ, on the other hand, showed lower consistency and were excluded.

Using principal component analysis, two stable factors were extracted based on the valid single indicators: a physiological stress dimension consisting of skin conductance response and heart rate, and a cognitive stress dimension consisting of NASA-TLX, Physical Wellbeing, and self-reported stress. These aggregated composite indicators showed stable loading patterns across 14 different test conditions. Thus, compared to

single measures, they offer improved explanatory strength, higher reliability, and more consistent correlations with situational demands. Their validity was further supported by their relation to salivary cortisol.

The transferability and reproducibility of the two stress composites were validated in Paper No. 6 (Appendix B.3). In this study, it was examined whether lavender scent had a stress-reducing effect during critical simulator drives. The two dimensions, Cognitive Reaction and Physiological Reaction, replicated with the same loading structure as in Paper No. 5. The analysis showed that lavender scent particularly reduced the cognitive stress response, but only when the scent was consciously perceived.

Overall, the results show that the integration of physiological and cognitive single indicators into the two composite indicators Cognitive Reaction and Physiological Reaction offers significant added value for practical stress measurement. The cognitive indicator consists of only eight items and can be easily integrated into studies while providing high interpretive power regarding subjective stress responses. If a multi-method approach is feasible, the physiological composite indicator also offers a more reliable and less disturbance-prone way to capture objective stress responses than individual physiological metrics.

In conclusion, the results of the present study show that the use of (driving) simulators represents a conceptually viable alternative to real-world driving, especially with regard to the measurement of physiological and cognitive (stress) responses. The findings provide differentiated insights into the validity of driving simulators, their application potential in the acceptance research of new mobility technologies, and the potential of forming valid composite indicators for physiological and cognitive responses. This dissertation thus contributes to the further development of multimodal approaches in mobility research and opens up perspectives for future studies in which human–machine interactions can be analyzed realistically under controlled conditions.

In view of rapid digitalization and the growing need for resource-efficient solutions, such valid and simulated test environments are becoming increasingly relevant for both basic research and the user-centered development and evaluation of existing and new mobility concepts.

Appendix A: Driving simulator validity

A.1 Research Paper No. 1: Investigating simulator validity by using physiological and cognitive stress

Authors: Czaban, M. & Himmels, C. (2025)

Citation: Czaban, M., & Himmels, C. (2025). Investigating simulator validity by using physiological and cognitive stress indicators. Transportation Research Part F: Traffic Psychology and Behaviour, 114, 831–851.

Doi: https://doi.org/10.1016/j.trf.2025.07.006

Abstract: Driving simulators are indispensable tools in modern automotive research and development. However, the transferability of findings to real-world driving, and thus, the validity of simulator-based results, cannot be assumed without empirical validation.

In this study, we examined physiological (Galvanic Skin Response-based measures, Electrocardiogram-based measures, salivary cortisol) and cognitive (NASA Task Load Index, Short Stress State Questionnaire, single-item ratings) stress indicators by comparing a real-world driving circuit with seven distinct sections to a medium-fidelity driving simulator, applying a Bayesian analytical approach. The results present a mixed picture, with both absolute and relative validity observed for certain physiological and cognitive stress indicators. Overall, our findings suggest that stress responses in the simulator and real-world driving are comparable, although the simulator was subjectively perceived as more stressful.

These results provide valuable insights into the validity of simulators for stress research and underscore the need to consider individual differences, experimental conditions, and methodological approaches in future studies.

Keywords: Driving Simulator Validation, Physiological Measurement, Stress Measurement, Cognitive Workload, Galvanic Skin Response, Electrocardiogram, Salivary Cortisol

1 Introduction

The automotive industry has advanced rapidly over the past decades, with significant advancements particularly in the areas of driving automation and electrification. Today, purchasing decisions are often influenced not only by the technical capabilities of a vehicle, but the user experience has become increasingly relevant. Customer centricity is a key element in designing systems that reflect the expectations, attitudes, and behaviors of users. According to the User-Centered Design Process (ISO 13407), users must be incorporated into the development process at very early stages. User studies are key to this end.

Driving simulators are commonly used to enable user studies due to several advantages. User behavior can be studied in an inherently controllable and safe test environment here (Caird & Horrey, 2016; Winter et al., 2012), which is often not possible on real roads. Furthermore, driving simulators allow testing at early development stages, relieving the requirement for physical prototypes (Xue et al.,

2023). To derive meaningful insights about user behavior in the real world, however, it must be guaranteed that results achieved in the simulator can be transferred to the real world. The matter at hand is understood as driving simulator validity.

Driving simulator validity has been subdivided into different constructs. The literature distinguishes between physical and behavioral validity (e.g., Bella et al., 2014). While physical validity describes the alignment of the simulator with a real car (Klüver et al., 2016), behavioral validity concerns the correspondence of driver behavior. Behavioral validity has been further subdivided into absolute and relative validity (Blaauw, 1982). Absolute validity is given when the numerical observation values in both environments are identical (Blaauw, 1982; Kaptein et al., 1996). Relative validity exists when the effects in the simulator take the same direction as in the real world (Blaauw, 1982). Behavioral validity has been suggested to be the more important quality compared to physical validity (Blaauw, 1982; Blana, 2001; Godley et al., 2002; Terumitsu et al., 2007).

To examine the validity of a simulator, validation studies are usually performed in which relevant outcome variables are compared between corresponding simulator and real-world drives (Klüver, 2016; I. M. Zöller, 2015). Validity hereby depends on the use case (Ahlström et al., 2012; Bella, 2008; Engen, 2008; Parduzi, 2021; Wynne et al., 2019), the outcome variables of interest (Himmels, Venrooij, et al., 2024; Wynne et al., 2019), and the simulator (Fischer et al., 2015; Himmels, Venrooij, et al., 2024). In a recent systematic literature review, Wynne et al. (2019) identified 44 studies directly comparing simulator and real-world driving. While the considered outcome variables largely varied across these 44 studies, the vast majority of studies considered driver output variables, such as speed or speed variation (21 studies), lane position or variation in lane position (13 studies), line crossing and lane change behavior (four studies), or overall driving performance and errors (10 studies). Few studies considered outcome variables underlying the observed driving behavior.

This is understandable, as variables underlying behavior are naturally more difficult to observe and interpret than driver behavior directly. However, several authors also noted the requirement to consider variables underlying behavior. The correspondence of perception between the driving simulator and the real world, for instance, is frequently mentioned (Blana, 2001; Boer, 2000). The simple idea here is that if the perception in the simulator corresponds to that in reality, the same driver's behavior should result. Vienne et al. (2014) suggested the term psychological validity, referring to the correspondence of processes underlying behavior.

Generally, different sensoric inputs can produce the same driver behavior (Espié et al., 2005). Meanwhile, perceptual biases can distort driver behavior (Espié et al., 2005). If existing biases in perception are disregarded or even exploited, the risk is that this will have unforeseeable effects on variables other than the particular considered ones, which ultimately leads to invalid results. In fact, invalid results occur frequently and the causes for this can often not be conclusively clarified. Taking into account variables underlying driver behavior could contribute to a better understanding of invalid outcomes and, in the long run, to closing the gap between reality and simulation.

Addressing the current research gap, stress will be considered in the present study. Stress is understood as a physiological and cognitive response to situations where a discrepancy is perceived between one's own capabilities and the external demands of a task (Cannon, 1929; Koolhaas et al., 2011; Selye, 1950, 1978, 1983; Zhou et al., 2022). Stress is typically categorized into positively perceived eustress and negatively perceived distress (Lazarus, 1966; Selye, 1976), with the latter being the predominant form in simulator studies (e.g., Daviaux et al., 2020; Matthews et al., 1998; Perello-March et al., 2022). In the context of a driving task, stress is defined as a situation perceived as challenging or dangerous (Francis, 2018; Gulian et al., 1989; Healey & Picard, 2005; Zhong et al., 2022).

While a driving task in a test is objectively the same for all participants, it can be experienced and evaluated differently depending on the individual's predisposition. This situational experience of stress can influence driver behavior and perception. Stressed drivers are more likely to make incorrect decisions (Kontogiannis, 2006; Westerman & Haigney, 2000).

2. Theoretical background

2.1 Validation studies related to variables underlying behavior

Table 1 provides an overview of previous physiological and cognitive validation studies investigating the validity of driving simulators. Notably, the sample sizes in most studies are relatively small, limiting the generalizability of the findings.

Table 1. Previous validation studies with a focus on physiological and/or cognitive indicators; Legend: EEG = Electroencephalogram; HR = Heart Rate; HRV = Heart Rate Variability; SCL = Skin Conductance Level

Authors	Validation	Variables	n	Validity?
Johnson et al. (2011)	Physiological	HR; Oxygen Consumption; Ventilation	9	None for HR Absolute & Relative for Oxygen Consumption and Ventilation
Mueller (2015)	Physiological	HR; HRV (Not specified); SCL; Pupil Diameter; Gaze-related Variables	34	Relative for HR, HRV, Gaze-related Variables None for SCL, Pupil Diameter
Reimer and Mehler (2011)	Physiological	HR; SCL	26	Absolute & Relative for HR Relative for SCL
Fors et al. (2013)	Physiological	EEG; Blink Duration, ECG; Gaze-related Variables	20	None (only blink data and gaze data reported)
Milleville-Pennel and Charron (2015)	Physiological	HR	14	None

Lobjois et al. (2021)	Physiological	Blinkrate	24	Relative
Li et al. (2013)	Physiological	EEG; HR	15	Absolute
Carter and Laya (1998)	Physiological	Scan Paths	16	Relative
Mueller (2015)	Cognitive	NASA-TLX	34	Relative (not for all items)
Lobjois et al. (2021)	Cognitive	NASA-TLX	24	Relative
Milleville-Pennel and Charron (2015)	Cognitive	NASA-TLX; Questionnaire of Psychological Feeling (QPF)	14	Relative for NASA-TLX (not for all items) Relative for QPF (not for all items)
Diels et al. (2011)	Cognitive	NASA-TLX	10	Relative (not for all items)
Galante et al. (2018)	Cognitive	Rotated Figures Task (RFT); NASA-TLX; SSSQ	100	Relative for NASA-TLX (Sumscore) Relative for all Dimensions of SSSQ; Absolute for Distress (SSSQ)

The overview reveals substantial variation in the validity results: some studies report relative or absolute validity for specific parameters (e.g., Johnson et al., 2011, for oxygen consumption), while others fail to demonstrate validity (e.g., Fors et al., 2013, for EEG). Additionally, there is a strong focus on isolated road segments or specific traffic contexts, without addressing a broader range of driving situations such as rural, urban, and highway driving. Johnson et al. (2011) identified this as a critical research gap that has been inadequately addressed in previous studies.

2.2 Measurement of stress

Stress responses can be assessed both physiologically (Perello-March et al., 2022) and cognitively (de Witte et al., 2021). Among the most frequently used physiological indicators are the Galvanic Skin Response (GSR; Bitterman & Holtzman, 1952; Sharma & Gedeon, 2012; Shi et al., 2007) and Electrocardiogram (ECG; Lohani et al., 2019). GSR measures the skin's electrical conductivity, which is influenced by minute sweat secretion (Boucsein, 2012; Giorgos Giannakakis et al., 2022). This sweating, referred to as arousal sweating (Darrow, 1933; Wilcott, 1967), is linked to stimuli that are novel, intense, and emotionally charged (Dawson et al., 2011). The intensity of emotional arousal triggers sweat gland activity (Kyriakou et al., 2019), making the skin more conductive and promoting electrical current flow (Caruelle et al., 2019; Navea et al., 2019; Stern et al., 2001). The activity of the eccrine sweat glands, one of three types of sweat glands, is measured and is solely innervated by the sympathetic nervous system (SNS; Critchley, 2002; Setz et al., 2010). The SNS is responsible for fight or flight responses, which is why GSR measurements exclusively reflect sympathetic activation, with no recording of relaxation responses regulated by the parasympathetic system (Fowles, 1986; Poh et al., 2010). Therefore, changes in GSR clearly indicate arousal and physiological preparation for stress (Boucsein, 2012;

McCorry, 2007; Norman et al., 2016). However, it should be noted that GSR activity, initially understood as a measure of arousal, only becomes a stress indicator in the context of a stressful situation (Healey & Picard, 2005; Labbé et al., 2007). GSR measurements are frequently used in automotive research (Caruelle et al., 2019).

The GSR measurement can be divided into two main components: tonic level (skin conductance level (SCL)) and phasic response (skin conductance response (SCR); Andreassi, 2010; Boucsein, 2012). SCL represents the slowly changing trend in skin conductivity and is calculated as an average over a specific period (Boucsein, 2012; Sharma & Gedeon, 2012), whereas SCR is a time-specific response to a particular stimulus. This reaction is reflected in a sudden increase in skin conductivity, known as GSR peaks, which represent short-term arousal. A related parameter, Peak Amplitude (PA), measures the magnitude of a GSR peak, indicating the intensity of the physiological response to a stimulus (Boucsein, 2012; Giannakakis et al., 2022).

ECG measurement allows the recording of electrical impulses generated by the autonomic heart rate (Shaffer et al., 2014), enabling the calculation of heart rate (HR) and heart rate variability (HRV), essential parameters for describing heart activity and frequently used as stress indicators (Andreassi, 2010; Cacioppo et al., 2017; Giannakakis et al., 2022).

Heart activity is regulated by the autonomic nervous system (ANS), which includes both sympathetic and parasympathetic nerves. The parasympathetic system reduces heart rate (Giorgos Giannakakis et al., 2022; Hall & Hall, 2020), while sympathetic nerves increase heart rate, enhancing blood flow and oxygen supply, preparing the body for a fight or flight response (Giannakakis et al., 2022; Hall & Hall, 2020). This means that during stress, the SNS dominates, leading to increased heart activity, improving blood circulation, and preparing the organism for heightened energy demand (Andreassi, 2010; Engert et al., 2014; Hall & Hall, 2020; Selye, 1950; Sharma & Gedeon, 2012).

In terms of parameters, HR refers to the number of heart beats per minute. Previous findings indicate that stress significantly increases heart rate (Engert et al., 2014; Giannakakis et al., 2017; Reinhardt et al., 2012). HR is a simple and widely used parameter for measuring the arousal state and physiological response to stress (Reinhardt et al., 2012; Taelman et al., 2011).

In contrast, HRV (Berntson et al., 2008) examines fluctuations in the time intervals between successive heartbeats (Electrophysiology, 1996). HRV is assessed using the mean RR-Interval, which measures the time between two R-peaks in the heart rate in milliseconds (RR-Int), as well as the Root Mean Square of Successive Differences (RMSSD) and the Standard Deviation of the intervals between R-peaks (SDNN) (Hall & Hall, 2020). While HR is an indicator of arousal, increasing with higher levels of stress (Reinhardt et al., 2012), HRV tends to decrease as internal arousal increases (Bernardi et al., 2000).

In addition to the physiological measurements of GSR and ECG, a stress reaction can also be assessed biologically. When a person experiences stress, two primary pathways in the body are activated: the sympathetic adrenal medullary system (SAM) and the hypothalamus-pituitary-adrenal axis (HPAA; (Andreassi, 2010; Reinhardt et

al., 2012). Activation of the HPAA triggers the release of the hormone corticotropin, which in turn stimulates the release of adrenocorticotropic hormone (ACTH). This leads to the release of the stress hormones adrenaline, noradrenaline, and cortisol from the adrenal glands. These hormones increase blood sugar levels, providing energy to the body in stressful situations (Chrousos, 2009), and a rise in cortisol levels is considered a direct indicator of stress (Dickerson & Kemeny, 2004). Cortisol can be measured in saliva (Salivary cortisol; SCT).

2.3 The present study

In the present study, a multi-method approach was employed to measure stress, including cognitive measures and objective physiological indicators. Experiments will be conducted in the driving simulator and the real world, employing a diverse mix of traffic scenarios (urban, rural, highway) that realistically reflect the situational cognitive demands of various driving situations. The study is aimed at determining to what extent driving simulators can replicate the driving experience in the real world, considering physiological and cognitive stress indicators.

We hypothesize:

• Physiological stress indicators correspond between the simulator and the real world (**H1**). We considered both absolute (**H1a**) and relative validity (**H1b**).

• Cognitive stress indicators correspond between the simulator and the real world (**H2**).

3. Method

3.1 Participants

A total of 72 participants were recruited. After excluding incomplete datasets due to device malfunctions, the final analysis included 68 participants. The sample consisted of 39 females (54.20 %) and 33 males (45.80 %) aged between 18 and 63 years (M = 30.07, SD = 11.58).

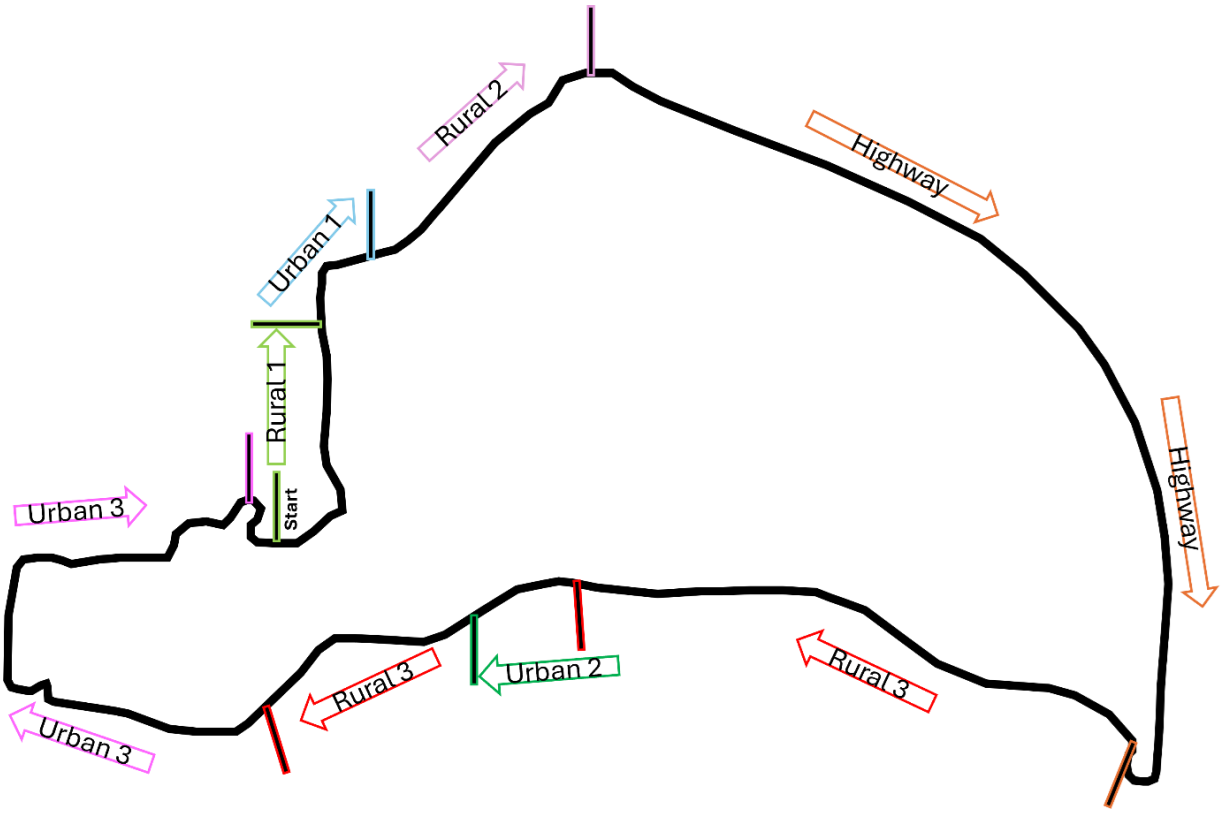
Regarding the participants' residential backgrounds, 43.1 % identified as living in rural areas, 44.4 % in small to medium-sized towns, and 12.5 % in urban environments.

Data collection was conducted using a convenience sampling approach with quotas based on age and gender to ensure diversity. The study employed a within-subject design and was conducted in the third quarter of 2023. Participants received compensation in the form of a travel reimbursement of 30 euros for their participation.

3.2 Route and Simulator

A circular route spanning approximately 23 kilometers was chosen for the study, comprising seven distinct test sections designed to include urban driving (5.3 km), rural roads (9.7 km), and highway driving (8 km). This segmentation was intentionally structured to reflect diverse environmental conditions, each imposing unique demands on the driver and influencing both physiological and cognitive load. The sequence of the individual segments was as follows: Rural 1, Urban 1, Rural 2, Highway, Rural 3, Urban 2, Urban 3.

Figure 1. Abstract route with segment subdivision



A circular route spanning approximately 23 kilometers was chosen for the study, comprising seven distinct test sections designed to include urban driving (5.3 km), rural roads (9.7 km), and highway driving (8 km). This segmentation was intentionally structured to reflect diverse environmental conditions, each imposing unique demands on the driver and influencing both physiological and cognitive load. The sequence of the individual segments was as follows: Rural 1, Urban 1, Rural 2, Highway, Rural 3, Urban 2, Urban 3.

Rural 3 is briefly interrupted by Urban 2, which consists of a small urban section belonging to a residential district that intersects the rural road.

All participants followed the same route in the same sequence (Figure 1).

The naming of the segments (e.g., Rural 1–3) was based on their chronological appearance along the route and reflects the classification of the road type at that point (e.g., rural, urban, or highway), not geographical proximity or functional differences. The naming follows the actual course layout.

The route was designed as a closed loop to ensure both practical feasibility and a high degree of situational variety. This approach made the course suitable for accurate replication in the driving simulator. The high situational variability was intended to help assess which types of driving environments are more or less suited for simulation.

Figure 2. Real (top) and simulated (bottom) sections

By including various driving sections, driver stress can be evaluated in diverse driving scenarios. Weather conditions during the real-world study were favorable, with smooth traffic flow throughout the testing period.

The driving task was intentionally designed as a regular, non-manipulated drive through real traffic environments to capture naturally occurring stress responses. The use of stress inducing methods was deliberately avoided due to ethical and safety considerations associated with real-world traffic.

The simulated route was programmed using Silab 7.1, a driving simulation software developed by the Würzburg Institute of Traffic Sciences (WIVW). SILAB is a professional simulation environment that allows the realistic replication of driving routes and complex traffic scenarios. The software is not publicly available, but further information can be found at: https://wivw.de/en/silab-2/

The virtual route was implemented as a digital twin of the real-world test course. Reconstruction was based on OpenStreetMap data and preserved original dimensions and topographical features. Figure 2 illustrates a comparison between the real and virtual versions of the route.

A medium-fidelity driving simulator (as defined by Wynne et al., 2019) was used for the simulator study (Figure 3). The simulator featured:

- An original driver's seat,
- A force-feedback steering wheel, pedals, turn signals, and dashboard,
- A mockup mounted on a D-Box system with 3 degrees of freedom, simulating road surface feedback, and

• Three 55” LCD screens providing a 180° horizontal field of view, offering an immersive driving experience.

The D-Box system is a professional motion platform using hydraulic actuators to simulate physical road feedback (https://www.d-box.com/en#tab_2). It provides motion cues such as vibrations or tilting to replicate road texture, acceleration, and braking. For example, during braking, the rear actuators lift slightly, pushing the driver forward to mimic real vehicle deceleration. This enhances the sense of realism during simulated driving.

Figure 3. Medium fidelity driving simulator used for the study

3.3 Measurements

3.3.1 Physiological measurements

On the physiological measurement side, we selected GSR measurements (including SCR, SCL, PA) as well as ECG measurements (HR and HRV). These measures have been established as reliable stress indicators in prior driving simulation studies (e.g., Daviaux et al., 2020; Manseer & Riener, 2014; Milardo et al., 2022; Scherz et al., 2023).

To capture the biological stress response, we employed salivary cortisol tests. Cortisol reflects Hypothalamic-Pituitary-Adrenal (HPA) axis activity and is considered a direct hormonal stress indicator (Kirschbaum & Hellhammer, 1994), whereas elevated HR and GSR activity serve as situational markers of stress (Lohani et al., 2019).

Concerning GSR, we collected the skin conductance response (SCR; peaks per minute), peak amplitude (PA), and the tonic skin conductance level (SCL). Measurements were conducted using the exosomatic method with direct current (Boucsein et al., 2012), utilizing a Shimmer 3 GSR+ sensor. Two electrodes were placed on the palm of each participant.

For ECG measurements, we recorded standard parameters, including heartbeats per minute (heart rate, HR) and various heart rate variability (HRV) metrics. These included the mean RR-interval (RR-Int), the root mean square of successive differences (RMSSD), and the standard deviation of NN intervals (SDNN).

Physiological responses were recorded using iMotions software (version 9.4).

Salivary cortisol (SCT) was collected using Salivettes. Sample analysis was carried out by Dresden Labservice GmbH Saliva samples were frozen and stored at -20 degrees Celsius until analysis. After thawing, samples were centrifuged at 3,000 rpm for 5 min, which resulted in a clear supernatant of low viscosity. Salivary concentrations were measured using commercially available chemiluminescence immunoassay with high sensitivity (Tecan - IBL International, Hamburg, Germany; catalogue number R62111). The intra and interassay coefficients of variance were below 9%.

Due to high inter-individual variability of GSR (SCR is more robust than SCL) and cortisol values (Boucsein, 2012; Hellhammer et al., 2009), interpretation of absolute levels is limited. Consequently, our analyses focus on intra-individual changes within a within-subject design. Each participant serves as their own reference point, enabling detection of relative stress reactivity across driving conditions.

3.3.2 Cognitive stress related variables

To assess the participants' cognitive responses, we measured various variables and constructs related to perceived stress using a semantic differential.

For measuring cognitive reactions, we employed an 11-point scale ranging from 0 (not at all) to 10 (very much). This scale is intuitive for participants to understand (Lewis, 2021), enhances data variance (Dawes, 2002), tends to produce normally distributed data (Leung, 2011), and facilitates the use of parametric tests (Chyung et al., 2018).

Table 2 presents the items used in the applied scales. The questions were translated from English to German by the authors. The self-formulated single-item measurements we developed were originally created in German and were translated to English for demonstration in this paper.

According to the Short Stress State Questionnaire, which consists of 24 items measuring three subdimensions—Engagement, Distress, and Worry—we shortened the questionnaire for our study by selecting the four items with the highest factor loadings for each dimension, as suggested by Helton (2004).

Table 2. Scales and single items used and their translation

Scale (Abbr.)	English Wording	German Wording	Source
Vehicle Operation (VO)	• How well did you manage operating the vehicle?	• Wie gut sind Sie mit der Bedienung des Fahrzeuges zurechtgekommen ?	Self-developed Single Item
Physical Wellbeing (PW)	• How was your physical well-being during the ride?	• Wie war Ihr körperliches Wohlbefinden	Self-developed

		während der Fahrt?	Single Item
STRESS	• To what extent did you experience stress during the drive?	• In welchem Ausmaß haben Sie während der Fahrt Stress empfunden?	Self-developed Single Item
Short State Stress Questionnaire (SSSQ)	1. I feel dissatisfied 2. I am committed to attaining my performance goals 3. I'm trying to figure myself out 4. I feel impatient 5. I am motivated to do the task 6. I'm reflecting about myself 7. I feel angry 8. I feel confident about my abilities 9. I feel concerned about the impression I am making 10. I feel irritated 11. Generally, I feel in control of things 12. I thought about how others have done on this task	1. Ich bin unzufrieden. 2. Ich bin entschlossen, meine Leistungsziele zu erreichen. 3. Ich versuche, mir selbst auf die Spur zu kommen. 4. Ich empfinde Ungeduld. 5. Ich bin motiviert, die Aufgabe zu erledigen. 6. Ich denke über mich selbst nach. 7. Ich bin wütend. 8. Ich bin mir sicher über meine Fähigkeiten. 9. Ich frage mich, welchen Eindruck ich hinterlasse. 10. Ich fühle mich irritiert. 11. Im Allgemeinen habe ich das Gefühl, die Dinge im Griff zu haben. 12. Ich überlege, wie andere bei dieser Aufgabe abschneiden.	Helton, 2004
NASA-TLX	1. How mentally demanding was the task? (Mental Demand) 2. How physically demanding was the task? (Physical Demand) 3. How hurried or rushed was the	1. Wie viel geistige Anforderung war bei der Fahrt erforderlich? (Mental Demand) 2. Wie viel körperliche Anforderung war bei der Fahrt	Hart, 2006; Hart & Staveland, 1988

pace of the task? (Temporal Demand)
4. How successful were you in accomplishing what you were asked to do? (Performance)
5. How hard did you have to work to accomplish your level of performance? (Effort)
6. How insecure, discouraged, irritated, stressed, and annoyed were you? (Frustration)

erforderlich? (Physical Demand)
3. Wie viel Zeitdruck empfanden Sie während der Fahrt? (Temporal Demand)
4. Wie zufrieden waren Sie mit Ihrer Leistung im Zusammenhang mit der Fahrt? (Performance)
5. Insgesamt betrachtet: Wie groß war die von Ihnen empfundene Anstrengung bei der Fahrt? (Effort)
6. Wie frustriert fühlten Sie sich während der Fahrt? (Frustration)

3.4 Procedure

At the beginning of data collection, the participant was welcomed, and the first salivary cortisol sample was collected (SCT/0). Subsequently, an introductory pre-survey was conducted to assess the participant's current stress state (SSSQ/0).

The next step involved attaching the physiological measurement devices to the participant.

Before the driving session, participants were informed that the route would be a circular course of approximately 25 minutes, consisting of different segments including urban, rural, and highway sections. The driving task itself was structured similarly to a driving school setup: participants were guided in real time by the experimenter, who gave timely navigation instructions (e.g., turn left, continue straight) throughout the entire drive. In the simulator there were also navigation arrows.

Following this, the participant accompanied the experimenter to the vehicle and completed the real-world driving session (GSR/1; ECG/1).

After completing the drive, a post-drive survey was conducted inside the vehicle. This survey assessed perceived stress (SSSQ/1; STRESS/1), perceived workload (NASA-TLX/1), situational strain related to vehicle operation (VO/1), and physical well-being (PW/1).

Upon returning to the laboratory, a second salivary cortisol sample was collected (SCT/1). Participants were then introduced to the driving simulator following the protocol outlined in the introduction package by Hoffmann et al. (2003). This introduction included three practice tracks (about 15 min) to familiarize participants with

the simulator. After the practice sessions, the placement of the physiological devices was checked to ensure they remained correctly attached. The participant then completed a simulated drive on a digital replica of the previously driven real-world route (GSR/2; ECG/2).

Upon completing the simulated drive, a post-simulation survey was conducted to capture cognitive responses (NASA-TLX/2; SSSQ/2; STRESS/2; VO/2; PW/2). Before the participant was dismissed, a final salivary cortisol sample was collected (SCT/3).

The total testing time per participant was approximately two hours, with an average driving duration of around 25 minutes for the real-world drive and 23 minutes for the simulated drive. The difference in duration between the real and simulated journey can be attributed to the traffic flow, which was standardized in the simulator (e.g. traffic lights) but could not be influenced on the real route. The order of the drives was fixed, with the real-world drive always preceding the simulated drive.

Before the start of the study, participants received a standardized oral briefing about the purpose, procedure and data protection regulations. Informed consent was obtained orally, in accordance with the approved procedure. The study was approved by the Ethics Committee of the University of Bayreuth. Participation was voluntary, and participants were informed that they could withdraw at any time without giving reasons.

3.5 Statistical analysis

Previous studies on simulator validation employed null-hypothesis significance tests (NHSTs), such as t-tests or regression analyses (Klüver et al., 2016; Losa et al., 2013; Törnros, 1998; I. Zöller et al., 2019). However, NHSTs can only identify effects. Non-significant results may result from either low statistical power or true equivalence, an important distinction that NHSTs cannot make. This limitation is particularly problematic in studies with small sample sizes.

To address these issues, we adopt a Bayesian approach (Himmels, Weigl, et al., 2024). In contrast to frequentist p, the Bayes factor BF10 is a relative indicator for the probability of H0 compared to H1. In this way, evidence can be provided not only for differences, but also for equivalence, which would indicate simulator validity.

We conducted Bayesian repeated-measures ANOVA using JASP (van Doorn et al., 2021) and predefined priors (Rouder et al., 2017) for Bayesian analyses.

Evidence from the Bayes Factor will be interpreted following (Jeffreys, 1998). Accordingly, a BF10 (or BFincl) > 3 is considered evidence for an effect, and a BF10 (or BFincl) < 0.3 is considered evidence for equivalence. Since Bayes Factors are relative indicators unlike p-values, they are informative even when they do not precisely follow these recommendations. If a measurement yields a BF10 (or BFincl) > 0.3 but < 1, this is referred to as anecdotal evidence of equivalence (van Doorn et al., 2021). A BF10 (or BFincl) of > 1 and < 3 signifies anecdotal evidence for an effect, which means that the evidence is considered weak or inconclusive, but still leans slightly in favor of the alternative hypothesis. In both cases, the term "anecdotal" reflects the limited strength of the statistical support, rather than anecdotal in a colloquial sense.

Furthermore, for the physiological variables, absolute validity was inferred from absence of a main effect of the environment (real vs. simulator), and relative validity from the absence of an Environment*Section interaction, as indicated by the corresponding BFincl values.

4. Results

Summarizing, SCT and the SSSQ were inquired before the real-world drive, after the real-world drive and after the simulator drive, GSR and ECG were inquired throughout the drives, and STRESS, the NASA-TLX, and PW were inquired after the real-world drive and after the simulator drive.

Bayesian dependent t-tests including the factor test environment (real vs. simulator) were conducted for cortisol, SSSQ, STRESS, NASA-TLX, VO, and PW. For this variables we can only consider absolute validity.

As GSR and ECG were recorded continuously throughout the drive, these were analyzed in a two-factor Bayesian repeated-measures ANOVA including the factors environment and Scenario. This approach allowed us to also consider relative validity for ECG and GSR.

Note that relative validity could only be assessed for continuously recorded physiological measures, as these allowed comparisons across driving sections. In contrast, the cognitive variables and de cortisol measures were only collected after the entire drive, thus preventing any within-drive section-level analysis.

Physiological

The physiological data (e.g., noise filtering, HRV calculation) was performed using R Notebooks in iMotions. As GSR and ECG were recorded continuously throughout the drive, mean scores for SCR, SCL, PA, HR, RR-Interval, RMSSD, and SDNN were calculated section-wise (Rural 1, Urban 1, Rural 2, Highway, Rural 3, Urban 2, Urban 3).

Prior to each driving environment (real and simulated), baselines were recorded for all physiological indicators (SCR, SCL, PA, HR, RR-Interval, RMSSD, and SDDN). For SCL (µS) (real: 12.08 simulator:13.92; p<.001), PA (µS) (real: 0.25 simulator: 0.33 ;p<.001) HR (bpm) (real:79.88 simulator: 77.30 ;p=.005) and RR-Interval (ms) (real: 777.44 simulator: 803.12 ;p=.007) small but statistically significant differences were observed, possibly indicating anticipators responses. However, since our primary analyses relied on within-subject comparisons across scenarios, the baseline differences do not confound the observed intra-individual physiological patterns.

For SCR, there is evidence for the absence of an effect of the test environment, indicating that values do not differ between the real world and the simulator (Table 3, Figure 4). Furthermore, there is evidence for the absence of an interaction between environment and section (Table 3).

Regarding SCL, there is evidence for a main effect of environment, indicating a significant difference between real and simulated driving, with higher SCL values in the simulator (Table 3, Figure 5). However, there is evidence for equivalence concerning the interaction between environment and section (Table 3).

For PA, the data indicate evidence of an effect of the test environment (Table 3). PA is higher in the simulator compared to the real world (Figure 6). Additionally, there is evidence for an interaction effect between environment and section (Table 3). Differences between the real world and the simulator are larger in Highway, Rural 3, and Urban 2, while there are only marginal differences in Urban 3 (Figure 6).

Regarding HR, there is evidence for both an environment effect and an interaction effect between environment and section (Table 3). HRs are higher in the real world compared to the simulator (Figure 7), while the differences are differently pronounced across the scenarios.

For RR-Interval, evidence suggests both an environment effect and an interaction effect between environment and section (Table 3). The RR-Intervals are higher in the simulator than the real car (Figure 8), with especially large differences in Rural 1.

For RMSSD, the data provide anecdotal evidence for equivalence across environments (Table 3; Figure 9). Additionally, there is moderate evidence for the absence of an interaction effect (Table 3).

Regarding SDNN, the data indicate evidence for equivalence across environments (Table 3; Figure 10). Furthermore, there is evidence for the absence of an interaction effect between environment and section (Table 3).

Since SCT was collected only after each complete drive (rather than after each section), only the environment effect could be examined here. The data provide evidence for the absence of an effect of the environment (Table 3; Figure 11).

Table 3. Statistical results for Bayesian ANOVAs, Post Hoc tests, and Bayesian paired t-test. Evidence for equivalence is marked green, evidence for effects is marked red.

SCR					
BANOVA	P(incl)	P(excl)	P(incl|data)	P excl|data)	BF_{incl}
Environment	0.600	0.400	0.125	0.875	0.095
Section	0.600	0.400	0.125	1.110×10^{-16}	$6.005 \times 10^{+15}$
Environment*Section	0.200	0.800	0.001	0.999	0.006
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	0.050	0.050	0.437

SCL					
BANOVA	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Environment	0.600	0.400	0.971	0.029	22.073
Section	0.600	0.400	0.968	0.032	20.288
Environment*Section	0.200	0.800	0.009	0.991	0.037
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	$4.765 \times 10^{+10}$	$4.765 \times 10^{+10}$	4.122×10^{-17}

PA					
BANOVA	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Environment	0.600	0.600	1.000	6.695×10^{-14}	$9.958 \times 10^{+10}$
Section	0.600	0.400	1.000	0.000	∞
Environment*Section	0.200	0.800	1.000	2.073×10^{-9}	$1.029 \times 10^{+9}$
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	$1.530 \times 10^{+21}$	$1.530 \times 10^{+21}$	7.542×10^{-28}

HR					
BANOVA	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Environment	0.600	0.400	1.000	0.000	∞
Section	0.600	0.400	1.000	0.000	∞
Environment*Section	0.200	0.800	1.000	1.315×10^{-11}	$3.043 \times 10^{+11}$
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	$2.974 \times 10^{+44}$	$2.974 \times 10^{+44}$	5.673×10^{-47}

RR-Int					
BANOVA	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Environment	0.600	0.400	1.000	2.309×10^{-9}	$2.887 \times 10^{+8}$
Section	0.600	0.400	1.000	2.662×10^{-4}	2504.136
Environment*Section	0.600	0.400	1.000	2.664×10^{-4}	15010.813
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	$7.347 \times 10^{+28}$	$7.347 \times 10^{+28}$	7.172×10^{-36}

RMSSD					
BANOVA	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Environment	0.600	0.400	0.325	0.675	0.322
Section	0.600	0.400	1.000	2.330×10^{-13}	$2.861 \times 10^{+12}$
Environment*Section	0.200	0.800	0.002	0.998	0.009
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	3.592	3.592	0.007

SDNN					
BANOVA	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Environment	0.600	0.400	0.214	0.786	0.181
Section	0.600	0.400	1.000	1.179×10^{-9}	$5.654 \times 10^{+8}$
Environment*Section	0.200	0.800	0.001	0.999	0.005
Post Hoc		Prior	Posterior	$BF_{10,U}$	error %
Comparison - Environment		Odds	Odds		
Real vs. Simulator		1.000	0.227	0.227	0.100

SCT Bayesian paired t-test (Factor: Environment)	BF_{10}	error%
	0.134	0.097

Figure 4. Descriptives for Skin Conductance Response (with 95% credible interval)

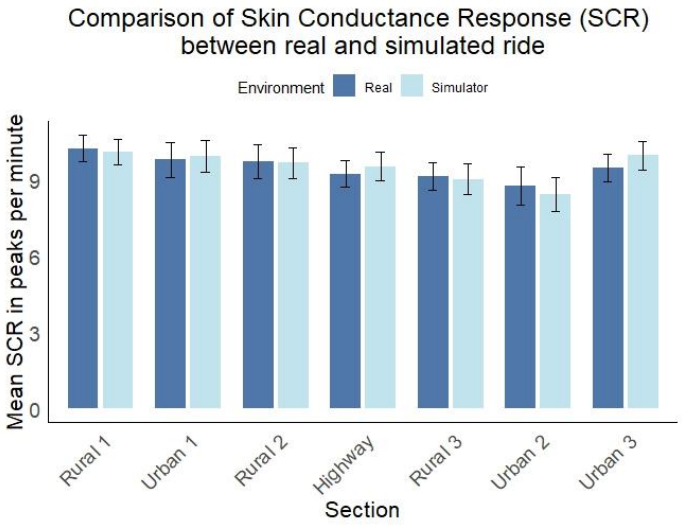


Figure 5. Descriptives for Skin Conductance Level (with 95% credible interval)

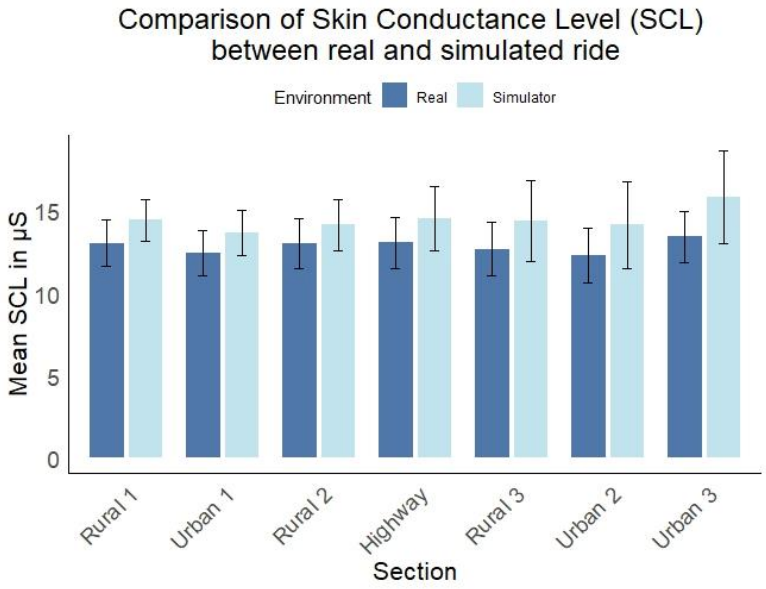


Figure 6. Descriptives for Peak Amplitude (with 95% credible interval)

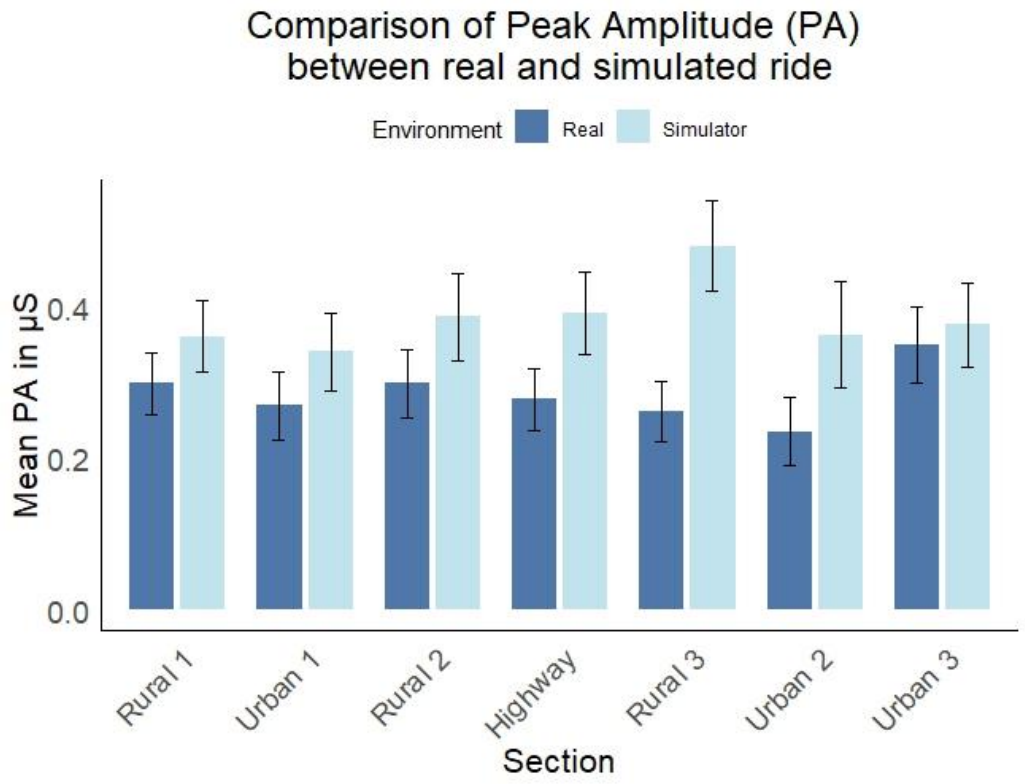


Figure 7. Descriptives for Heart Rate (with 95% credible interval)

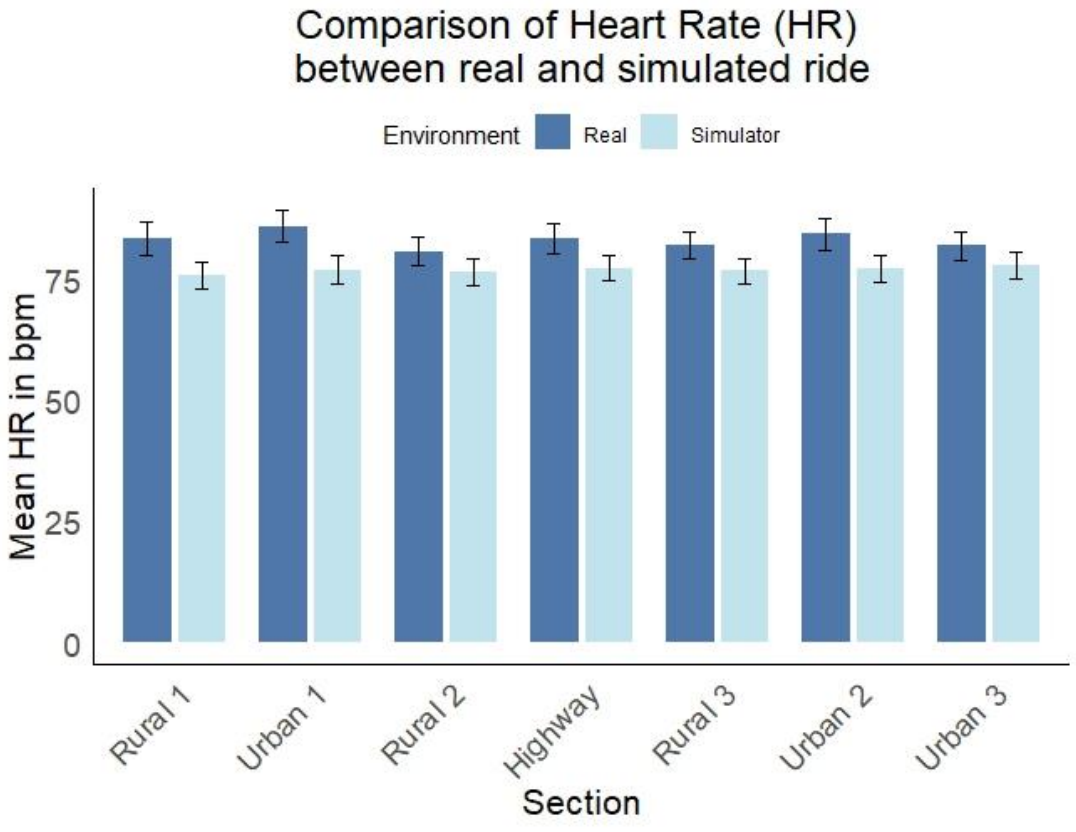


Figure 8. Descriptives for RR-Interval (with 95% credible interval)

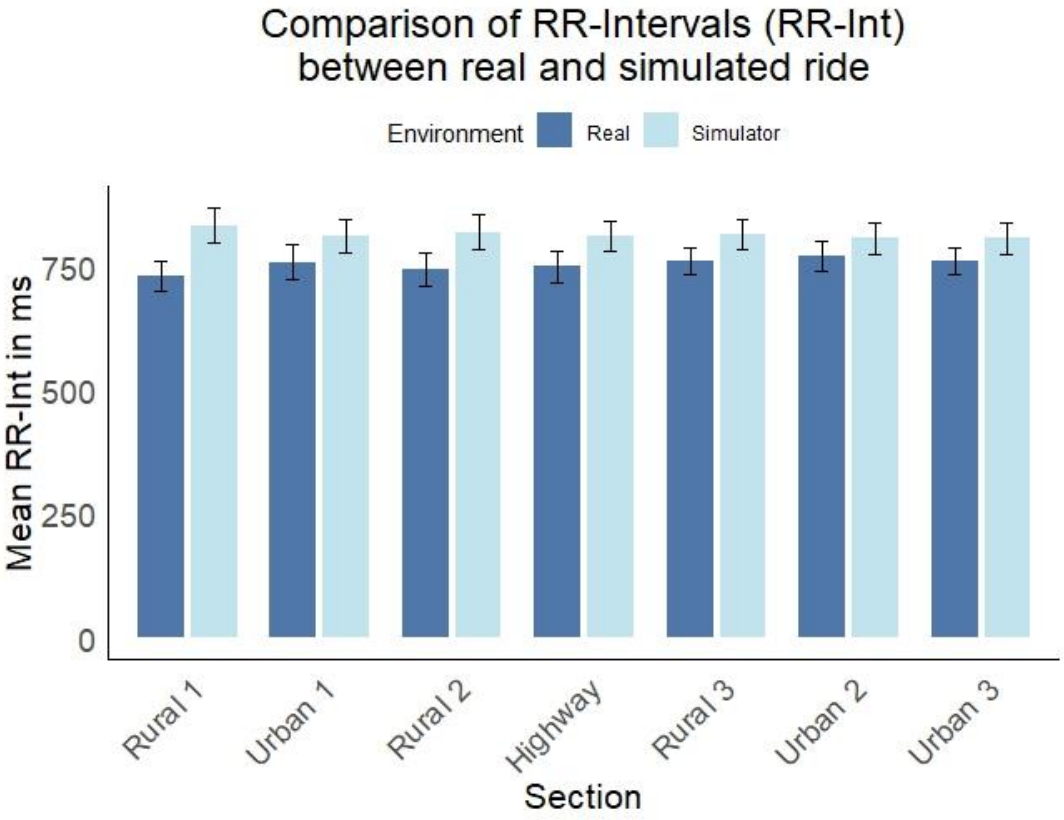


Figure 9. Descriptives for Root Mean Square of Successive Differences (with 95% credible interval)

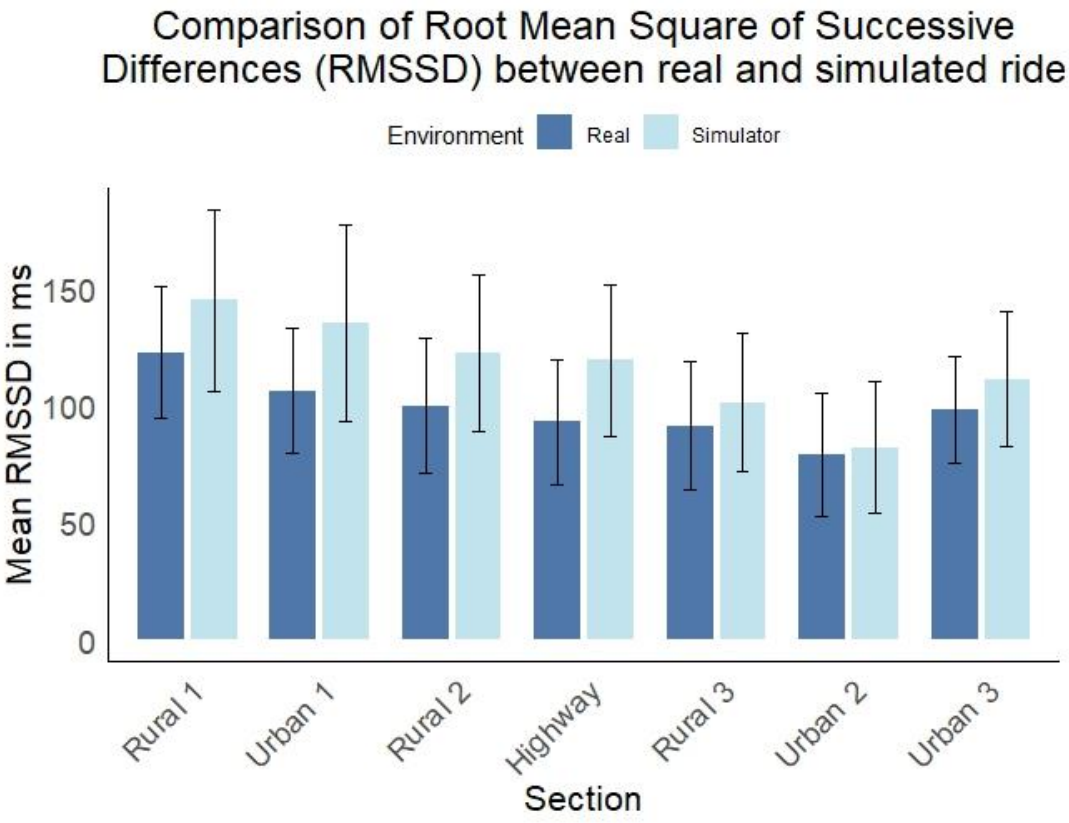


Figure 10. Descriptives for Standard Deviation of the NN Interval (with 95% credible interval)

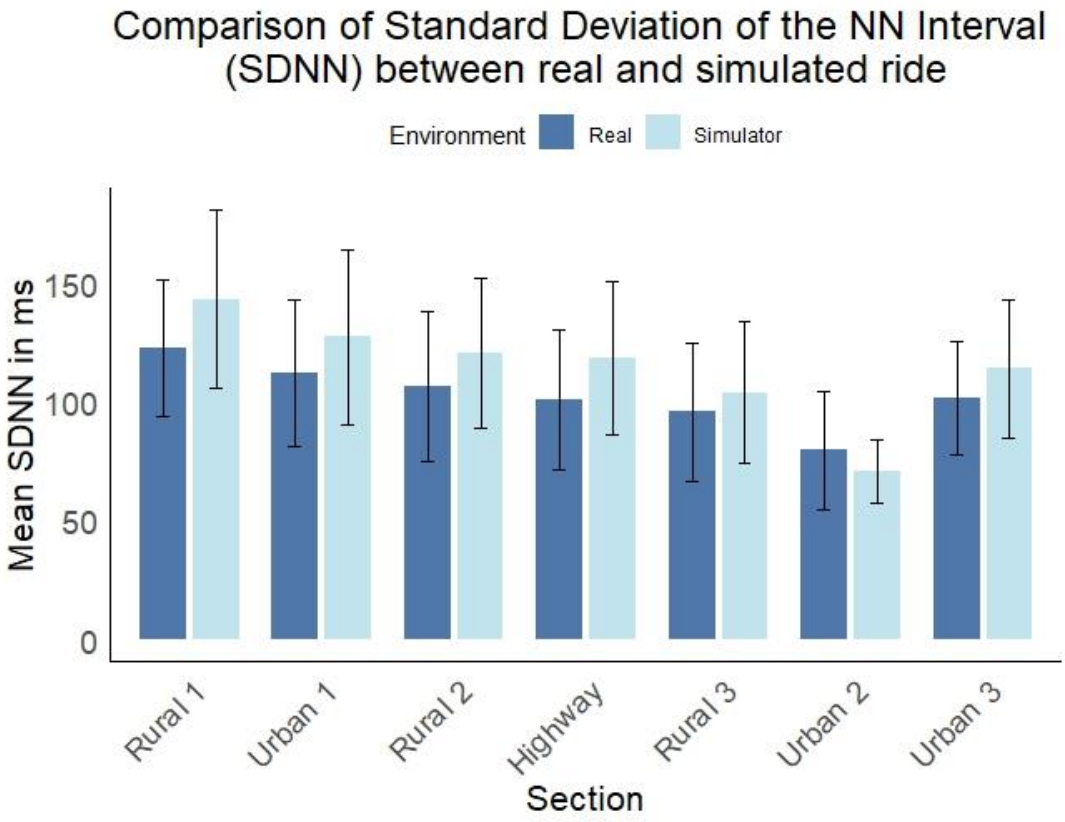
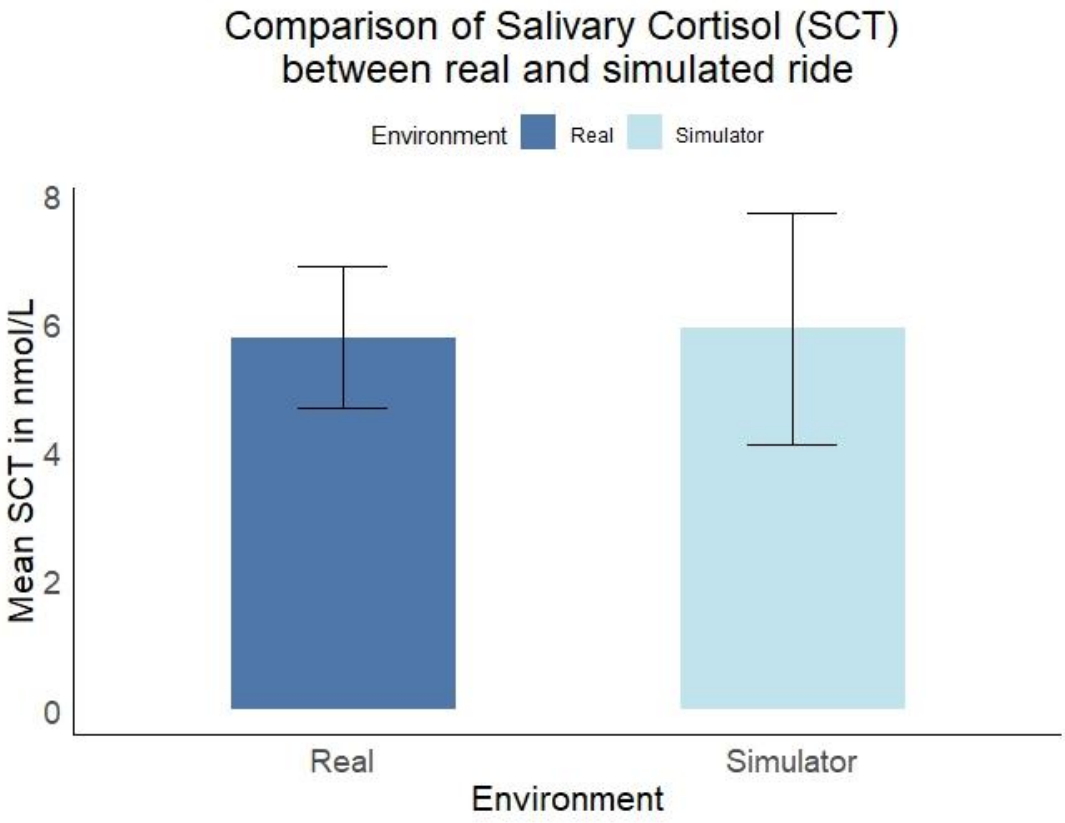


Figure 11. Descriptives for Salivary Cortisol (with 95% credible interval)



Cognitive

The subdimensions of the SSSQ (Worry, Engagement, and Distress) were calculated following Helton (2004). For the NASA-TLX, a mean index was calculated, showing an internal consistency of α = .72 (real) and α = .79 (simulator), respectively.

For VO, the NASA-TLX, Distress, Engagement, STRESS, and PW, there was evidence for an effect of the test environment (Table 4). VO and PW were rated higher in the real world compared to the simulator (Figures 12, 18). The NASA-TLX, Distress, Engagement, and STRESS achieved lower scores in the real world compared to the simulator (Figures 13,15-17). For Worry, there was evidence for equivalence across the real world and the simulator (Table 4; Figure 14).

Table 4. Statistical results for Bayesian paired t-test. Evidence for equivalence is marked green, evidence for effects is marked red.

Bayesian paired t-test (Factor: Environment)	BF_{10}	error %
VO	$4.269 \times 10^{+22}$	4.104×10^{-26}
NASA-TLX	$2.375 \times 10^{+12}$	2.427×10^{-18}
Worry	0.186	0.078
Distress	8450.038	5.213×10^{-11}
Engagement	36081.394	6.757×10^{-7}
STRESS	$4.088 \times 10^{+7}$	1.071×10^{-13}
PW	$1.505 \times 10^{+12}$	2.522×10^{-18}

Figure 12. Descriptives for Vehicle Operation (with 95% credible interval)

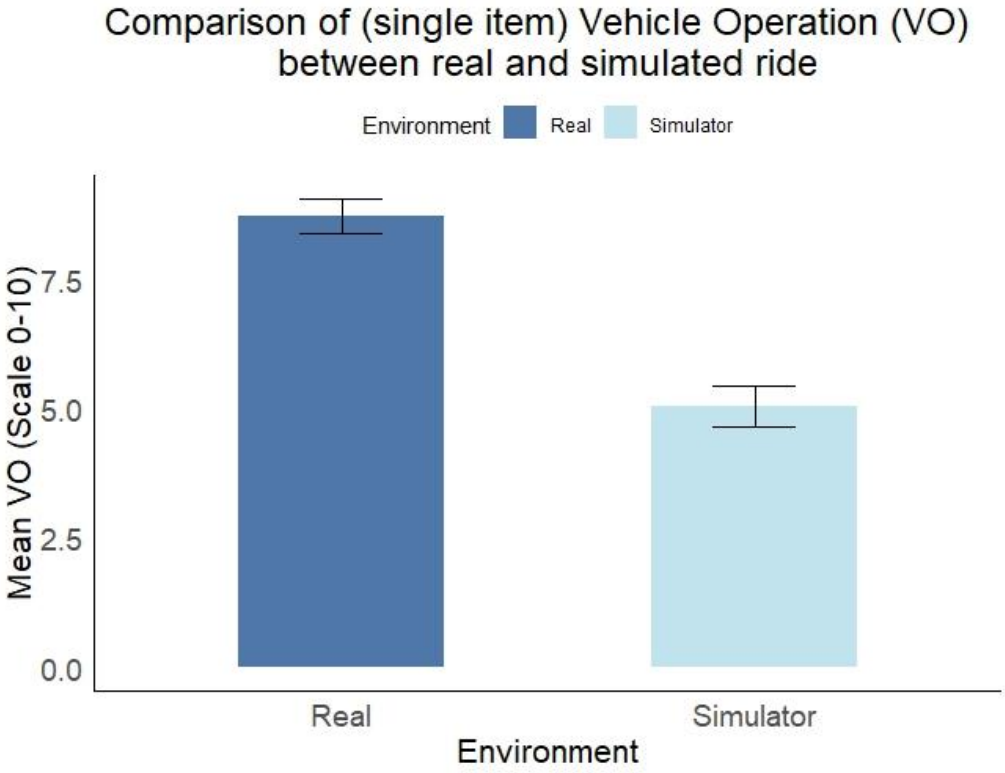


Figure 13. Descriptives for NASA Task Load Index (with 95% credible interval)

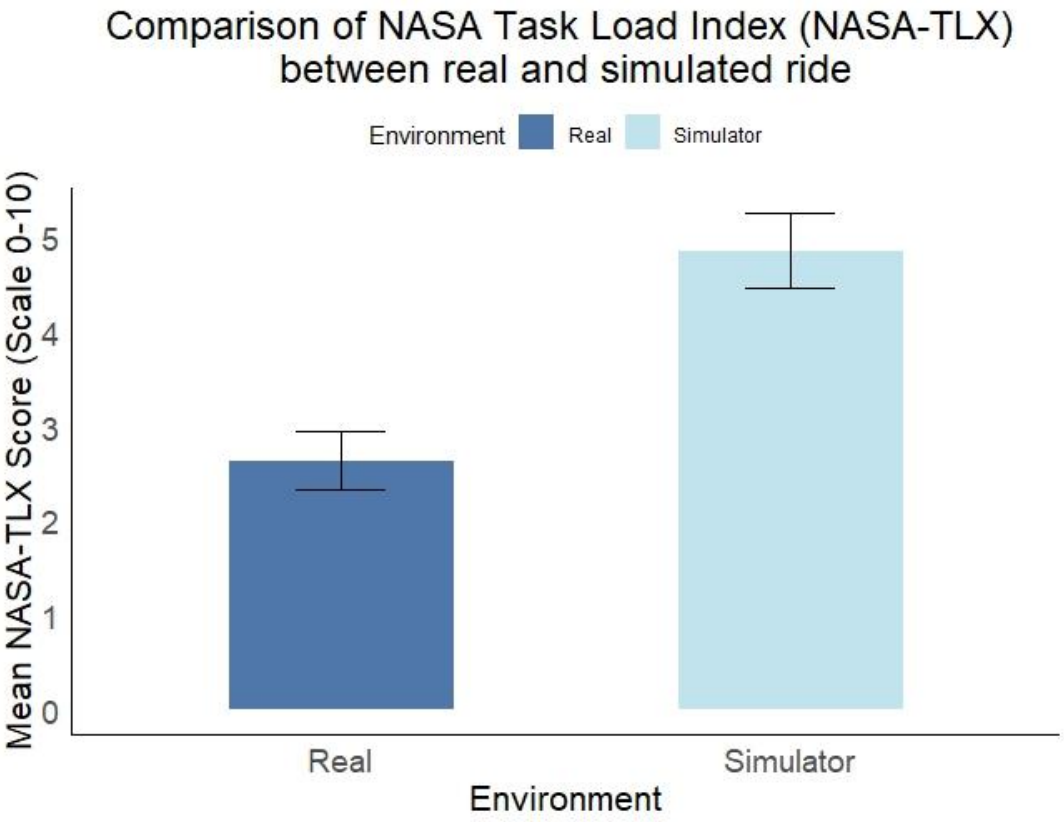


Figure 14. Descriptives for dimension Worry (SSSQ) (with 95% credible interval)

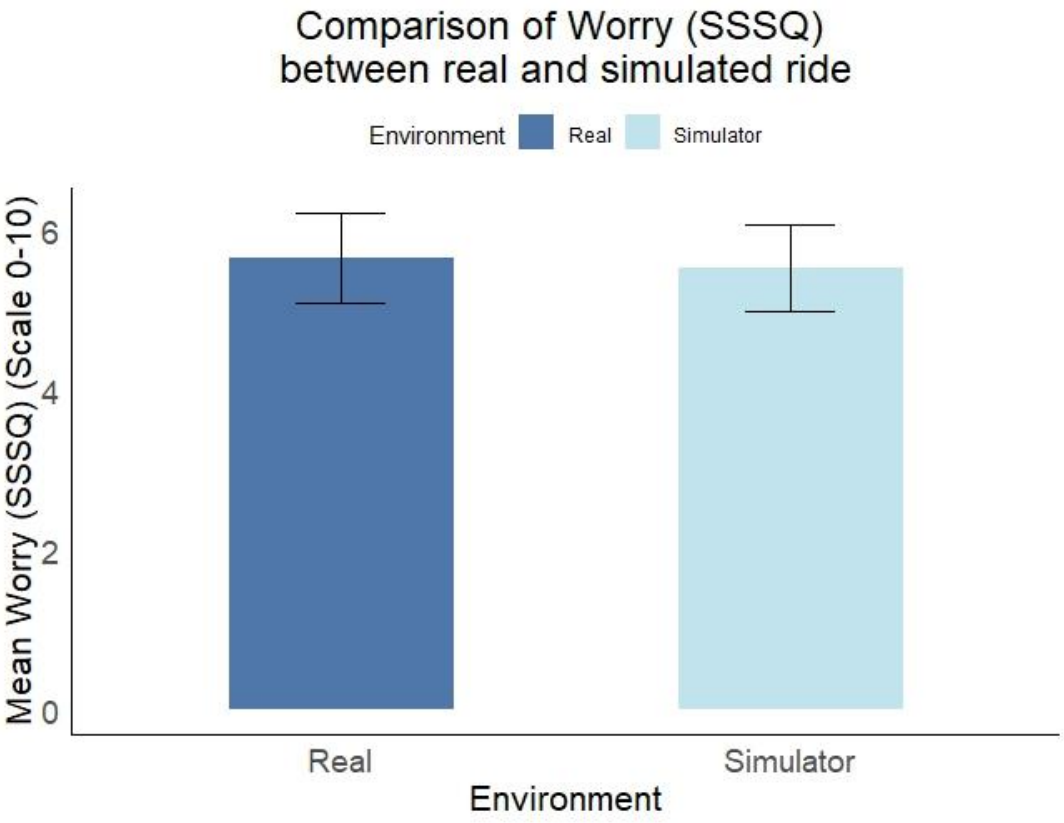


Figure 15. Descriptives for dimension Distress (SSSQ) (with 95% credible interval)

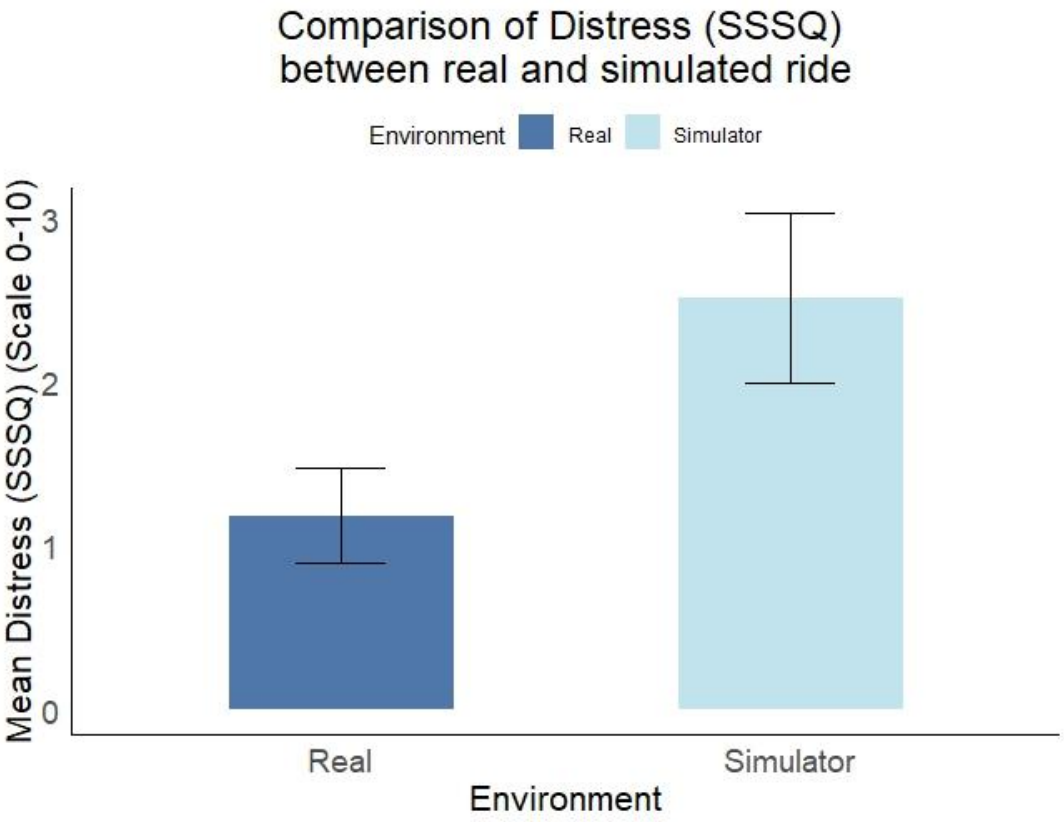


Figure 16. Descriptives for dimension Engagement (SSSQ) (with 95% credible interval)

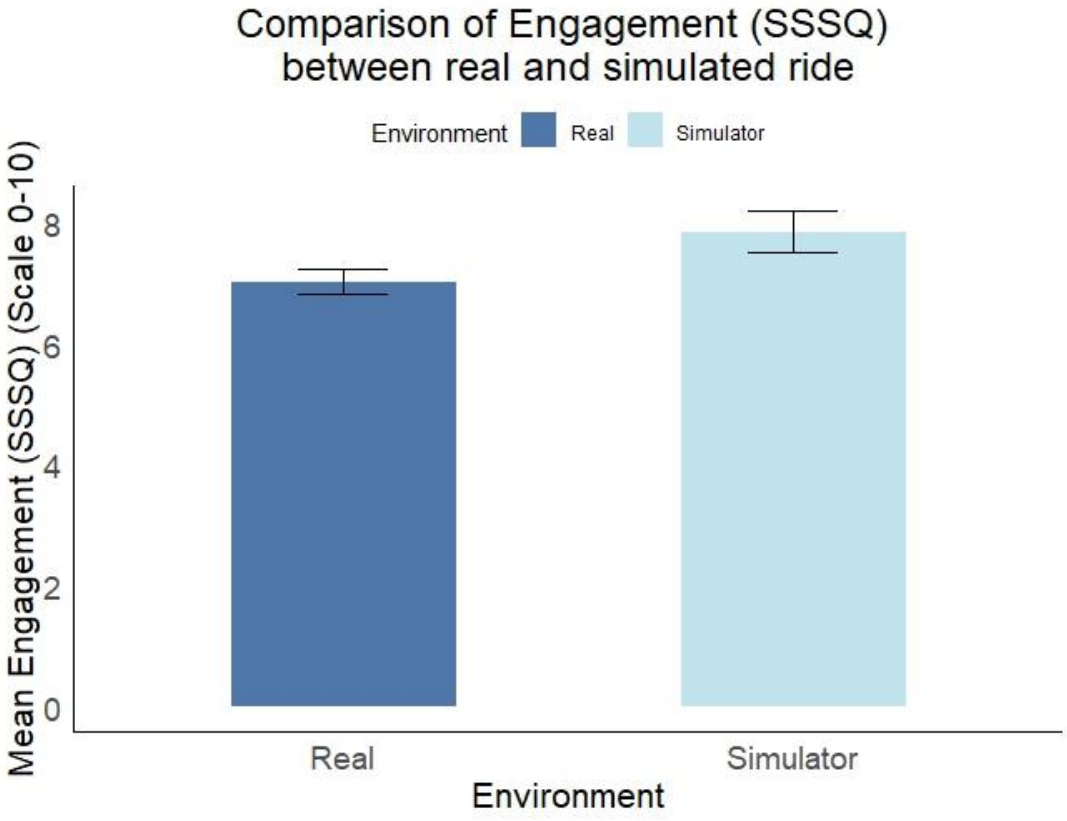


Figure 17. Descriptives for Stress (with 95% credible interval)

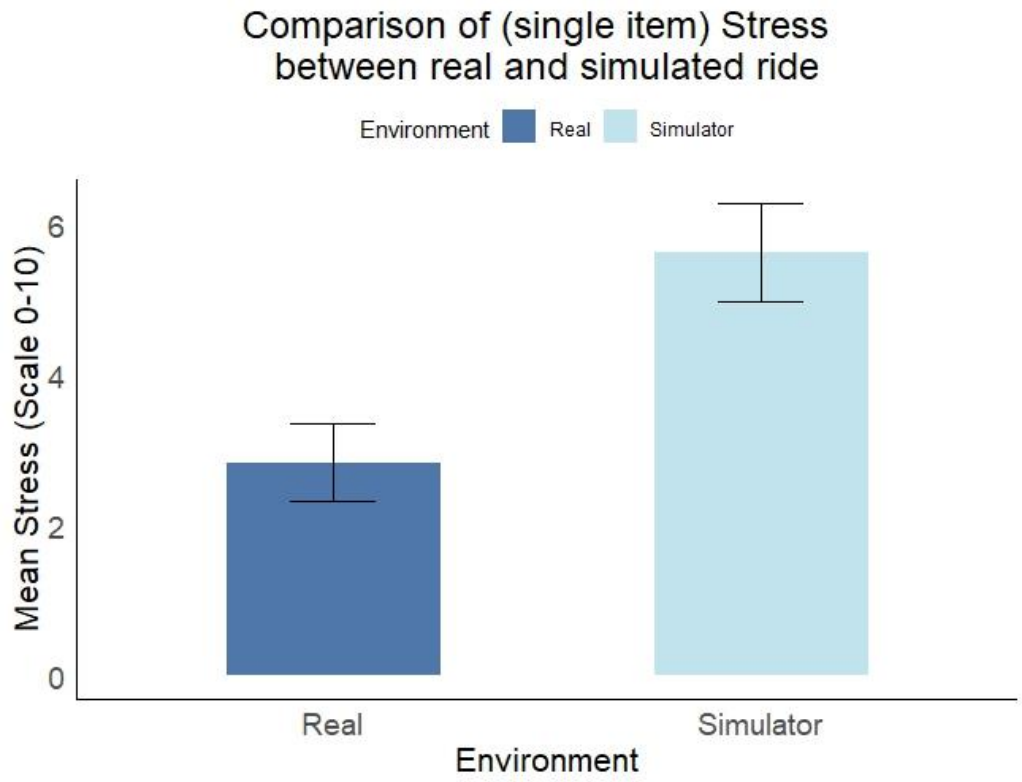
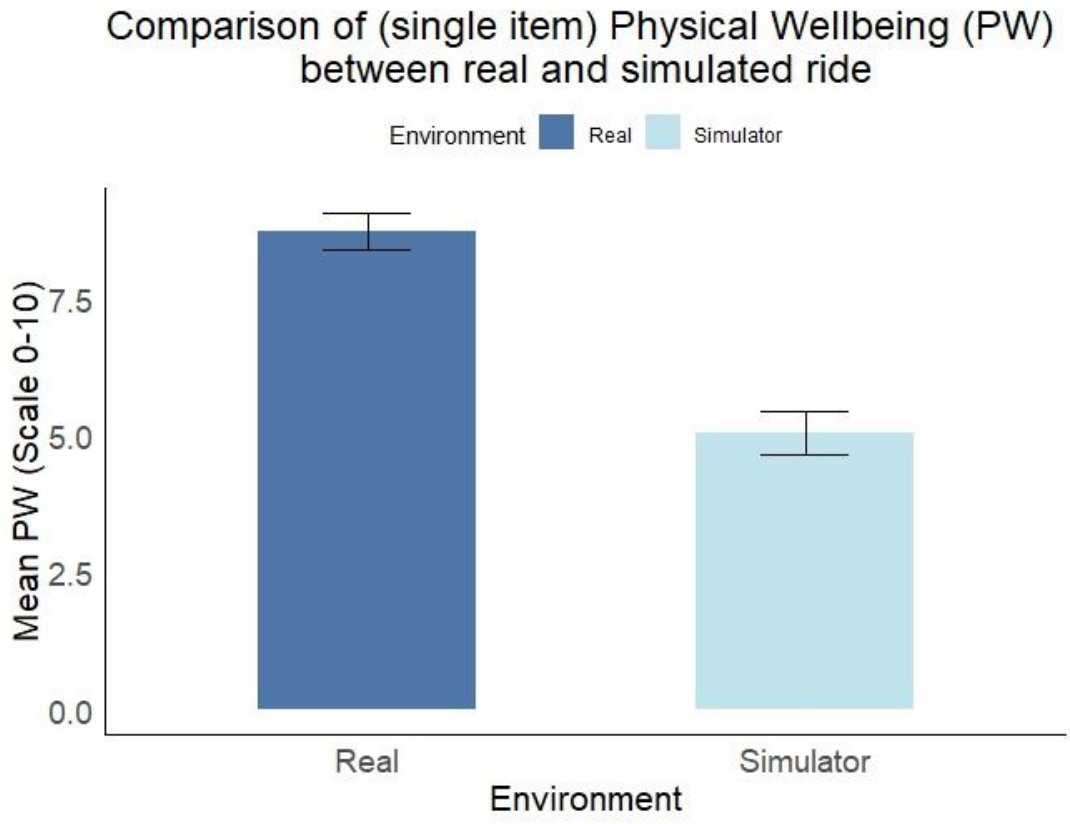


Figure 18. Descriptives for Physical Wellbeing (with 95% credible interval)



5. Discussion

In the present study, we investigated the validity of physiological (H1) and cognitive stress indicators (H2) relying on a within-subject design using a multi-method approach on a 23 km driving route (comprising urban, rural, and highway segments) between a real vehicle and a medium-fidelity simulator. In contrast to the majority of previous studies, we employed a Bayesian analysis approach and considered various driving scenarios.

For physiological data, which was recorded continuously throughout the drive, both absolute (H1a) and relative validity (H1b) were considered. Absolute validity was concluded from the absence of an effect of the test environment (real world vs. simulator), while relative validity will be concluded from the absence of an interaction effect between the test environment and the driving scenario.

Our results regarding GSR-related variables indicate both absolute and relative validity for SCR, meaning that physiological responses do not differ between simulated and real-world driving for this parameter. No absolute validity was found for SCL. However, we found evidence for relative validity for SCL, similar to the findings reported by Reimer and Mehler (2011), whereas Mueller (2015) did not report validity. For PA there was no absolute nor relative validity. To our knowledge, no prior validation studies considered SCR and PA, making it difficult to compare our results to existing literature.

Regarding ECG parameters, our results suggest anecdotal evidence for absolute validity for RMSSD, as well as absolute validity for SDNN. Furthermore, both HRV parameters (RMSSD and SDNN) demonstrated relative validity. Consequently, for these parameters, a drive in the simulator elicits comparable physiological responses to a drive in a real vehicle. However, no such findings were observed for RR-interval, as neither absolute nor relative validity was confirmed. Similarly, we could not establish absolute or relative validity for HR.

For HRV parameters RMSSD and SDNN, our findings align with Mueller (2015), who also reported relative validity for HRV. However, it should be noted that this study did not specify which HRV parameter was investigated. For HR, Reimer and Mehler (2011) and Li et al. (2013) reported absolute validity, while Reimer and Mehler (2011) and Mueller (2015) also found relative validity. In contrast, Johnson et al. (2011) and Milleville-Pennel and Charron (2015) found no validity for HR. Fors et al. (2013) reported collecting ECG data but did not analyze ECG parameters in their study.

Furthermore, there was absolute validity regarding SCT, indicating a similar biological stress level after driving in the two environments. To the best of our knowledge, SCT has not been empirically compared across simulator and real world driving in the past.

One potential reason why PA did not show valid results could be that it is an intensity-based measure, especially sensitive to sudden and unexpected stimuli. For most participants, this study was their first experience with a driving simulator, which could have been perceived as a novel stimulus and hence may have increased PA. On the other hand, this result also corresponds to the fact that cognitive stress indicators also indicated a higher stress level in the simulator.

A possible explanation for the lack of validity concerning RR-Int could be that RR-Int is more susceptible to artifacts, such as micro-movements within the vehicle (e.g., braking, turning), whereas the other two parameters (RMSSD and SDNN) are more stable as they are averaged over longer time periods.

The reason for higher HR values in the real vehicle could result from perceived risk. While participants are theoretically exposed to real dangers during a real drive, they are likely aware that there are no physical consequences to accidents in the simulator (Caird & Horrey, 2016; Vlakveld, 2005).

Concluding our findings regarding the stated hypotheses, we cannot generally accept H1a or H1b in light of inconsistent findings. Absolute validity (H1a) was given for SCR, RMSSD, SDNN, and SCT, while there was no absolute validity regarding SCL, PA, HR, and RR-Int. Relative validity (H1b) was given for SCR, SCL, RMSSD, and SDNN but must be rejected for PA, HR, RR-Int, and SCT. Caird and Horrey (2016) already emphasized that validity at least in parts depends on the dependent variables considered, which we can hence confirm regarding physiological stress.

Regarding the cognitive stress indicators, we found absolute validity only for the Worry dimension of the SSSQ. A comparable study by Galante et al. (2018), however, found absolute validity for the Distress dimension instead. Meanwhile, Galante et al. (2018), Diels et al. (2011), Milleville-Pennel and Charron (2015), Mueller (2015), and Lobjois et al. (2021) found relative validity for the overall NASA-TLX score. While we did not directly test for relative validity, no absolute validity was found regarding the NASA-TLX in our study.

Absolute validity was neither confirmed regarding VO and PW. Since these items were developed by us, they cannot be directly compared with previous research.

Participants reported significantly higher cognitive stress in the simulator, which could be due to various factors. For most of the sample, driving a simulator was a completely new situation, where steering, braking, and speed perception differ from real vehicles, which is also reflected in the VO variable. Although we did not assessed simulator sickness using standardized tools, a certain degree could have influenced stress perception.

Summarizing, regarding H2, we can neither make a final conclusion. While absolute validity was given for Worry, the other cognitive stress indicators point at higher stress in the simulator compared to the real world.

Table 5 summarizes the results regarding the validity of the dependent variables and allows for an integrative interpretation. A pattern emerges: physiological indicators such as SCR, RMSSD and SDDN show consistent validity on both absolute and relative levels. This supports their applicability in driving simulators, particularly for analyzing intra-individual responses. Other measures, such as HR or PA, demonstrate less consistency, which may indicate greater context sensitivity or interindividual variability.

In contrast, the cognitive measures show a more inconsistent pattern. This may be due to the fact that the data were not collected after each section, but retrospectively for the entire drive. Section-specific results might have turned out differently. The overview

of table 5 thus highlights that continuously recorded physiological data are more suitable for validating driving simulators, whereas cognitive measure may require methodological adoption to yield comparably robust results.

Table 5. Overview of the validation results found

"Level"	Parameters	Absolute	Relative
Physiological	SCR	✓	✓
	SCL	X	✓
	PA	X	X
	HR	X	X
	RR-Int	X	X
	RMSSD	✓	✓
	SDNN	✓	✓
	SCT	✓	/
Cognitive	VO	X	/
	NASA-TLX	X	/
	Worry	✓	/
	Distress	X	/
	Engagement	X	/
	STRESS	X	/
	PW	X	/

6. Limitations

Objective performance data (driving data) were not considered in our analysis as intra-individual processes were the focus. Hence, we cannot conclude whether differences in stress actually would have induced differences in driving behavior. Note, however, that driving parameters have been frequently considered in driving simulator validation studies in the past (Wynne et al., 2019).

The decision not to include behavioral driving data was primarily due to technical constraints. The laptop system used for data acquisition was battery-powered and already processing multiple physiological signals in real-time. Adding additional data streams, such as driving behavior, would have risked overloading the system, particularly in the real-world driving context, where no fixed power supply was available.

Nevertheless, we acknowledge the value of integrating both physiological and behavioral data to better understand the relationship between stress and driving performance. Future studies should aim to include driving performance measures in combination with physiological and cognitive stress data for a more comprehensive analysis.

There are mixed results for the physiological and cognitive stress indicators. While SCR shows both absolute and relative validity, this is not the case for PA although these indicators are linked to each other. This may suggest that some indicators are more strongly influenced by factors such as unfamiliarity with the simulator, leading to stronger reactions in the participants. However, this variability is more likely due to situational sensitivity and individual differences rather than a systematic error of the

simulator. Nevertheless, future studies should consider exploring the potential effects of simulator novelty and the influence of individual differences more closely to better understand their impact on stress indicators.

For cortisol measurements, a setting is recommended in which participants complete the two tests on different days, ideally at the same time of day, as the degradation of stress hormones that occurs can influence the results of multiple tests within a day. Since cortisol secretion follows a circadian rhythm, with the highest levels in the morning and a gradual decline throughout the day, variations in measurement timing could significantly affect the outcomes. While our sample was larger than those in other studies, the stability of the indicators we found needs to be empirically verified in further studies.

Furthermore, simulator sickness was not systematically assessed in the present study. Although none of the participants showed overt symptoms or had to discontinue the experiment due to simulator sickness, we cannot rule out the possibility that some individuals experienced mild discomfort, which may have influenced their stress responses. Future studies should systematically measure simulator sickness to better understand its potential impact on physiological and subjective stress indicators.

Additionally, there is variance in the real-world driving environment induced by variations in the weather, traffic, or similar, that is systematically absent in the simulator, which limits the comparability of the two test environments.

Moreover, the order of the simulator and real vehicle conditions was not counterbalanced for practical reasons. We acknowledge that this may introduce a potential order effect. This should be addressed in future studies.

The study was conducted on public roads with traffic conditions typical for this semi-rural area. While the traffic volume remained consistently low during the test, this reflects the usual traffic patterns for the area, where high congestion or heavy traffic is not commonly encountered. This controlled traffic environment ensured safety and comparability between real-world and simulated driving scenarios. However, it may limit the ecological validity when extrapolating the findings to more stressful traffic conditions, such as those encountered in urban areas or during rush hours. Stressors such as heavy congestion, unpredictable driver behavior or adverse weather were not present and were not simulated in the study. Therefore, repeating the test under varying environmental conditions, including those that induce greater stress, such as heavy traffic or poor weather conditions, would be advisable.

In addition to this, while the current study did not focus on age or gender effects, future research should examine whether and how these factors moderate physiological stress responses in different driving environments.

7. Conclusion

In our study, we examined the validity of physiological and cognitive stress indicators in a medium-fidelity driving simulator compared to real-world driving, with the simulated drive being an exact replica of the real route. Using a multi-method approach and a Bayesian analysis, we assessed both absolute and relative validity for various parameters.

Our results indicate absolute validity for the physiological stress indicators SCR, RMSSD, SDNN, and SCT, suggesting that these measures reflect similar physiological responses in the simulator as in real-world driving. Additionally, we found relative validity for SCL, RMSSD, and SDNN, supporting the intra-individual comparability between simulated and real driving. However, no validity was found for PA, HR, and RR-Int.

Regarding cognitive stress indicators, only the Worry dimension of the SSSQ demonstrated absolute validity, whereas all other cognitive parameters, except for our single-item measure of stress, exhibited relative validity. Overall, our findings suggest that stress experiences in the simulator are comparable to those in real-world driving, although the simulator is subjectively perceived as more stressful.

Despite certain limitations, such as the lack of environmental variability in the simulator and an unbalanced sequence of real and simulated drives, our study provides valuable insights into the validation of physiological and cognitive stress indicators. Future research should further investigate these findings under varied driving conditions and consider individual differences more explicitly.

References

Ahlström, C., Bolling, A., Sörensen, G., Eriksson, O., & Andersson, A. (2012). *Validating speed and road surface realism in VTI driving simulator III*. Statens väg-och transportforskningsinstitut.

Andreassi, J. L. (2010). *Psychophysiology*. Psychology Press. https://doi.org/10.4324/9780203880340

Bella, F. (2008). Driving simulator for speed research on two-lane rural roads. *Accident; Analysis and Prevention*, *40*(3), 1078–1087. https://doi.org/10.1016/j.aap.2007.10.015

Bella, F., Calvi, A., & D'Amico, F. (2014). Analysis of driver speeds under night driving conditions using a driving simulator. *Journal of Safety Research*, *49*, 45–52. https://doi.org/10.1016/j.jsr.2014.02.007

Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., & Sleight, P. (2000). Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology*, *35*(6), 1462–1469. https://doi.org/10.1016/S0735-1097(00)00595-7

Berntson, G. G., Norman, G. J., Hawkley, L. C., & Cacioppo, J. T. (2008). Cardiac autonomic balance versus cardiac regulatory capacity. *Psychophysiology*, *45*(4), 643–652. https://doi.org/10.1111/j.1469-8986.2008.00652.x

Bitterman, M. E., & Holtzman, W. H. (1952). Conditioning and extinction of the galvanic skin response as a function of anxiety. *Journal of Abnormal Psychology*, *47*(3), 615–623. https://doi.org/10.1037/h0057190

Blaauw, G. J. (1982). Driving Experience and Task Demands in Simulator and Instrumented Car: A Validation Study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(4), 473–486. https://doi.org/10.1177/001872088202400408

Blana, E. (2001). *The behavioural validation of driving simulators as research tools: a case study based on the Leeds Driving Simulator.* University of Leeds. https://etheses.whiterose.ac.uk/id/eprint/11329/

Boer, E. R. (2000). Experiencing the same road twice: A driver centered comparison between simulation and reality. In *Proceeding of Driving Simulation Conference DSC 2000*.

Boucsein, W. (2012). *Electrodermal Activity*. Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2017). *Handbook of Psychophysiology*. Cambridge University Press. https://doi.org/10.1017/9781107415782

Caird, J. K., & Horrey, W. J. (2016). A review of novice and teen driver distraction. *Handbook of Teen and Novice Drivers*, 189–210.

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, *9*(3), 399–431. https://doi.org/10.1152/physrev.1929.9.3.399

Carter, C. J., & Laya, O. (1998). Drivers visual search in a field situation and in a driving simulator. In A. G. Gale, I. D. Brown, C. M. Haselgrave, & S. P. Taylor (Eds.), *Vision in vehicles VI.* Elsevier Science Ltd.

Caruelle, D., Gustafsson, A., Shams, P., & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research*, *104*, 146–160. https://doi.org/10.1016/j.jbusres.2019.06.041

Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature Reviews. Endocrinology*, *5*(7), 374–381. https://doi.org/10.1038/nrendo.2009.106

Chyung, S. Y. Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement*, *57*(5), 38–48. https://doi.org/10.1002/pfi.21763

Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *8*(2), 132–142. https://doi.org/10.1177/107385840200800209

Darrow, C. W. (1933). Considerations for evaluating the galvanic skin reflex. *American Journal of Psychiatry*, *90*(2), 285–298. https://doi.org/10.1176/ajp.90.2.285

Daviaux, Y., Bonhomme, E., Ivers, H., Sevin, É. de, Micoulaud-Franchi, J.-A., Bioulac, S., Morin, C. M., Philip, P., & Altena, E. (2020). Event-Related Electrodermal Response to Stress: Results From a Realistic Driving Simulator Scenario. *Human Factors*, *62*(1), 138–151. https://doi.org/10.1177/0018720819842779

Dawes, J. (2002). Five point vs. eleven point scales: Does it make a difference to data characteristics? *Australasian Journal of Market Reserach*(10(1)), 39–47.

Dawson, M. E., Schell, A. M., & Courtney, C. G. (2011). The skin conductance response, anticipation, and decision-making. *Journal of Neuroscience, Psychology, and Economics*, *4*(2), 111–116. https://doi.org/10.1037/a0022619

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, *130*(3), 355–391. https://doi.org/10.1037/0033-2909.130.3.355

Diels, C., Robbins, R., & Reed, N. (2011). Behavioural validation of the TRL driving simulator DigiCar: phase 1: speed choice. In (Vol. 5, pp. 429–446).

Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. (1996). Heart Rate Variability. *Circulation*, *93*(5), 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043

Engen, T. (2008). Use and validation of driving simulators.

Engert, V., Merla, A., Grant, J. A., Cardone, D., Tusche, A., & Singer, T. (2014). Exploring the use of thermal infrared imaging in human stress research. *PloS One*, *9*(3), e90782. https://doi.org/10.1371/journal.pone.0090782

Espié, S., Gauriat, P., & Duraz, M. (2005). Driving simulators validation: The issue of transferability of results acquired on simulator. In *Driving Simulation Conference North-America (DSC-NA 2005), Orlondo, FL*.

Fischer, M., Labusch, A., Bellmann, T., & Seehof, C. (2015). A task-oriented catalogue of criteria for driving simulator evaluation. In *Proceedings of the Driving Simulation Conference 2015*.

Fors, C., Ahlström, C., & Anund, A. (2013). *Simulator validation with respect to driver sleepiness and subjective experiences : final report of the project SleepEYE II, part 1* (ViP publication : ViP - Virtual Prototyping and Assessment by Simulation 2013-1). Swedish National Road and Transport Research Institute, Human-vehicle-transport system interaction.

Fowles, D. C. (1986). The eccrine system and electrodermal activity. *Psychophysiology: Systems, Processes, and Applications*, *1*, 51–96.

Francis, A. L. (2018). The Embodied Theory of Stress: A Constructionist Perspective on the Experience of Stress. *Review of General Psychology*, *22*(4), 398–405. https://doi.org/10.1037/gpr0000164

Galante, F., Bracco, F., Chiorri, C., Pariota, L., Biggero, L., & Bifulco, G. N. (2018). Validity of Mental Workload Measures in a Driving Simulation Environment. *Journal of Advanced Transportation*, *2018*, 1–11. https://doi.org/10.1155/2018/5679151

Giannakakis, G [G.], Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., Marias, K., & Tsiknakis, M [M.] (2017). Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, *31*, 89–101. https://doi.org/10.1016/j.bspc.2016.06.020

Giannakakis, G [Giorgos], Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M [Manolis] (2022). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, *13*(1), 440–460. https://doi.org/10.1109/TAFFC.2019.2927337

Godley, S. T., Triggs, T. J., & Fildes, B. N. (2002). Driving simulator validation for speed research. *Accident; Analysis and Prevention*, *34*(5), 589–600. https://doi.org/10.1016/S0001-4575(01)00056-2

Gulian, E., Matthews, G., Glendon, A. I., Davies, D. R., & Debney, L. M. (1989). Dimensions of driver stress. *Ergonomics*, *32*(6), 585–602. https://doi.org/10.1080/00140138908966134

Hall, J. E., & Hall, M. E. (2020). *Guyton and Hall Textbook of Medical Physiology* (14th Revised Edition). Elsevier - Health Sciences Division.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology. Human Mental Workload* (Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey, J. A., & Picard, R. W [R. W.] (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. https://doi.org/10.1109/TITS.2005.848368

Hellhammer, D. H [Dirk H.], Wüst, S., & Kudielka, B. M. (2009). Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology*, *34*(2), 163–171. https://doi.org/10.1016/j.psyneuen.2008.10.026

Helton, W. S. (2004). *PsycTESTS Dataset.* https://doi.org/10.1037/t57758-000

Himmels, C., Venrooij, J., Parduzi, A., Peller, M., & Riener, A. (2024). The bigger the better? Investigating the effects of driving simulator fidelity on driving behavior and perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *101*, 250–266. https://doi.org/10.1016/j.trf.2024.01.007

Himmels, C., Weigl, K., Riener, A., & Venrooij, J. (2024). Establishing driving simulator validity: drawbacks of null-hypothesis significance testing when compared to equivalence tests and Bayes factors. *Theoretical Issues in Ergonomics Science*, *25*(5), 546–566. https://doi.org/10.1080/1463922X.2023.2286478

Hoffmann, S., Krüger, H. P., & Buld, S. (2003). Avoidance of simulator sickness by training the adaptation to the driving simulation. *VDI Berichte*, 385–404.

ISO 13407. *Human-centred design processes for interactive systems*.

Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.

Johnson, M. J., Chahal, T., Stinchcombe, A., Mullen, N., Weaver, B., & Bédard, M. (2011). Physiological responses to simulated and on-road driving. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *81*(3), 203–208. https://doi.org/10.1016/j.ijpsycho.2011.06.012

Kaptein, N., Theeuwes, J., & van der Horst, R. (1996). Driving Simulator Validity: Some Considerations. *Transportation Research Record: Journal of the Transportation Research Board*, *1550*, 30–36. https://doi.org/10.3141/1550-05

Kirschbaum, C., & Hellhammer, D. H [D. H.] (1994). Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. *Psychoneuroendocrinology*, *19*(4), 313–333. https://doi.org/10.1016/0306-4530(94)90013-2

Klüver, M. (2016). *Can we trust driving simulator studies: The behavioral validity of the Daimler AG driving simulators* [, Johannes-Gutenberg-Universität Mainz]. BibTeX.

Klüver, M., Herrigel, C., Heinrich, C., Schöner, H.-P., & Hecht, H. (2016). The behavioral validity of dual-task driving performance in fixed and moving base driving simulators. *Transportation Research Part F: Traffic Psychology and Behaviour*, *37*, 78–96. https://doi.org/10.1016/j.trf.2015.12.005

Kontogiannis, T. (2006). Patterns of driver stress and coping strategies in a Greek sample and their relationship to aberrant behaviors and traffic accidents. *Accident; Analysis and Prevention*, *38*(5), 913–924. https://doi.org/10.1016/j.aap.2006.03.002

Koolhaas, J. M., Bartolomucci, A., Buwalda, B., Boer, S. F. de, Flügge, G., Korte, S. M., Meerlo, P., Murison, R., Olivier, B., Palanza, P., Richter-Levin, G., Sgoifo, A., Steimer, T., Stiedl, O., van Dijk, G., Wöhr, M., & Fuchs, E. (2011). Stress revisited: A critical evaluation of the stress concept. *Neuroscience and Biobehavioral Reviews*, *35*(5), 1291–1301. https://doi.org/10.1016/j.neubiorev.2011.02.003

Kyriakou, K., Resch, B., Sagl, G., Petutschnig, A., Werner, C., Niederseer, D., Liedlgruber, M., Wilhelm, F., Osborne, T., & Pykett, J. (2019). Detecting Moments of Stress from Measurements of Wearable Physiological Sensors. *Sensors (Basel, Switzerland)*, *19*(17). https://doi.org/10.3390/s19173805

Labbé, E., Schmidt, N., Babin, J., & Pharr, M. (2007). Coping with stress: The effectiveness of different types of music. *Applied Psychophysiology and Biofeedback*, *32*(3-4), 163–168. https://doi.org/10.1007/s10484-007-9043-9

Lazarus, R. S. (1966). *Psychological stress and the coping process*. *Psychological stress and the coping process*. McGraw-Hill.

Leung, S.-O. (2011). A Comparison of Psychometric Properties and Normality in 4-, 5-, 6-, and 11-Point Likert Scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Lewis, J. R. (2021). Measuring User Experience With 3, 5, 7, or 11 Points : Does It Matter? *Human Factors*, *63*(6), 999–1011. https://doi.org/10.1177/0018720819881312

Li, J., Zhao, X., Xu, S., Ma, J., & Rong, J. (2013). The Study of Driving Simulator Validation for Physiological Signal Measures. *Procedia - Social and Behavioral Sciences*, *96*, 2572–2583. https://doi.org/10.1016/j.sbspro.2013.08.288

Lobjois, R., Faure, V., Désiré, L., & Benguigui, N. (2021). Behavioral and workload measures in real and simulated driving: Do they tell us the same thing about the validity of driving simulation? *Safety Science*, *134*, 105046. https://doi.org/10.1016/j.ssci.2020.105046

Lohani, M., Payne, B. R., & Strayer, D. L. (2019). A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. *Frontiers in Human Neuroscience*, *13*, 57. https://doi.org/10.3389/fnhum.2019.00057

Losa, M., Frendo, F., Cofrancesco, A., & Bartolozzi, R. (2013). A procedure for validating fixed-base driving simulators. *Transport*, *28*(4), 420–430. https://doi.org/10.3846/16484142.2013.867281

Manseer, M., & Riener, A. (2014). Evaluation of Driver Stress while Transiting Road Tunnels. In S. Osswald, B. Pearce, D. Szostak, A. L. Kun, L. N. Boyle, E. Miller, & Y. Wu (Eds.), *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1–6). ACM. https://doi.org/10.1145/2667239.2667269

Matthews, G., Dorn, L., Hoyes, T. W., Davies, D. R., Glendon, A. I., & Taylor, R. G. (1998). Driver stress and performance on a driving simulator. *Human Factors*, *40*(1), 136–149. https://doi.org/10.1518/001872098779480569

McCorry, L. K. (2007). Physiology of the autonomic nervous system. *American Journal of Pharmaceutical Education*, *71*(4), 78. https://doi.org/10.5688/aj710478

Milardo, S., Rathore, P., Amorim, M., Fugiglando, U., Santi, P., & Ratti, C. (2022). Understanding Drivers' Stress and Interactions With Vehicle Systems Through Naturalistic Data Analysis. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14570–14581. https://doi.org/10.1109/TITS.2021.3130438

Milleville-Pennel, I., & Charron, C. (2015). Driving for Real or on a Fixed-Base Simulator: Is It so Different? An Explorative Study. *Presence: Teleoperators and Virtual Environments*, *24*(1), 74–91. https://doi.org/10.1162/PRES_a_00216

Mueller, J. A. (2015). *Driving in a simulator versus on-road: The effect of increased mental effort while driving on real roads and a driving simulator(PhD dissertation)*. Montana State University.

Navea, R. F., Buenvenida, P. J., & Cruz, C. D. (2019). Stress Detection using Galvanic Skin Response: An Android Application. *Journal of Physics: Conference Series*, *1372*(1), 12001. https://doi.org/10.1088/1742-6596/1372/1/012001

Norman, G. J., Necka, E., & Berntson, G. G. (2016). The Psychophysiology of Emotions. In *Emotion Measurement* (pp. 83–98). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00004-7

Parduzi, A. (2021). *Bewertung der Validität von Fahrsimulatoren anhand vibro-akustischer Fahrzeugschwingungen* [, Dissertation, Berlin, Technische Universität Berlin, 2021]. BibTeX.

Perello-March, J. R., Burns, C. G., Birrell, S. A., Woodman, R., & Elliott, M. T. (2022). Physiological measures of risk perception in Highly Automated Driving. IEEE Transactions on Intelligent Transportation Systems, 23(5), 4811–4822. https://doi.org/10.1109/TITS.2022.3146793

Poh, M.-Z., Swenson, N. C., & Picard, R. W [Rosalind W.] (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Bio-Medical Engineering*, *57*(5), 1243–1252. https://doi.org/10.1109/TBME.2009.2038487

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: A field study and simulation validation. *Ergonomics*, *54*(10), 932–942. https://doi.org/10.1080/00140139.2011.604431

Reinhardt, T., Schmahl, C., Wüst, S., & Bohus, M. (2012). Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). *Psychiatry Research*, *198*(1), 106–111. https://doi.org/10.1016/j.psychres.2011.12.009

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304.

Scherz, W. D., Seepold, R., & Ortega, J. A. (2023). Experiment design for Stress data collection while driving in a simulator. *Procedia Computer Science*, *225*, 4381–4388. https://doi.org/10.1016/j.procs.2023.10.435

Selye, H. (1950). Stress and the general adaptation syndrome. *British Medical Journal*, *1*(4667), 1383–1392. https://doi.org/10.1136/bmj.1.4667.1383

Selye, H. (1976). Forty years of stress research: Principal remaining problems and misconceptions. *Canadian Medical Association Journal*, *115*(1), 53–56.

Selye, H. (1978). *The stress of life, Rev. ed*. *The stress of life, Rev. ed*. Mcgraw Hill.

Selye, H. (1983). *Selye's Guide to Stress Research*. *Selye's Guide to Stress Research*. Van Nostrand Reinhold. https://books.google.de/books?id=lOIgAQAAIAAJ

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *14*(2), 410–417. https://doi.org/10.1109/TITB.2009.2036164

Shaffer, F., McCraty, R., & Zerr, C. L. (2014). A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability. *Frontiers in Psychology*, *5*, 1040. https://doi.org/10.3389/fpsyg.2014.01040

Sharma, N., & Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, *108*(3), 1287–1301. https://doi.org/10.1016/j.cmpb.2012.07.003

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In M. B. Rosson & D. Gilmore (Eds.), *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2651–2656). ACM. https://doi.org/10.1145/1240866.1241057

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording*. Oxford University Press, USA.

Taelman, J., Vandeput, S., Vlemincx, E., Spaepen, A., & van Huffel, S. (2011). Instantaneous changes in heart rate regulation due to mental load in simulated office work. *European Journal of Applied Physiology*, *111*(7), 1497–1505. https://doi.org/10.1007/s00421-010-1776-0

Terumitsu, H., Tetsuo, Y., & Tsuyoshi, T. (2007). Development of the driving simulation system MOVIC-T4 and its validation using field driving data. *Tsinghua Science and Technology*, *12*(2), 141–150. https://doi.org/10.1016/S1007-0214(07)70021-4

Törnros, J. (1998). Driving behavior in a real and a simulated road tunnel--a validation study. *Accident; Analysis and Prevention*, *30*(4), 497–503. https://doi.org/10.1016/S0001-4575(97)00099-7

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Vienne, F., Caro, S., Auberlet, J.-M., Rosey, F., & Dumont, E. (2014). Driving simulator: an innovative tool to test new road infrastructures. In

Vlakveld, W. P. (2005). The use of simulators in basic driver training. In *Humanist TFG Workshop on the Application of New Technologies to Driver Training, Brno, Czech Republic. Available at: www. escope. info/download/research_and_development/HUMANISTA_13Use. pdf*.

Westerman, S., & Haigney, D. (2000). Individual differences in driver stress, error and violation. *Personality and Individual Differences*, *29*(5), 981–998. https://doi.org/10.1016/S0191-8869(99)00249-4

Wilcott, R. C. (1967). Arousal sweating and electrodermal phenomena. *Psychological Bulletin*, *67*(1), 58–72. https://doi.org/10.1037/h0024140

Winter, D. J. de, van Leeuwen, P., & Happee, R. (2012). Advantages and Disadvantages of Driving Simulators: A Discussion. *Proceedings of Measuring Behavior*, 47–50.

Wynne, R. A., Beanland, V., & Salmon, P. M. (2019). Systematic review of driving simulator validation studies. *Safety Science*, *117*, 138–151. https://doi.org/10.1016/j.ssci.2019.04.004

Xue, H., Previati, G., Gobbi, M., Mastinu, G., & others (2023). Research and Development on Noise, Vibration, and Harshness of Road Vehicles Using Driving Simulators-A Review. *SAE INTERNATIONAL JOURNAL of VEHICLE NOISE and VIBRATION*, *7*(4), 555–577.

Zhong, S., Fu, X., Lu, W., Tang, F., & Lu, Y. (2022). An Expressway Driving Stress Prediction Model Based on Vehicle, Road and Environment Features. *IEEE Access*, *10*, 57212–57226. https://doi.org/10.1109/ACCESS.2022.3165570

Zhou, X., Ma, L., & Zhang, W. (2022). Event-related driver stress detection with smartphones among young novice drivers. *Ergonomics*, *65*(8), 1154–1172. https://doi.org/10.1080/00140139.2021.2020342

Zöller, I., Abendroth, B., & Bruder, R. (2019). Driver behaviour validity in driving simulators – Analysis of the moment of initiation of braking at urban intersections. *Transportation Research Part F: Traffic Psychology and Behaviour*, *61*, 120–130. https://doi.org/10.1016/j.trf.2017.09.008

Zöller, I. M. (2015). Analyse des Einflusses ausgewählter Gestaltungsparameter einer Fahrsimulation auf die Fahrerverhaltensvalidität.

A.2 Research Paper No. 2: Comparison of gaze behavior in real and simulated driving

Authors: Czaban, M. & Purucker, C. (2025)

Citation: Czaban, M., & Purucker, C. (2025). Comparison of Gaze Behavior in Real- and Simulated Driving. In: Davis, F.D., Riedl, R., vom Brocke, J., Léger, PM., Randolph, A.B., Müller-Putz, G.R. (eds) Information Systems and Neuroscience. NeuroIS 2025. Lecture Notes in Information Systems and Organisation, Cham

URL-Preprint: https://www.neurois.org/wp-content/uploads/2025/05/NeuroIS-Retreat-2025-Preprint-Proceedings.pdf

Abstract: Driving simulators are essential for the development of vehicle systems, as they enable safe and efficient user engagement. Their validity determines the extent to which the results from empirical user studies obtained in driving simulators can be transferred to real-world driving situations.

This study examines the gaze behavior of participants in a within-subject design, both in a real vehicle and in a driving simulator with a digital road replica. Using gaze-point plots and expert ratings, we compare fixation patterns across three road sections (City Drive, Rural Drive, Highway). The results visually indicate a moderate to high similarity in gaze distributions, suggesting consistent fixation patterns in both environments, with some notable exceptions on an individual level and generally highest matches in the City Drive.

However, further statistical analyses are necessary to quantitatively con-firm similarities and assess systematic differences.

Keywords: Driving Simulator Validity, Eye-Tracking, Gaze Behavior, Fixation Patterns

1 Introduction

The automotive industry is undergoing a major transition toward automation, accompanied by rapid advancements in driver assistance systems (Stoma et al., 2021). For these innovations to gain public acceptance, they must align with user needs in accordance with the principle of customer centricity (Kleinaltenkamp et al., 2022; Riedl et al., 2024).

Driving simulators are widely used in automotive research, providing cost-efficient, safe, and standardized environments for the development and evaluation of driver assistance systems (Drosdol & Panik, 1985; Pawar et al., 2022). Simulator studies are intended to support the design of safer vehicles and to improve our understanding of driver behavior—ultimately contributing to accident reduction (Carroll et al., 2023).

However, the use of simulators is not without challenges. Despite extensive re-search over the past two decades, the impact of confounding variables, such as the complexity of the road environment, traffic density, or visibility, on driver behavior (e.g., vehicle control, monitoring of the driving scene) remains only partially understood. Moreover, even if these influences were fully known, replicating them accurately in simulation environments is often not feasible due to technical or cost-related constraints. This

raises a central question: which variables can, and should, be realistically reproduced in simulators to ensure meaningful research outcomes (Caird & Horrey, 2011).

It is generally accepted that the validity of a driving simulator depends on whether the behavior observed in the simulation corresponds to that observed in real-world driving (Wang et al., 2010; Wynne et al., 2019). According to transfer-of-training theory, knowledge gained through simulations can only be generalized to real-world contexts if behavioral patterns, such as gaze behavior, are comparable across both settings (Blume et al., 2010; Liu et al., 2023).

Gaze behavior plays a central role in driving, as visual attention is essential for maintaining situational awareness and performing driving tasks (Martin et al., 2018). While simulator validity research has traditionally focused on performance metrics such as speed or lane keeping (Wynne et al., 2019), gaze behavior has received comparatively less attention. Only a few studies (Carter & Laya, 1998; Fors et al., 2013; Mueller, 2015) have attempted to validate eye-tracking data between simulated and real driving conditions, and while overall, they found differences between the real car and the simulator, their findings are inconsistent: some report narrower fixation patterns in simulators, while others observe increased gaze dispersion. These discrepancies point to a gap in understanding the ecological validity of simulator-based eye-tracking data.

To address this gap, the present study investigates whether gaze behavior systematically differs between simulated and real-world driving conditions. Based on the assumption that visual attention is influenced by environmental fidelity and task complexity, we compare gaze behavior using eye-tracking data in a within-subject design. Participants drove the same 1:1 replicated route—including urban (City Drive), rural (Rural Drive), and highway (Highway Drive) segments—in both a real vehicle and a high-fidelity driving simulator. Our research question is:

RQ: Does gaze behavior systematically vary between different road types (urban, rural, highway) and between real and simulated driving conditions?

2 Theoretical background and hypothesis development

2.1 Driving simulator validity

The primary goal of simulator validation studies is to determine whether a simulated driving environment provides a valid representation of reality, allowing reliable insights to be drawn for real vehicles (Himmels et al., 2024; Pawar et al., 2022). To achieve this, relevant outcome variables are compared between real and simulated driving (Klüver, 2016; Zöller, 2015).

The validity of driving simulators is divided into physical validity and behavioral validity (Bella et al., 2014). Physical validity describes the degree of correspondence between the simulator and the real vehicle (Klüver, 2016) though a higher level of similarity does not necessarily lead to valid study results (Himmels et al., 2024).

Behavioral validity refers to participants' driving behavior and response data (Blaauw, 1982).

Naturally for studies on driver behavior, behavioral validity is considered to be more important than physical validity, as driving behavior is crucial for the transferability of results, whereas an exact physical replication is not always necessary (Blana, 2001; Terumitsu et al., 2007).

One behavioral parameter, which is discussed in the context of driving simulator validity is gaze behavior, assessed via eye-tracking. In terms of our study behavior validity refers to whether gaze patterns observed in the simulator resemble those in real-world driving.

2.2 Eye-Tracking in driving simulations

Eye-tracking enables detailed analysis of drivers' visual attention and cognitive processing by capturing fixations and saccades (Calvi et al., 2023; Duchowski, 2002). Especially in complex, fast-changing traffic scenarios, drivers rely primarily on foveal vision to identify relevant objects and events, while peripheral vision supports spatial orientation and scene organization (Fisher et al., 2011). The rule of thumb that only directly fixated elements are typically recognized and cognitively processed is therefore widely accepted for eye-tracking measures employed in traffic and driver research.

In the context of simulator validation, eye-tracking metrics offer valuable insights into behavioral realism. Commonly used indicators include fixation duration, gaze dispersion, and the spatial distribution of fixations, particularly in high-relevance areas such as the central roadway (Johansson et al., 2001). These parameters enable direct comparisons of attention allocation between real and simulated environments and are critical for assessing ecological validity (Calvi et al., 2023).

While prior studies consistently show that gaze behavior differs between real-world and simulated driving, the form of these deviations remains inconsistent. Some studies report more concentrated fixations and reduced dispersion in simulators (Carter & Laya, 1998), while others observe broader scan paths or increased fixation frequency (Mueller, 2015). In this sense, the existence of a discrepancy appears robust, but the direction and nature of these differences vary across findings. This heterogeneity may stem from variations in simulator fidelity, interface design, and the perceived cognitive demands of the simulated task (Angell, 2011).

Additionally, research suggests that these deviations tend to diminish as simulator fidelity increases, particularly for metrics like glance durations and visual scanning behavior (Angell, 2011). Accordingly, high-fidelity simulations with realistic environmental modeling are more likely to evoke gaze behavior that mirrors real-world driving.

In the present study, we use gaze-point plots to compare visual attention patterns between real and simulated driving across three distinct road types (urban, rural, highway). By analyzing similarities and differences in fixation distributions, we aim to assess whether gaze behavior in the simulator reflects real-world patterns in a context-sensitive and differentiated manner.

2.3 Research gap and research question

Most validation studies on driving simulators focus on performance data like speed or lane position (Wynne et al., 2019), with fewer examining physiological parameters such as heart rate (e.g., Johnson et al., 2011) or skin conductance (e.g., Reimer & Mehler, 2011). However, eye-tracking data are rarely validated (Calvi et al., 2023). Wynne et al. (2019) call for greater emphasis on these measures, as some researchers believe cognitive demands in simulation mirror those in real driving.

Studies comparing gaze behavior in simulation and reality consistently report differences, yet these vary in direction and magnitude, depending on factors such as simulator fidelity and task context. According to Angell (2011), lower-fidelity simulators may lead to altered visual behavior due to reduced realism and cognitive engagement. For example, Carter & Laya (1998) observed more concentrated fixation areas in the simulator than in real traffic. Similarly, Fors et al. (2013) reported more frequent fixations, albeit within a narrower visual radius. In contrast, Mueller (2015) found greater gaze dispersion in the simulator, both horizontally and vertically. These divergent outcomes suggest that gaze patterns between real and simulated driving are not directly interchangeable and point to unresolved questions regarding the ecological validity of simulator-based eye-tracking data.

Besides simulation fidelity associated to the graphics and dynamic properties involved, these differences might result from the simulated traffic environment or even varying road sceneries. As our research environment replicates the exact properties of the road scenery around the campus in Hof, Bavaria, we chose to address the following research question: RQ: Does gaze behavior systematically vary between different road types (urban, rural, highway) and between real and simulated driving conditions?

To answer this RQ, we compare participants' gaze-point plots while driving the same route in both a real vehicle and a driving simulator with a digital twin. We distinguish between the road sections City Drive (CD), Rural Drive (RD), and High-way (HD). By incorporating sections with distinct visual and task-related demands, we aim to examine whether simulator validity is consistent across different types of real-world scenarios. This approach enhances the ecological validity of the study and improves the potential generalizability of the findings.

3 Method

3.1 Experimental design

Participants. The study follows a within-subject design with 12 participants (7 women, 58.3%; 5 men, 41.7%), averaging 29.3 years (SD = 12.9, range: 18–59). Regarding residence, 58.3% live in rural areas, 41.7% in small or medium-sized towns.

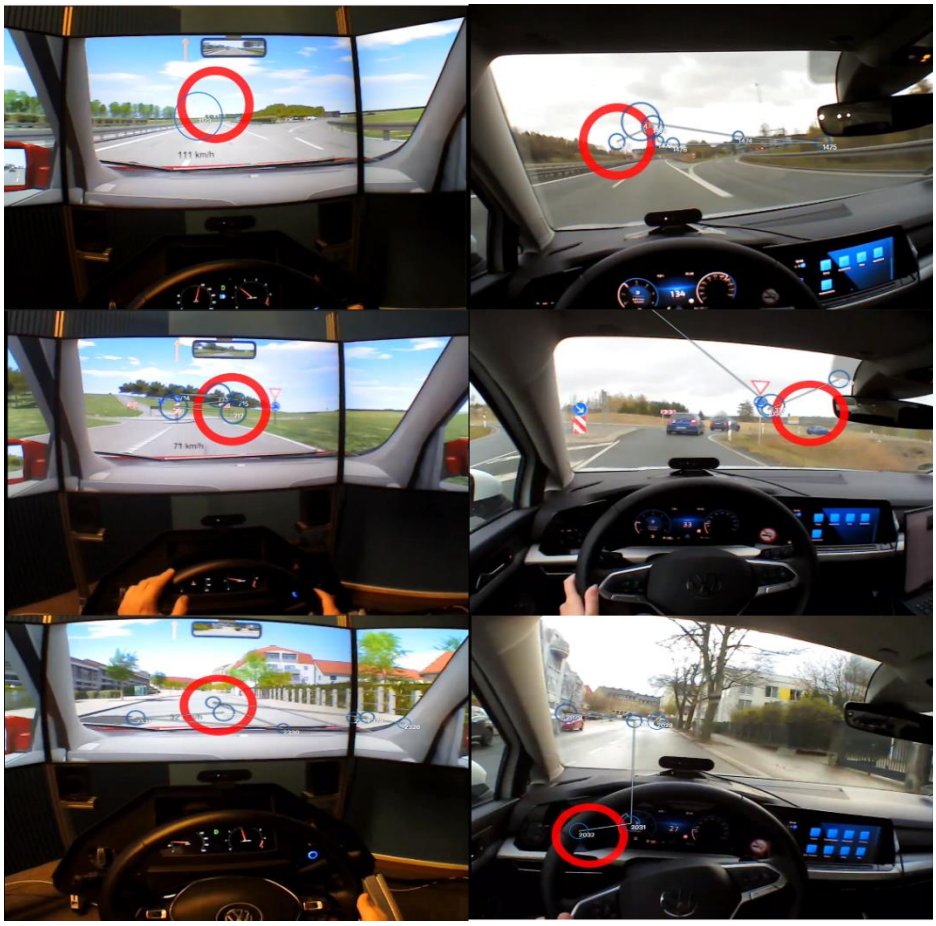
Eye-Tracking Device. To record eye movements, the Pupil Labs Invisible was used, which are mobile eye-tracking glasses with a scene camera resolution of 1088 × 1080 pixels at a frame rate of 30 Hz. The scene camera's field of view is 82° horizontally and 82° vertically. The system's gaze accuracy is 4.6° (uncalibrated).

3.2 Materials, procedure and data processing

Vehicles. The real-world test vehicle used in this study was a VW Golf 8 with 110 kW. The driving simulator, classified as a medium-fidelity simulator in accordance with Wynne et al. (2019) (Wynne et al., 2019), was equipped with a VW Golf 7 steering wheel and pedals, as well as a three-degree-of-freedom motion platform to enhance realism during simulated driving.

Track. The driving route covered a 23 km loop and included three different road types: urban roads (5.3 km), rural roads (9.7 km), and highways (8.0 km). For maximum comparability, the route was replicated in the simulator as a digital twin of the real-world track. Figure 1 shows representative images of urban, rural, and highway sections, each displayed for both the driving simulator (left) and the real vehicle (right row). All participants started the drive at the same location and followed the same sequence of road sections. As the track was designed as a closed loop, it was not feasible, without disproportionate effort and logistical complexity, to randomize or counterbalance the order of the road segments. Additionally, all participants first completed the real drive before performing the same route in the simulator. This fixed order was chosen to ensure that participants had a real-world reference, minimizing disorientation in the simulator. Although this introduces a potential learning effect, the analysis focused on spatial gaze patterns rather than performance metrics, which are more susceptible to such effects.

Figure 1. Comparison of visual scenes from the simulator (left row) and the real vehicle (right row) across highway (top), rural (middle) and urban (down) segments.



Procedure. Upon arrival at the lab, participants were first introduced to the real vehicle and fitted with mobile eye-tracking glasses. They then completed the real-world drive while refraining from speaking, except when receiving instructions. After returning to the lab, participants completed a short familiarization drive in the simulator to minimize the risk of simulator sickness. Following this, they completed the full simulated drive, again wearing the eye-tracking glasses and driving the same route as before. On average, the real-world drive lasted 25 min, the simulator drive 23 min, and the entire experimental session took approximately 90 minutes per participant.

Data Processing. Gaze-point plots were generated for each participant to visualize fixation density across the three road sections (City Drive, Rural Drive, Highway Drive). These visualizations were based on horizontal and vertical gaze coordinates and illustrated the distribution of visual attention in each condition. The data were analyzed both descriptively and exploratorily.

In addition, an expert rating procedure was conducted. Three independent experts reviewed the gaze-point plots for each participant, comparing the real-world and simulated conditions for the entire route as well as for each road section individually. Each expert assigned a similarity score from 0 (no similarity in gaze distribution) to 10 (very high similarity). The average of the three expert ratings was calculated for each participant and for each condition (whole route and individual road sections).

4 Results and discussion

Figure 2 exemplary shows the gaze distributions of Participant 4 and Participant 7 in the road sections CD, RD, and HD. The gaze data are represented in pixel coordinates relative to the scene camera image, where the X-axis denotes the horizontal and the Y-axis the vertical position of the gaze within the video frame, with (0,0) located in the top-left corner of the image.

The plots for CD show similar fixation patterns, with high density in the central visual field and wide dispersion in the periphery. The focus was primarily on the roadway and other road users in both real and simulated drives. However, the real drive had wider fixation dispersion, possibly due to more frequent and richer peripheral stimuli, such as pedestrians. In the simulator, the fixation density appeared more compact, indicating reduced environmental stimuli or a stronger focus on the road. Additionally, the values seemed vertically shifted downward, showing a notable downward spread, suggesting differences in the human-machine interface: in the simulator, relevant driving information may be further from the road scene.

Similar fixation patterns were observed in the RD, with the primary focus again on the roadway. Once again, the horizontal dispersion was smaller in the simulator, likely because more environmental stimuli, such as oncoming traffic, were observed in the real environment. Compared to CD, the gaze dispersion seemed somewhat reduced in both the real vehicle drive and the simulator drive.

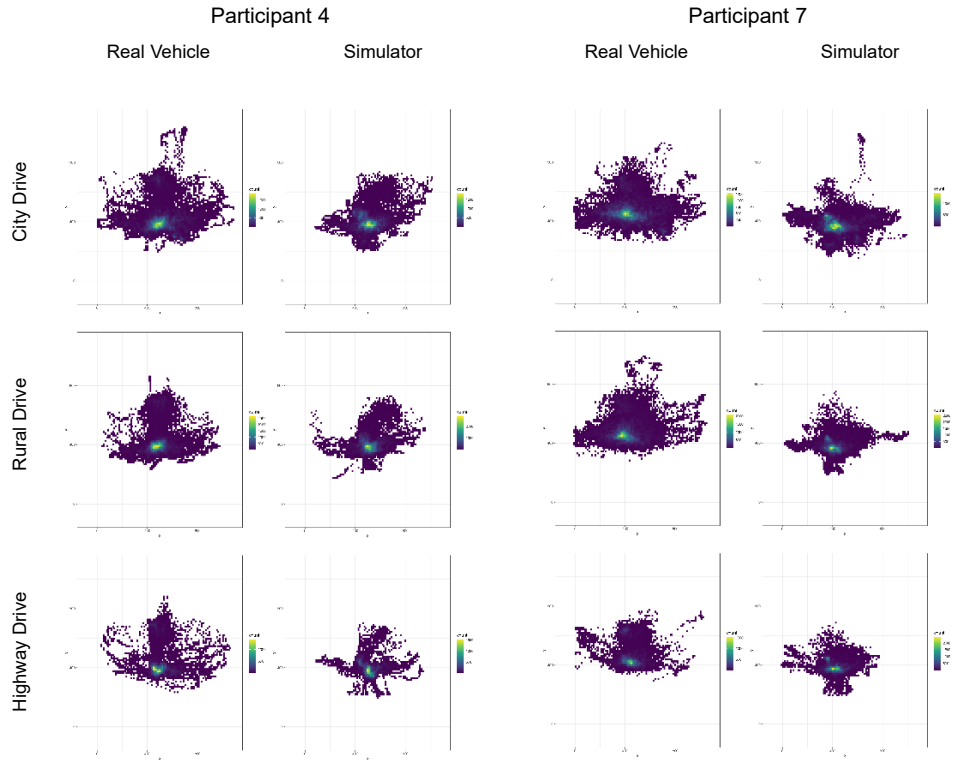
On the HD, the gaze in both driving environments was even more focused on the central area than in the other two conditions. The peripheral dispersion was notably lower, suggesting that the focus was primarily on vehicles ahead or lane markings.

Across all driving conditions, both participants showed changes in gaze patterns, observed in both the real vehicle and the simulator, supporting the assumption of relative validity between the two environments. However, when examining the absolute gaze distribution values, systematic differences between the two environments become evident: gaze dispersion in the simulator is generally more restricted, and the horizontal gaze points are systematically shifted downward.

When comparing both participants, the corresponding experimental environments and driving conditions appear more similar, with participant-specific patterns being less distinguishable. This observation is mostly consistent across the entire sample, though a few participants exhibit distinct gaze patterns, which are more isolated cases.

Overall, the gaze-point plots show a high degree of similarity between real and simulated driving. Slight differences in horizontal and vertical fixation dispersion can be explained by context-dependent environmental factors. A similar pattern was observed with other participants as well.

Figure 2. Comparison of the gaze data from participant 4 and participant 7 in the sections City Drive, Rural Drive, and Highway Drive between real vehicle and simulator (The X- and Y-axes represent pixel coordinates).



To systematically evaluate the visual impression of the gaze-point plots, three independent experts assessed the similarity of gaze patterns between real and simulated driving for each participant. They provided ratings on a scale from 0 (no similarity) to 10 (very high similarity) for the entire route as well as for the individual segments (city, rural, highway). The resulting mean scores and standard deviations are presented in Table 1.

Overall, the average scores, ranging between approximately 5 and 6, suggest a moderate level of similarity between real and simulated gaze behavior. As a tendency, the highest similarity ratings were found for the entire route, followed closely by the highway section. City and rural segments showed slightly lower average similarity scores.

Interestingly, although the highway segment received the highest mean similarity rating, it also exhibited the largest standard deviation. This indicates substantial interindividual variability in gaze similarity for this segment. One possible explanation is that while some participants displayed nearly identical gaze behavior in both conditions, others adapted their gaze patterns more strongly depending on whether they were driving in the simulator or in the real car.

This variability might be partly explained by the nature of the highway section, which is typically more monotonous than urban or rural environments. In the real vehicle, participants may still have experienced a heightened sense of risk due to the presence of other vehicles and real-world consequences of failure, whereas in the simulator, this sense of risk was likely diminished. The relatively neutral visual design of the simulated highway may also have resulted in reduced visual exploration for some participants, thereby increasing perceived similarity. However, the large standard deviation suggests that this effect was not consistent across the sample.

Table 1. Expert ratings for the similarity of the gaze-point plots.

Participant	Whole Track	City	Country	Highway
1	7.67	6.67	6.00	6.67
2	7.33	7.33	6.00	4.67
3	7.00	6.67	4.00	7.00
4	7.33	5.00	6.67	4.67
5	6.33	5.33	6.67	5.00
6	3.67	2.33	1.67	1.00
7	6.00	5.67	5.00	5.33
8	6.67	5.67	5.67	4.33
9	7.67	4.00	6.33	6.33
10	6.33	7.33	7.33	7.33
11	6.67	5.67	3.33	2.67
12	4.67	6.33	5.33	5.00
$Mean_{Experts}$	5.58	4.92	4.92	5.00
$SD_{Experts}$	1.38	1.80	1.85	2.20

Overall, the combination of visual inspection and expert ratings suggests that gaze behavior in real and simulated driving is largely comparable. At the same time, intra- and interindividual differences—particularly pronounced on the highway segment—likely reflect varying perceptions of task relevance and environmental realism. Minor differences in fixation dispersion and vertical gaze orientation are more plausibly attributed to environmental and interface-related factors (e.g., dis-play resolution, realism of scenery) rather than fundamental behavioral divergence.

5 Conclusion and limitations

The gaze-point analysis revealed a moderate degree of visual similarity in fixation patterns between real and simulated driving, yet with a progressive narrowing of gaze distributions from the CD through the RD to the HD. At the same time, two clear absolute differences emerged: simulated gaze data exhibited reduced peripheral dispersion and a consistent downward shift. These observations reconcile apparently contradictory reports in the literature, which have documented both more centralized (Fors et al., 2013) and more widely dispersed fixations (Mueller, 2015) in simulation studies.

Collectively, our findings underscore the context dependence of gaze behavior: both the experimental environment and the driving scenario (CD, RD, HD) exert a influence on fixation distribution. Participant-specific effects, while modest on average, can be substantial in individual cases and therefore warrant consideration in studies of driver state or personalized assistance systems. Notably, the lowest similarity between real and simulated conditions was observed in the highway segment likely a consequence of the simulator's perceived monotony and absence of genuine risk, which diminished visual exploration compared to the real-world drive.

A primary limitation of this study is its relatively small sample size, which constrains the generalizability of our results. Furthermore, potential learning effects may have influenced the results, as all participants completed the road segments in the same order and always began with the real-world drive. This fixed sequence could have introduced systematic biases. Future studies should at least counterbalance the order of driving conditions (real vs. simulated), and ideally also vary the sequence of road segments, although the latter may be difficult to implement in practice.

Moreover, the lack of robust statistical analysis of the gaze data represents a further constraint: although exploratory spatial scan statistics (Benjamin Allévius, 2018; Purucker et al., 2013) were applied, these methods are overly sensitive to central-field differences and ill-suited for peripheral pattern analysis. While expert ratings provided valuable qualitative insights, their inherent subjectivity underscores the need for objective, quantitative similarity metrics in future work.

For future research, we recommend (1) evaluating simulators of varying fidelity to determine whether higher realism promotes closer convergence of simulated gaze patterns with those observed in real driving, or whether extreme fidelity levels produce larger divergences, and (2) adopting advanced statistical approaches (e.g., heatmap-based similarity measures, cluster analyses etc.) to rigorously validate and extend the visually and experientially derived findings reported here.

References

Angell, L. S. (2011). Surrogate Methods and Measures. In Donald L. Fisher, Matthew Rizzo, Jeffrey K. Caird, & John D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology.* CRC Press.

Bella, F., Calvi, A., & D'Amico, F. (2014). Analysis of driver speeds under night driving conditions using a driving simulator. *Journal of Safety Research*, *49*, 45.e1-52. https://doi.org/10.1016/j.jsr.2014.02.007

Benjamin Allévius (2018). scanstatistics: space-time anomaly detection using scan statistics. *Journal of Open Source Software*, *3*(25), 515. https://doi.org/10.21105/joss.00515

Blaauw, G. J. (1982). Driving Experience and Task Demands in Simulator and Instrumented Car: A Validation Study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(4), 473–486. https://doi.org/10.1177/001872088202400408

Blana, E. (2001). *The behavioural validation of driving simulators as research tools: a case study based on the Leeds Driving Simulator.* University of Leeds.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of Training: A Meta-Analytic Review. *Journal of Management*, *36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Caird, J. K., & Horrey, W. J. (2011). Twelve Practical and Useful Questions About Driving Simulation: 12. In Donald L. Fisher, Matthew Rizzo, Jeffrey K. Caird, & John D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology.* CRC Press.

Calvi, A., D'Amico, F., & Vennarucci, A. (2023). Comparing Eye-tracking System Effectiveness in Field and Driving Simulator Studies. *The Open Transportation Journal*, *17*(1), Article e187444782301191. https://doi.org/10.2174/18744478-v17-e230404-2022-49

Carroll, M., Rebensky, S., Chaparro Osman, M., & Deaton, J. (2023). Justification for Use of Simulation. In D. A. Vincenzi, M. Moloua, P. A. Hancock, J. A. Pharmer, & J. C. Ferraro (Eds.), *Human Factors in Simulation and Training* (pp. 65–90). CRC Press. https://doi.org/10.1201/9781003401360-2

Carter, C. J., & Laya, O. (1998). Driver's visual search in a field situation and in a driving simulator. *Vision in Vehicles*, *6*, 21–31.

Drosdol, J., & Panik, F. (1985). The Daimler-Benz Driving Simulator A Tool for Vehicle Development. In *SAE Technical Paper Series, SAE Technical Paper Series.* SAE International400 Commonwealth Drive, Warrendale, PA, United States. https://doi.org/10.4271/850334

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 455–470. https://doi.org/10.3758/BF03195475

Fisher, D. L., Pollatsek, A., & Horrey, W. J. (2011). Eye behaviors: how driving simulators can expand their role in science and engineering. In Donald L. Fisher, Matthew Rizzo, Jeffrey K. Caird, & John D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology.* CRC Press.

Fors, C., Ahlström, C., & Anund, A. (2013). *Simulator validation with respect to driver sleepiness and subjective experiences: final report of the project SleepEYE II, part 1*. Statens väg-och transportforskningsinstitut.

Himmels, C., Venrooij, J., Parduzi, A., Peller, M., & Riener, A. (2024). The bigger the better? Investigating the effects of driving simulator fidelity on driving behavior and perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *101*, 250–266. https://doi.org/10.1016/j.trf.2024.01.007

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye–Hand Coordination in Object Manipulation. *The Journal of Neuroscience*, *21*(17), 6917–6932. https://doi.org/10.1523/JNEUROSCI.21-17-06917.2001

Johnson, M. J., Chahal, T., Stinchcombe, A., Mullen, N., Weaver, B., & Bédard, M. (2011). Physiological responses to simulated and on-road driving. *International Journal of Psychophysiology*, *81*(3), 203–208. https://doi.org/10.1016/j.ijpsycho.2011.06.012

Kleinaltenkamp, M., Eggert, A., Kashyap, V., & Ulaga, W. (2022). Rethinking customer-perceived value in business markets from an organizational perspective. *Journal of Inter-Organizational Relationships*, *28*(1-2), 1–18. https://doi.org/10.1080/26943980.2022.2129545

Klüver, M. (2016). *Can we trust driving simulator studies: The behavioral validity of the Daimler AG driving simulators.* Johannes-Gutenberg-Universität Mainz.

Liu, D., Yu, J., Macchiarella, N. D., & Vincenzi, D. A. (2023). Simulation Fidelity. In D. A. Vincenzi, M. Moloua, P. A. Hancock, J. A. Pharmer, & J. C. Ferraro (Eds.), *Human Factors in Simulation and Training* (pp. 91–108). CRC Press. https://doi.org/10.1201/9781003401360-3

Martin, S., Vora, S., Yuen, K., & Trivedi, M. M. (2018). Dynamics of Driver's Gaze: Explorations in Behavior Modeling and Maneuver Prediction. *IEEE Transactions on Intelligent Vehicles*, *3*(2), 141–150. https://doi.org/10.1109/TIV.2018.2804160

Mueller, J. A. (2015). *Driving in a simulator versus on-road: The effect of increased mental effort while driving on real roads and a driving simulator*. Montana State University.

Pawar, N. M., Velaga, N. R., & Sharmila, R. B. (2022). Exploring behavioral validity of driving simulator under time pressure driving conditions of professional drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *89*, 29–52. https://doi.org/10.1016/j.trf.2022.06.004

Purucker, C., Landwehr, J. R., Sprott, D. E., & Herrmann, A. (2013). Clustered insights. *International Journal of Market Research*, *55*(1), 105–130. https://doi.org/10.2501/IJMR-2013-009

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics*, *54*(10), 932–942. https://doi.org/10.1080/00140139.2011.604431

Riedl, J., Wengler, S., Czaban, M., Mohr, S. V., & Steudtel, S. (2024). *Studies on the Human-Machine-Interface in Advanced Driver Assistance Systems towards Autonomous Driving*. Hochschule Hof. https://doi.org/10.57944/1051-147

Stoma, M., Dudziak, A., Caban, J., & Droździel, P. (2021). The Future of Autonomous Vehicles in the Opinion of Automotive Market Users. *Energies*, *14*(16), 4777. https://doi.org/10.3390/en14164777

Terumitsu, H., Tetsuo, Y., & Tsuyoshi, T. (2007). Development of the driving simulation system MOVIC-T4 and its validation using field driving data. *Tsinghua Science and Technology*, *12*(2), 141–150. https://doi.org/10.1016/S1007-0214(07)70021-4

Wang, Y., Mehler, B., Reimer, B., Lammers, V., D'Ambrosio, L. A., & Coughlin, J. F. (2010). The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing. *Ergonomics*, *53*(3), 404–420. https://doi.org/10.1080/00140130903464358

Wynne, R. A., Beanland, V., & Salmon, P. M. (2019). Systematic review of driving simulator validation studies. *Safety Science*, *117*, 138–151. https://doi.org/10.1016/j.ssci.2019.04.004

Zöller, I. M. (2015). Analyse des Einflusses ausgewählter Gestaltungsparameter einer Fahrsimulation auf die Fahrerverhaltensvalidität.

A.3 Extended Abstract Research Paper No. 3: User Interaction with digital twins: how comparable are simulation and reality

Authors: Czaban, M.; Sultanow, E.; Chircu, A.; Czarnecki, C.; Riedl, J.; Wengler, S. – Under review

Abstract: This paper investigates the physiological responses of individuals driving both on a real route and within a simulator designed as a digital twin of that route. The analysis of observed data patterns in stress response bio signals provide sufficient evidence of similarity to validate the driving simulation digital twin as a reliable replacement for real-world experiences in controlled and consistent settings, or when overall trends of physiological variables, rather than specific variable levels, are of interest. The findings also stress the need for optimizing the precision of digital twins in complex settings. These findings support the broader application of digital twins in fields where real world interactions are unfeasible, providing foundational insights for future digital twin design and use.

Keywords: Digital twins; Physiological measurement; Vehicle Simulation; Stress response

1 Introduction

The concept of the Digital Twin (DT) has gained increasing importance across various industries (Apte & Spanos, 2021; Barricelli et al., 2019; Jones et al., 2020). A DT represents the digital replication of a physical entity, process, or system that can accurately reflect real-world dynamics. This enables interaction with the virtual model as if it were the physical counterpart (Semeraro et al., 2021).

This concept holds transformative potential, as it opens up new possibilities for analysis, prediction, and performance optimization under different conditions—without the need for physical testing in every scenario. The advantages of a DT therefore lie not only in cost and risk reduction but also in providing new insights that can enhance decision-making processes (Attaran et al., 2023; Singh et al., 2022).

For the application of DTs to be meaningful and for their potential to be fully realized, it is essential to understand how users interact with such digital models—and whether this interaction can replicate or replace real-world experience. This leads to the central research question of whether interaction with a DT is sufficiently realistic to substitute direct experience with the physical entity.

In fields such as healthcare, manufacturing, or the automotive industry, physical testing is often time-consuming, costly, and resource-intensive (Piromalis & Kantaros, 2022; Atalay et al., 2022; Voigt et al., 2021). If DTs can serve as realistic replications, this would offer significant advantages—they could be used as flexible alternatives or complements to physical tests. Especially in early development stages, DTs can reduce costs and accelerate processes (Attaran et al., 2023). Moreover, they enable testing of complex or extreme scenarios in a safe, controlled environment that would be difficult or costly to realize in the real world (Mihai et al., 2022). The data obtained can provide valuable insights for product optimization and improvement of user experience (Lo et al., 2021).

To address the research question, this study employs a driving simulator. The physiological responses of participants were examined while driving on a real track and while driving within the DT of the same track in the simulator. The aim was to determine the degree of similarity between physiological stress responses in both environments (real-world driving vs. simulated driving) using mean value and time series analyses.

The results indicate significant correlations between both environments, supporting the applicability of the DT concept in the context of driving simulation.

2 Theoretical background and research questions

2.1 Digital twins: concept and application

Digital Twins (DTs) are understood as virtual systems or computer-generated models that replicate or “mirror” the lifecycle of a physical entity—such as an object, process, or person. A key characteristic of the DT concept is the continuous, bidirectional data integration between the physical entity and its digital counterpart (Fuller et al., 2020; Barricelli et al., 2019).

In the literature, different maturity levels of data integration are distinguished (Fuller et al., 2020; Botín-Sanabria et al., 2022):

- **Digital Model:** Simulates the physical system but without real-time data transfer.
- **Digital Shadow:** Includes a unidirectional data transfer from the physical entity to the virtual system to improve the simulation.
- **True Digital Twin:** Features bidirectional data flows, allowing the virtual system to accurately represent the physical entity, predict its behavior, and send decisions back to the real system.

DTs enable cost- and time-efficient simulations and are particularly suitable for analyzing complex or risky scenarios that are difficult or impossible to replicate in the real world (Chircu et al., 2023). The range of DT applications is broad, spanning from smart cities, transportation, manufacturing, healthcare, and product design to agriculture and societal modeling (Barricelli et al., 2019; Semeraro et al., 2021; Jones et al., 2020). For instance, in the medical field, DTs are used to develop personalized treatment strategies (Voigt et al., 2021).

In the automotive industry, DTs support vehicle design development, traffic management, and the validation of vehicle systems (Deng et al., 2023). Driving simulators are often used in this context—for example, for component testing (e.g., batteries), validation of autonomous driving functions, or studies of driver behavior such as distraction or stress (Shoukat et al., 2024; Ma et al., 2024).

Despite significant progress, challenges remain in modeling real-world complexity, validating models, and ensuring data security. Moreover, standardized reference frameworks are often lacking (Sharma et al., 2020). In many domains, the fidelity of DTs has not yet been fully achieved. For example, driving simulators often fail to capture the subtleties of real driving environments—such as sudden traffic patterns or the unpredictable behavior of pedestrians—accurately enough (Piromalis & Kantaros,

2022). This so-called simulation-to-reality gap therefore represents a central challenge (Stocco, 2022).

2.2 Simulation vs. reality

Virtual Environments (VEs) and Virtual Reality (VR) enable the replication of real-world scenarios under controlled conditions. They are used in disciplines such as psychology, design, medicine, and training to systematically study behavior and perception (Bishop & Rohrmann, 2003). This approach allows researchers to combine ecological validity (as in field studies) with experimental control (as in laboratory studies) (Loomis et al., 1999; Weibel et al., 2018).

A key criterion for the quality of a simulation is behavioral realism—the extent to which reactions within the simulation correspond to reactions in the real environment (Freeman et al., 2000; IJsselsteijn et al., 2000). Studies have shown that physiological responses in simulations often resemble those observed in real environments, while psychological responses sometimes differ considerably (Higuera-Trujillo et al., 2017; Hu et al., 2011).

In the context of driving behavior, driving simulators are already widely used. They serve to investigate parameters such as driving dynamics, attention, or stress responses (Bella, 2014; Veldstra et al., 2015). These can be validated using subjective, objective, and physiological measures (Johnson et al., 2011; Li et al., 2013). A distinction is made between absolute and relative validity (Törnros, 1998; Pawar et al., 2022): While absolute validity requires identical values in both environments, relative validity is achieved when the trends between environments are consistent.

Previous studies have shown that driving simulators elicit similar physiological patterns in heart rate, gaze behavior, and skin conductance as real driving (Johnson et al., 2011; Carter & Laya, 1998).

This leads to the research question of the present study:

To what extent do the physiological stress responses of users while driving in a Digital Twin reflect those observed in real driving contexts?

3 Method

The study was conducted using a within-subject design with a total of n = 68 participants. The sample consisted of 37 women (54.4%) and 31 men (45.6%) (M = 30.07, SD = 11.58).

For the experiment, a Digital Twin (DT) of a 23 km driving route was created, including urban, rural, and highway sections. To achieve a high level of simulator realism and to replicate the real vehicle (VW Golf 8) as accurately as possible, components from a VW Golf 7 (steering wheel, pedals, seat) were integrated into the driving simulator. These components were mounted on a D-Box motion system with three degrees of freedom. The driving environment was displayed on three 55-inch screens.

Physiological measurement:

- ECG (Electrocardiogram): Heart Rate (HR) and Heart Rate Variability (HRV; RR interval, RMSSD, SDNN)

- GSR (Galvanic Skin Response): Skin Conductance Response (SCR), Skin Conductance Level (SCL) und Peak Amplitude (PA)

Participants first completed the real-world drive, followed by the simulated drive in the DT.

Data analysis:

The data were analyzed using descriptive statistics, t-tests for individual segments, and time series analyses for the entire drive. Correlations between real and simulated measurements were also computed.

4 Results

4.1 Mean value analysis

The mean comparisons between real-world driving and driving simulation revealed nuanced differences. The GSR indicators (SCR, SCL, PA) were overall higher during the simulated drive, indicating stronger sympathetic activation and increased emotional arousal.

In contrast, heart rates (HR) were higher in certain segments of the real drive, suggesting greater physical exertion and more intense physiological strain. The RR intervals were longer during the simulation, indicating a more relaxed cardiovascular response.

The HRV parameters (RMSSD and SDNN) showed no significant differences between the two environments, suggesting comparable autonomic regulation. Overall, the findings indicate that real driving induces higher physical strain, while simulated driving elicits stronger emotional responses.

4.2 Time series analysis

To capture dynamic patterns that are smoothed out by mean values, the time series—exemplified by the SCL—were normalized. Subsequently, the Pearson correlation coefficient was calculated between the time series of the real-world drive and the simulated drive for urban and rural sections.

The results show a moderate linear similarity: the correlation for rural sections was $r = 0.34$, and for urban sections $r = 0.31$. The visual analysis of the time series revealed strong parallels in shape and progression, indicating similar dynamics of physiological stress responses in both environments.

5 Discussion and conclusion

The aim of this study was to examine the extent to which physiological parameters during real-world driving and simulated driving in a Digital Twin are comparable. The results reveal both significant similarities and differences but indicate that the driving simulator serves as a reliable tool for capturing general trends and dynamic patterns of physiological responses—though less so for exact absolute values.

The higher GSR responses in the simulation can be attributed to emotional or slightly unsettling aspects of the virtual environment, whereas the higher HR observed during the real drive reflects greater physical strain. This suggests that stress perception in

the simulator is more cognitively and emotionally driven, while in the real environment it is more physiologically and physically induced.

Mean value analyses provide initial insights but do not capture the temporal dynamics and contextual dependence of physiological processes. Time series analyses are therefore particularly suitable for identifying similarities between real and simulated drives, especially regarding temporal patterns and trends—a finding supported by the present results.

From a practical perspective, Digital Twins can complement or partially replace real-world tests, particularly in controlled, standardized, and low-risk environments. However, discrepancies still exist in complex urban scenarios, highlighting the need for further development of adaptive simulations with real-time data integration and bidirectional feedback.

Overall, the results demonstrate that Digital Twins can reproduce key physiological patterns observed in real driving. While absolute values in some parameters differ, overarching trends and stress responses can be reliably examined under controlled conditions. Based on these findings, this study contributes to the advancement and validation of Digital Twins as a tool for research, development, and training.

References

Apte, P. P., & Spanos, C. J. (2021). The Digital Twin Opportunity. MIT Sloan Management Review, 63(1), 15–17.

Atalay, M., Murat, U., Oksuz, B., Parlaktuna, A. M., Pisirir, E., & Testik, M. C. (2022). Digital twins in manufacturing: Systematic literature review for physical–digital layer categorization and future research directions. International Journal of Computer Integrated Manufacturing, 35(7), 679–705. https://doi.org/10.1080/0951192X.2021.2022762

Attaran, M., Attaran, S., & Celik, B. G. (2023). The impact of digital twins on the evolution of intelligent manufacturing and Industry 4.0. *Advances in Computational Intelligence*, *3*(3), 11. https://doi.org/10.1007/s43674-023-00058-y

Barricelli, B. R., Casiraghi, E., & Fogli, D. (2019). A survey on digital twin: Definitions, characteristics, applications, and design implications. IEEE Access, 7, 167653–167671.

Bella, F., Calvi, A., & D'Amico, F. (2014). Analysis of driver speeds under night driving conditions using a driving simulator. Journal of Safety Research, 49, 45–52. https://doi.org/10.1016/j.jsr.2014.02.007

Bishop, I., & Rohrmann, B. (2003). Subjective responses to simulated and real environments: A comparison. Landscape and Urban Planning, 65(4), 261–277. https://doi.org/10.1016/S0169-2046(03)00070-7

Botín-Sanabria, D. M., Mihaita, A. S., Peimbert-García, R. E., Ramírez-Moreno, M. A., Ramírez-Mendoza, R. A., & Lozoya-Santos, J. d. J. (2022). Digital twin technology challenges and applications: A comprehensive review. Remote Sensing, 14(6), 1335.

Carter, C. J., & Laya, O. (1998). Drivers visual search in a field situation and in a driving simulator. In A. G. Gale, I. D. Brown, C. M. Haselgrave, & S. P. Taylor (Eds.), Vision in vehicles VI. Elsevier Science Ltd.

Chircu, A., Czarnecki, C., Friedmann, D., Pomaskow, J., & Sultanow, E. (2023). Towards a digital twin of society. In Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, January 3–6.

Deng, S., Ling, L., Zhang, C., Li, C., Zeng, T., Zhang, K., & Guo, G. (2023). A systematic review on the current research of digital twin in automotive application. Internet of Things and Cyber-Physical Systems, 3, 180–191.

Freeman, J., Avons, S. E., Meddis, R., Pearson, D. E., & IJsselsteijn, W. (2000). Using behavioral realism to estimate presence: A study of the utility of postural responses to motion stimuli. Presence: Teleoperators and Virtual Environments, 9(2), 149–164. https://doi.org/10.1162/105474600566691

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technologies, challenges and open research. IEEE Access, 8, 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358

Higuera-Trujillo, J. L., López-Tarruella Maldonado, J., & Llinares Millán, C. (2017). Psychological and physiological human responses to simulated and real environments: A comparison between photographs, 360° panoramas, and virtual reality. Applied Ergonomics, 65, 398–409. https://doi.org/10.1016/j.apergo.2017.05.006

Hu, B., Ma, L., Zhang, W., Salvendy, G., Chablat, D., & Bennis, F. (2011). Predicting real-world ergonomic measurements by simulation in a virtual environment. International Journal of Industrial Ergonomics, 41(1), 64–71. https://doi.org/10.1016/j.ergon.2010.10.001

IJsselsteijn, W. A., de Ridder, H., Freeman, J., & Avons, S. E. (2000). Presence: Concept, determinants, and measurement. In B. E. Rogowitz & T. N. Pappas (Eds.), Human vision and electronic imaging V (pp. 520-529). SPIE. doi: 10.1117/12.387188

Johnson, M. J., Chahal, T., Stinchcombe, A., Mullen, N., Weaver, B., & Bédard, M. (2011). Physiological responses to simulated and on-road driving. International Journal of Psychophysiology, 81(3), 203-208. doi: 10.1016/j.ijpsycho.2011.06.012

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the digital twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36-52.

Li, J., Zhao, X., Xu, S., Ma, J., & Rong, J. (2013). The study of driving simulator validation for physiological signal measures. Procedia - Social and Behavioral Sciences, 96, 2572–2583. doi: 10.1016/j.sbspro.2013.08.288

Lo, C. K., Chen, C. H., & Zhong, R. Y. (2021). A review of digital twin in product design and development. Advanced Engineering Informatics, 48, 101297. doi: 10.1016/j.aei.2021.101297

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. Behavior Research Methods, Instruments, & Computers, 31(4), 557–564. doi: 10.3758/BF03200735

Ma, Y., Du, R., Abdelraouf, A., Han, K., Gupta, R., & Wang, Z. (2024). Driver digital twin for online recognition of distracted driving behaviors. IEEE Transactions on Intelligent Vehicles, 9(2), 3168–3180.

Mihai, S., Yaqoob, M., Hung, D. V., Davis, W., Towakel, P., Raza, M., Karamanoglu, M., Barn, B., Shetve, D., Prasad, R. V., Venkataraman, H., Trestian, R., & Nguyen, H. X. (2022). Digital twins: A survey on enabling technologies, challenges, trends and future prospects. IEEE Communications Surveys & Tutorials, 24(4), 2255–2291. doi: 10.1109/COMST.2022.3208773

Pawar, N. M., Velaga, N. R., & Sharmila, R. B. (2022). Exploring behavioral validity of driving simulator under time pressure driving conditions of professional drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 89, 29-52. doi: 10.1016/j.trf.2022.06.004

Piromalis, D., & Kantaros, A. (2022). Digital twins in the automotive industry: The road toward physical-digital convergence. Applied System Innovation, 5(4), 65. doi: 10.3390/asi5040065

Semeraro, C., Lezoche, M., Panetto, H., & Dassisti, M. (2021). Digital twin paradigm: A systematic literature review. Computers in Industry, 130, 103469.

Sharma, A., Kosasih, E., Zhang, J., Brintrup, A., & Calinescu, A. (2024). Digital twin: Generalization, characterization and implementation. Decision Support Systems, 145, 113524.

Shoukat, M. U., Yan, L., Yan, Y., Zhang, F., Zhai, Y., Han, P., ... & Hussain, A. (2024). Autonomous driving test system under hybrid reality: The role of digital twin technology. *Internet of Things*, *27*, 101301. https://doi.org/10.1016/j.iot.2024.101301

Singh, M., Srivastava, R., Fuenmayor, E., Kuts, V., Qiao, Y., Murray, N., & Devine, D. (2022). Applications of digital twin across industries: A review. *Applied Sciences*, *12*(11), 5727. https://doi.org/10.3390/app12115727

Stocco, A., Pulfer, B., & Tonella, P. (2022). Mind the gap! a study on the transferability of virtual versus physical-world testing of autonomous driving systems. IEEE Transactions on Software Engineering, 49(4), 1928-1940. 10.1109/TSE.2022.3202311

Törnros, J. (1998). Driving behavior in a real and a simulated road tunnel–a validation study. Accident; Analysis and Prevention, 30(4), 497-503. doi: 10.1016/S0001-4575(97)00099-7

Veldstra, J. L., Bosker, W. M., de Waard, D., Ramaekers, J. G., & Brookhuis, K. A. (2015). Comparing treatment effects of oral THC on simulated and on-the-road driving performance: Testing the validity of driving simulator drug research. Psychopharmacology, 232(16), 2911–2919. doi: 10.1007/s00213-015-3927-9

Voigt, I., Inojosa, H., Dillenseger, A., Haase, R., Akgün, K., & Ziemssen, T. (2021). Digital twins for multiple sclerosis. Frontiers in Immunology, 12, 669811. doi: 10.3389/fimmu.2021.669811

Weibel, R. P., Grübel, J., Zhao, H., Thrash, T., Meloni, D., Hölscher, C., & Schinazi, V. R. (2018). Virtual reality experiments with physiological measures. Journal of Visualized Experiments, 138. doi: 10.3791/58318

Appendix B: Acceptance and stress measurement using simulators

B.1 Extended Abstract Research Paper No. 4: User Acceptance of autonomous shuttle systems: A UTAUT2 -based analysis with simulated driving tests and physiological measurement

Authors: Czaban, M. & Baier, D. (20xx) – Under review

Abstract: Autonomous shuttle buses offer significant potential for improving public transportation, enhancing traffic safety, and reducing environmental impact. However, their successful implementation depends not only on technological development but also crucially on user acceptance. While previous studies have primarily investigated acceptance based on surveys of individuals without direct experience, especially in critical traffic situations, our study addresses this gap through a simulated driving test incorporating such scenarios.

Using an extended UTAUT2 model and including both physiological (electrocardiogram, galvanic skin response) and cognitive (Perceived Stress Scale, NASA-TLX, and self-developed items) stress indicators, we examined factors influencing the behavioral intention to use autonomous shuttle buses. The results show that social influence, trust and perceived risk, and perceived usefulness positively affect usage intention, while cognitive stress has a negative impact. Physiological indicators also play a role: heart-related parameters show the expected negative association with usage intention, while electrodermal activity demonstrates a positive relationship, suggesting it may reflect general arousal rather than stress alone.

These findings highlight the importance of social context and emotional responses in the acceptance of new mobility technologies. In practical terms, users' subjective sense of safety, especially in stressful situations, may be as critical as actual technical safety. This study provides a more realistic contribution to understanding user acceptance and forms a basis for further research under real-world conditions. Future studies should explore physiological responses in real-life testing environments.

Keywords: UTAUT2, User Acceptance, Autonomous Shuttle Buses, Simulation Study, Streass measurement, Physiological Measurement

1 Introduction

The use of autonomous shuttle buses offers numerous advantages. In addition to improving traffic safety and reducing emissions, they can contribute to increasing the efficiency of public transportation systems (Bansal et al., 2016; Fagnant & Kockelman, 2015; Othman, 2023). As the use of such vehicles eliminates the need for driving personnel, automated and demand-responsive operations become possible, making public transport more flexible, cost-efficient, and inclusive (Ma et al., 2021; Othman, 2020; Millonig & Fröhlich, 2018).

However, a central challenge lies in the still limited public acceptance of autonomous vehicles (Korkmaz et al., 2022; Rejali et al., 2024). Concerns about safety, a lack of trust in the technology, and the perceived loss of control contribute to a pronounced technological skepticism that hinders the diffusion of such systems. Therefore, researching the acceptance of autonomous mobility solutions is essential.

Most existing studies examine acceptance through surveys in which participants have no real user experience and base their assessments on assumptions or mental images. Furthermore, traditional questionnaire-based acceptance models are increasingly criticized for offering limited new insights and for requiring extensions through new methodological approaches (Blut et al., 2022).

The present study addresses this research gap by providing participants with an actual usage experience through a shuttle bus simulator. It combines the theoretical framework of an extended UTAUT2 model (Venkatesh et al., 2012) with physiological and cognitive stress measurements to better understand which factors influence the behavioral intention to use autonomous shuttle buses in critical situations.

In doing so, this work follows the call by Davis & Granic (2024) to extend the classical acceptance model with a NeuroIS approach while simultaneously examining whether simulators are suitable for realistic acceptance research in the field of autonomous mobility.

2 Theoretical background and research questions

2.1 Acceptance models

The study of technology acceptance has a long tradition and encompasses various theoretical models. Early approaches include the Theory of Reasoned Action (TRA; Fishbein & Ajzen, 1975) and the Theory of Planned Behavior (TPB; Ajzen, 1991), which emphasize that behavior is determined by attitudes, subjective norms, and perceived control.

The Technology Acceptance Model (TAM; Davis, 1989) is considered one of the most influential models and is based on the core constructs of Perceived Usefulness and Perceived Ease of Use, which directly influence the intention to use a technology. The later Unified Theory of Acceptance and Use of Technology (UTAUT; Venkatesh et al., 2003) integrates several earlier models and defines four key constructs: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions.

For consumer contexts, this model was extended to UTAUT2 (Venkatesh et al., 2012) by adding the factors Hedonic Motivation, Price Value, and Habit. The constructs Price Value and Habit are not included in the present study, as autonomous shuttle buses are not yet available on the mass market, making a realistic assessment of these factors impossible.

Since the use of autonomous transport involves handing over control of the vehicle to the system, the model was extended by the variables Trust and Perceived Risk (Gefen et al., 2003; Featherman & Pavlou, 2003). These factors play a crucial role, particularly in the early stages of technological diffusion (Korkmaz et al., 2022; Salonen, 2018). Trust represents the feeling of safety and reliability, whereas Perceived Risk reflects the perceived uncertainty or potential danger.

2.2 Physiological and cognitive stressreactions

Stress arises when a situation is perceived as threatening and the available resources are considered insufficient to cope with it (Lazarus & Folkman, 1985). The resulting

stress response serves to restore physiological balance (homeostasis) and can be assessed both cognitively and physiologically (Witte et al., 2021).

Cognitive stress responses can be measured using established questionnaires such as the Perceived Stress Scale (PSS-10) (Cohen et al., 1983) or the NASA-TLX (Hart & Staveland, 1988), which primarily measures workload but is closely related to the experience of stress. However, subjective measures are prone to biases, such as social desirability bias (Nederhof, 1985).

To complement cognitive assessments, physiological indicators can be used, as they capture unconscious emotional responses in real time and are more objective (Dawson et al., 2007). The most common include:

- Electrocardiogram (ECG; cardiac activity): Records cardiac activity, particularly heart rate (HR) and heart rate variability (HRV). An increased HR and decreased HRV indicate sympathetic activation and thus stress (Reinhardt et al., 2012).
- Galvanic Skin Response (GSR): Measures changes in the electrical conductance of the skin caused by sweat gland activity. Since this activity is exclusively controlled by the sympathetic nervous system (SNS), GSR is considered a direct indicator of emotional arousal. It is differentiated into tonic and phasic skin conductance (Boucsein et al., 2012).

2.3 Research gap and questions

Numerous studies investigate the acceptance of autonomous shuttle buses using established models such as TAM or UTAUT. However, most of these studies are based on hypothetical assumptions, as participants have no real usage experience. Moreover, traditional acceptance models are often criticized for providing limited explanatory insights.

Therefore, the present study extends the acceptance model by incorporating cognitive and physiological stress responses to examine whether these factors enhance the predictive power of the behavioral intention to use autonomous shuttle buses.

Based on the UTAUT2 model and previous research, the following hypotheses were formulated:

- H1: Performance Expectancy has a positive influence on Behavioral Intention
- H2: Effort Expectancy has a positive influence on Behavioral Intention
- H3: Social Influence has a positive influence on Behavioral Intention
- H4: Facilitating Conditions have a positive influence on Behavioral Intention
- H5: Hedonic Motivation has a positive influence on Behavioral Intention
- H6: Trust & Perceived Risk has a positive influence on Behavioral Intention
- H7: Perceived Usefulness has a positive influence on Behavioral Intention

Furthermore, the following hypotheses are derived to address the research question:

- H8: Cognitive Reaction (CR) has a negative influence on Behavioral Intention
- H9: Cardiac Activation (CA) has a positive influence on Cognitive Reaction
- H10: Electrodermal Activation (EA) has a positive influence on Cognitive Reaction (CR)

3 Method

3.1 Sample and simulator situations

A total of n = 104 individuals participated in the study (58 women, 46 men; M = 29.6 years, SD = 13.0).

The experiment was conducted using a custom-developed autonomous shuttle bus simulator modeled after the Navya Arma. The simulation was built on the open-source platform CARLA and displayed on a 75-inch screen, complemented by two additional monitors providing relevant information to participants.

The virtual route included four critical driving scenarios specifically designed to elicit physiological stress responses:

1. Violation of right of way by another vehicle,
2. Blocked roadway,
3. Sudden pedestrian crossing,
4. Interaction task using gesture control.

In addition to the four situations that occurred during the drive, a stress-inducing situation also took place before the drive began. Before starting the drive, participants were required to book a ticket on a smartphone and validate it in the form of a QR code at the correct location to unlock the door and be allowed to enter. Thus, the participants were exposed to a total of five situations that had the potential to induce stress.

3.2 Measurements

Physiological measurements:

- ECG indicators (Cardiac Activation, CA): Heart rate (HR), Heart rate variability (RMSSD)
- GSR indicators (Electrodermal Activation, EA): Skin Conductance Response (SCR), Skin Conductance Level (SCL)

Cognitive measurements:

- UTAUT2-Constructs (extended): (Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Hedonic Motivation, Trust & Perceived Risk, Perceived Usefulness, Behavioral Intention)
- PSS10, NASA-TLX, Single item stressmeasurement, Single item physical wellbeing

3.4 Procedure and data analysis

After a preliminary survey, measurement sensors were attached, and a two-minute baseline of physiological data was recorded. Participants then completed the simulated test drive with the five critical scenarios. Following the drive, a post-survey was conducted to assess the cognitive indicators.

Combining the five scenarios with 104 participants resulted in 520 observations. Physiological values were normalized to the baseline. To test the hypotheses, a structural equation model (SEM) was calculated.

4 Results

4.1 Descriptives

On the physiological level, the strongest reactions were observed in Scenario 1 (violation of right of way by another vehicle). Both heart rate (HR), skin conductance response (SCR), and skin conductance level (SCL) reached their highest values, indicating strong activation.

Cognitively, the ride was evaluated as a whole experience, since the corresponding variables were measured ex post. Overall, the results show a positive evaluation of the technology and low levels of perceived stress and workload. The highest mean scores were obtained for Facilitating Conditions (M = 8.41) and Effort Expectancy (M = 8.37).

4.2 Path analysis

The final model showed good fit indices (AVE > 0.5; Cronbach's $\alpha > 0.7$).

The results of the structural equation model (SEM) indicate that Behavioral Intention is significantly and positively influenced by:

- Social Influence (H3: $\beta = 0.391$, strongest effect)
- Perceived Usefulness (H7: $\beta = 0.309$)
- Facilitating Conditions (H4: $\beta = 0.128$)
- Trust & Perceived Risk (H6: $\beta = 0.114$)

The cognitive stress response (CR) showed the expected negative effect on behavioral intention (H8: $\beta = –0.179$).

Hypotheses H1 (Performance Expectancy) and H2 (Effort Expectancy) were not supported.

Unexpectedly, Hedonic Motivation (H5: $\beta = –0.106$) showed a negative effect, contrary to the hypothesis.

Regarding physiological stress variables, a significant positive relationship was found between cardiac activation (CA) and cognitive response (H9: $\beta = 0.112$). In contrast, electrodermal activation (EA) showed a negative relationship (H10: $\beta = –0.191$).

The overall model explained 61.1% of the variance in behavioral intention ($R^2 = 0.611$).

5 Discussion and conclusion

This study examines the acceptance of autonomous shuttle buses by simulating critical driving situations in a realistic environment and integrating cognitive and physiological stress responses into an extended UTAUT2 model.

The findings support previous evidence regarding the importance of classical acceptance factors: Social Influence emerged as the strongest predictor, suggesting that the social environment plays a key role in the acceptance of autonomous systems (Kapser & Abdelrahman, 2020). Similarly, Perceived Usefulness as well as Trust & Perceived Risk were confirmed as significant influencing factors (Chen, 2019; Choi & Ji, 2015).

The integration of stress responses provides an important new contribution. As expected, higher cognitive load (CR) reduced the intention to use autonomous shuttle buses, highlighting that mental strain during the ride can be a critical barrier to acceptance.

For physiological indicators, differentiated findings emerged: increased cardiac activation (CA, higher HR, lower HRV) was associated with greater subjective stress, while increased electrodermal activation (EA) was unexpectedly linked to lower perceived stress. This suggests that GSR measurement in this study reflected not specific stress, but rather general emotional arousal—potentially triggered by curiosity, attention, or positive excitement.

The unexpected negative effect of Hedonic Motivation may indicate that the critical driving situations suppressed elements of enjoyment or curiosity. Alternatively, participants who found the ride entertaining may have been more aware of its simulated nature, leading them to evaluate real-world usage differently.

Overall, the study shows that acceptance decisions are not solely driven by rational factors but are also significantly influenced by emotional and physiological processes. Integrating physiological indicators into acceptance models thus provides a valuable approach to expanding and enhancing the realism of future research on autonomous mobility.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Bansal, P., Kockelman, K. M., & Singh, A. (2016). Assessing public opinions of and interest in new vehicle technologies: An Austin perspective. *Transportation Research Part C: Emerging Technologies*, *67*, 1–14. https://doi.org/10.1016/j.trc.2016.01.019

Blut, M., Chong, A. Y. L., Tsiga, Z., & Venkatesh, V. (2022). Meta-analysis of the unified theory of acceptance and use of technology (UTAUT): challenging its validity and charting a research agenda in the red ocean. In Symposium conducted at the meeting of Association for Information Systems.

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Chen, C.-F. (2019). Factors affecting the decision to use autonomous shuttle services: Evidence from a scooter-dominant urban context. *Transportation Research Part F: Traffic Psychology and Behaviour*, *67*, 195–204. https://doi.org/10.1016/j.trf.2019.10.016

Choi, J. K., & Ji, Y. G. (2015). Investigating the Importance of Trust on Adopting an Autonomous Vehicle. *International Journal of Human–Computer Interaction*, *31*(10), 692–702. https://doi.org/10.1080/10447318.2015.1070549

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, *24*(4), 385. https://doi.org/10.2307/2136404

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Dawson, M. E., Schell, A. M., Filion, D. L., & others (2007). The electrodermal system. *Handbook of Psychophysiology*, *2*, 200–223.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part a: Policy and Practice*, *77*, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: a perceived risk facets perspective. *International Journal of Human-Computer Studies*, *59*(4), 451–474. https://doi.org/10.1016/S1071-5819(03)00111-3

Fishbein, M., & Ajzen, I. (1975). *Beliefs, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Gefen, Karahanna, & Straub (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, *27*(1), 51. https://doi.org/10.2307/30036519

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology. Human Mental Workload* (Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Kapser, S., & Abdelrahman, M. (2020). Acceptance of autonomous delivery vehicles for last-mile delivery in Germany – Extending UTAUT2 with risk perceptions. Transportation Research Part C: Emerging Technologies, 111, 210–225. https://doi.org/10.1016/j.trc.2019.12.016

Korkmaz, H., Fidanoglu, A., Ozcelik, S., & Okumus, A. (2022). User Acceptance of Autonomous Public Transport Systems (APTS): Extended UTAUT2 Model. *Journal of Public Transportation*, *23*(1). https://doi.org/10.5038/2375-0901.23.1.5

Lazarus, R., & Folkman, S. (1985). Stress and coping. *New York*, *18*(31), 34–42.

Ma, X., Liu, X., & Qu, X [Xiaobo] (2021). Public Transit Planning and Operation in the Era of Automation, Electrification, and Personalization. *IEEE Transactions on Intelligent Transportation Systems*, *22*(4), 2345–2348. https://doi.org/10.1109/TITS.2021.3064090

Millonig, A., & Fröhlich, P. (2018). Where Autonomous Buses Might and Might Not Bridge the Gaps in the 4 A's of Public Transport Passenger Needs. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 291–297). ACM. https://doi.org/10.1145/3239060.3239079

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Othman, K. (2020). Benefits of Vehicle Automation for Public Transportation Operations. *Current Trends in Civil & Structural Engineering*, *6*(5). https://doi.org/10.33552/CTCSE.2020.06.000646

Othman, K. (2023). Public attitude towards autonomous vehicles before and after crashes: A detailed analysis based on the demographic characteristics. *Cogent*

Engineering, *10*(1), Article 2156063. https://doi.org/10.1080/23311916.2022.2156063

Reinhardt, T., Schmahl, C., Wüst, S., & Bohus, M. (2012). Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). *Psychiatry Research*, *198*(1), 106–111. https://doi.org/10.1016/j.psychres.2011.12.009

Rejali, S., Aghabayk, K., Mohammadi, A., & Shiwakoti, N. (2024). Evaluating public a priori acceptance of autonomous modular transit using an extended unified theory of acceptance and use of technology model. *Journal of Public Transportation*, *26*, 100081. https://doi.org/10.1016/j.jpubtr.2024.100081

Salonen, A. (2018). Passenger's subjective traffic safety, in-vehicle security and emergency management in the driverless shuttle bus in Finland. *Transport Policy*, *61*, 106–110. https://doi.org/10.1016/j.tranpol.2017.10.011

Venkatesh, Morris, & Davis (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425. https://doi.org/10.2307/30036540

Venkatesh, Thong, & Xu (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157. https://doi.org/10.2307/41410412

Witte, M. de, Kooijmans, R., Hermanns, M., van Hooren, S., Biesmans, K., Hermsen, M., Stams, G. J., & Moonen, X. (2021). Self-Report Stress Measures to Assess Stress in Adults With Mild Intellectual Disabilities-A Scoping Review. *Frontiers in Psychology*, *12*, 742566. https://doi.org/10.3389/fpsyg.2021.742566

B.2 Extended Abstract Research Paper No. 5: Single measurement vs composite indicators for user experience research

Authors: Czaban, M.; Riedl, J.; Wengler, S. (20xx) – Under review

Abstract: This study examines the suitability of established single indicators for measuring physical and cognitive user reactions to technology interactions. Driving tests serve as the application example, conducted both with a real vehicle on a real driving route with seven segments and on an identically modeled route using a professional driving simulator.

Data were collected on galvanic skin response, electrocardiogram, salivary cortisol, and various cognitive user reactions measured via questionnaires, including measures of demand, stress, and physical well-being.

The single indicators generally showed parallel, though not entirely consistent, measurements of participants' situational activation and stress, a pattern also observed in many other studies.

Aggregating indicators to enhance stability revealed two new dimensions: physiological reactions and cognitive reactions. In the driving tests, participants perceived the simulated drive as more challenging than the real drive; however, regardless of the variations in the reaction data observed, the two dimensions, physiological reactions and cognitive reactions, remained stable in their composition across 14 different test conditions, providing a reliable basis for analyzing reaction data.

These two composite indicators are therefore recommended for use in future user tests of all types, particularly when measuring participant activation and demand.

Keywords: User Experience Measurement; Physiological Reactions; Cognitive Reactions; Composite Indicators; Stress and Activation; Human Factors

1 Introduction

User activation during technology interaction is a central topic in user experience research. Especially for emerging technologies, the empirical investigation of user expectations, attitudes, and behaviors is essential.

An example of this can be found in driving tests within the automotive industry, which are increasingly conducted in simulators (Caird & Horrey, 2011). Driving simulators offer several advantages: in addition to providing controlled conditions (Hussain et al., 2019; Winter et al., 2012), they eliminate the need for specially trained test drivers. Moreover, critical driving situations can be reproduced under standardized conditions (Brookhuis & Waard, 2010; Mansi et al., 2021).

Driving tests are often perceived as stressful by participants (Engström et al., 2005). Accordingly, reactions are typically assessed through a combination of self-reported stress experiences (Hill & Boyle, 2007) and physiological measurements such as heart based measures, galvanic skin response, or cortisol (Koohestani et al., 2019; Li et al., 2013).

However, current research shows heterogeneous results across measurement methods, revealing a research gap regarding integrated and reliable composite indicators. The aim of the present study is therefore to develop composite indicators with higher explanatory power and stability for both academic and applied research.

2 Theoretical background and research questions

Usability is defined according to ISO as the effectiveness, efficiency, and satisfaction with which a system enables users to achieve specific goals. The measurement of user experience (UX) is based on objective user data, satisfaction assessments, and/or strain indicators.

In the automotive sector—particularly regarding Human-Machine Interfaces (HMIs)—these measurements play a crucial role, as usability strongly influences user acceptance (Albers et al., 2020; Biassoni & Gnerre, 2024). However, many HMIs are only partially intuitive for first-time users (S.-C. Lin et al., 2018; Orlovska et al., 2019).

There are various approaches to measuring user experience, each with its own advantages and limitations (Ganglbauer et al., 2009). Objective data such as braking behavior, for example, neglect intra-individual processes (Wynne et al., 2019). Physiological measurements (e.g., heart rate, galvanic skin response, cortisol) reflect activation or stress and are free from perception bias, but they require higher effort and do not always produce consistent results (Arza et al., 2019; Mauri et al., 2010). Questionnaires such as the NASA-TLX (Hart, 2006) or the Short Stress State Questionnaire (Helton, 2004) provide valuable insights into cognitive reactions but are susceptible to bias (Nederhof, 1985).

Previous research often reports inconsistent results across methods, posing a challenge for interpretation. Therefore, current studies advocate combining performance-based, subjective, and physiological data into valid, multidimensional UX indicators (Apraiz Iriarte et al., 2021; Leis & Lautenbach, 2020; Yu et al., 2016).

The assessment of stress responses is particularly suitable for analyzing participant reactions. Stress arises when situational demands exceed an individual's capabilities (Lazarus, 1990; Selye, 1980). It can be experienced as positive (eustress) or negative (distress). In the context of driving tests, stress is generally understood as distress—strain associated with loss of control and overload (Francis, 2018; Healey & Picard, 2005). The individual stress experience varies between persons and can be assessed on both cognitive and physiological levels (Witte et al., 2021).

The most commonly used physiological indicators in driving tasks are the galvanic skin response (GSR) and the electrocardiogram (ECG). The GSR measures arousal and indicates activation independent of valence (Caruelle et al., 2019). It can be divided into a phasic component (Skin Conductance Response, SCR) and a tonic component (Skin Conductance Level, SCL) (Andreassi, 2010).

The ECG allows for the measurement of heart rate (HR) and heart rate variability (HRV). HR refers to the number of heartbeats per minute, which increases under stress (Reinhardt et al., 2012). HRV reflects the variation in heartbeat intervals, which decreases as stress increases (Bernardi et al., 2000). On a biochemical level, cortisol serves as a direct marker of stress (Liebherr et al., 2021; Dickerson & Kemeny, 2004).

On the cognitive level, instruments such as the NASA-TLX (Hart, 2006), the Short Stress State Questionnaire (Helton, 2004), and visual analogue scales (Arza et al., 2019; Kabilmiharbi et al., 2022) are commonly used.

The main challenge in using physiological and cognitive indicators lies in the high heterogeneity of measurement methods, which limits reliability (Böhler et al., 2021). Therefore, the present study aims to develop stable combined indicators that integrate physiological and cognitive responses.

The study is based on the following hypotheses:

- H1: Physiological and cognitive stress indicators correlate positively with each other.
- H2: As situational demands increase, the combined stress indicator also increases.
- H3: The composition of the combined stress indicator remains stable across different situations.
- H4: The combined stress indicator and its physiological subcomponents correlate positively with cortisol levels.

The aim of this study is to establish robust, situation-dependent stress indicators that strengthen future UX research methodologically and improve the validity of user studies.

3 Method

The study sample consisted of n = 68 participants, making it larger than in comparable studies (e.g., Fors et al., 2013: n = 20; Li et al., 2013: n = 15; Johnson et al., 2011: n = 24). The sample included 37 women (54.4%) and 31 men (45.6%). Participants were between 18 and 63 years old (M = 30.07; SD = 11.58). The study followed a within-subject design, meaning that all participants completed both a real-world drive and a simulator drive, with the real drive always taking place first.

To capture different situational demands in the driving context (Healey & Picard, 2005), the driving route consisted of a 23 km circuit divided into seven defined segments of varying complexity (urban, rural, and highway sections). The real route was replicated 1:1 in a professional driving simulator.

Assessed cognitive constructs/indicators:

- Perceived situational demand: Single-item question “How well did you manage operating the vehicle?” (Vehicle Operation).
- Cognitive Load: NASA-TLX (6 items: mental, physical, temporal demand, performance, effort, frustration) (Hart, 2006; Yahoodik et al., 2020).
- Stress: Shortened SSSQ, 12 items across three subdimensions (Distress, Worry, Engagement) based on Helton, 2004; items selected based on highest factor loadings.
- Self-reported stress: Single-item visual analogue scale (“To what extent did you experience stress during the ride?”) (Barré et al., 2017).
- Physical Wellbeing: Single-item (“How was your physical wellbeing?”) – allowing differentiation between eustress and distress.

Scaling was conducted on a decimal scale from 0–10, enabling parametric analyses, increasing variance, and providing intuitive understanding for participants (Lewis, 2021; Chyung et al., 2018; Dawes, 2002; Leung, 2011).

Assessed physiological indicators:

- GSR: Skin Conductance Level (SCL, tonic), Skin Conductance Response (SCR, phasic), Peak Amplitude (PA) (Boucsein et al., 2012).
- ECG: Heart rate (HR) and heart rate variability parameters: RR interval, RMSSD, SDNN.
- Salivary cortisol

For data analysis, each of the seven route segments per person and condition was treated as a separate event, resulting in n = 476 observations per condition. To develop composite indicators, exploratory principal component analyses (PCA) were conducted.

4 Results

The descriptive analyses show that participants experienced the simulated drive as significantly more demanding and stressful than the real drive. This was reflected in higher NASA-TLX scores, lower physical wellbeing, and noticeably stronger physiological stress responses (e.g., SCL).

Testing H1:

The correlation analyses largely supported the hypothesis. Most physiological and cognitive single measures correlated significantly in the expected direction. However, the SSSQ subdimensions Worry and Engagement proved unsuitable for valid stress measurement. Therefore, the following valid single indicators were used for further analyses: SCR, HR, NASA-TLX, Physical Wellbeing, and Self-reported Stress.

Testing H2 and H3:

These hypotheses concern the formation of the composite indicators. Based on the previously selected individual indicators, the principal component analysis revealed an identical and stable two-factor structure:

1. Physiological Reaction (PR): composed of SCR and HR.
2. Cognitive Reaction (CR): composed of NASA-TLX, Physical Wellbeing (recoded so that higher values indicate stronger discomfort), and Self-reported Stress.

The stability of this structure was confirmed both across the entire drive and for all 14 route segments (7 real and 7 simulated). Both indicators correlated significantly with situational demand, thereby supporting H2: the poorer the driving task was performed, the stronger the stress responses on both cognitive and physiological levels.

Testing H4:

For validation, the developed indicators were correlated with cortisol levels. The hypothesis was supported only under the sufficiently demanding condition of the simulated drive: both PR and CR showed highly significant positive correlations with cortisol levels. In contrast, no significant correlations were found during the real drive,

which was less demanding. Overall, the cognitive indicator CR (r = .420) showed a markedly stronger correlation with cortisol levels than the physiological indicator PR (r = .196).

5 Discussion and conclusion

The findings of this study highlight the methodological weaknesses and limitations of relying solely on single indicators in UX research. The heterogeneity of individual variables can lead to misinterpretations.

Based on the results, the aggregation of valid single measures allows for the identification of overarching, stable, and reliable dimensions of user response. The use of the composite indicators Physiological Reaction (PR) and Cognitive Reaction (CR) is therefore recommended.

Across 14 different situations, this robust two-factor structure consistently emerged, providing researchers and practitioners with a reliable and methodologically sound foundation for measuring strain and stress. The use of the CR indicator, in particular, is recommended for future UX testing, as it can be assessed with only eight questions, is easy to administer, and minimizes participant burden.

Interestingly, the questionnaire-based indicator CR proved to be a better predictor of the biochemical stress marker cortisol than the physiologically measured indicator PR. Consequently, it can be concluded that, when a multimodal approach is not feasible, a carefully designed questionnaire may still yield valid insights into users' stress responses.

This study provides a practical, application-oriented solution for assessing physiological and cognitive user experience responses and offers a solid foundation for addressing the measurement challenges associated with single indicators.

References

Albers, D., Radlmayr, J., Loew, A., Hergeth, S., Naujoks, F., Keinath, A., & Bengler, K. (2020). Usability Evaluation—Advances in Experimental Design in the Context of Automated Driving Human–Machine Interfaces. *Information*, *11*(5), 240. https://doi.org/10.3390/info11050240

Andreassi, J. L. (2010). *Psychophysiology*. Psychology Press. https://doi.org/10.4324/9780203880340

Apraiz Iriarte, A. A., Erle, G. L., & Etxabe, M. M. (2021). Evaluation User Experience with physiological monitoring: A systematic literature review. *DYNA NEW TECHNOLOGIES*, *8*(1), [20 p.]-[20 p.]. https://doi.org/10.6036/NT10072

Arza, A., Garzón-Rey, J. M., Lázaro, J., Gil, E., Lopez-Anton, R., La Camara, C. de, Laguna, P., Bailon, R., & Aguiló, J. (2019). Measuring acute stress response through physiological signals: Towards a quantitative assessment of stress. *Medical & Biological Engineering & Computing*, *57*(1), 271–287. https://doi.org/10.1007/s11517-018-1879-z

Barré, R., Brunel, G., Barthet, P., & Laurencin-Dalicieux, S. (2017). The visual analogue scale: An easy and reliable way of assessing perceived stress. *Quality in Primary Health Care*, *1*(1), 1–5.

Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., & Sleight, P. (2000). Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology*, *35*(6), 1462–1469. https://doi.org/10.1016/S0735-1097(00)00595-7

Biassoni, F., & Gnerre, M. (2024). Understanding Elderly Drivers' Perception of Advanced Driver Assistance Systems: A Systematic Review of Perceived Risks, Trust, Ease of Use, and Usefulness. *Geriatrics (Basel, Switzerland)*, *9*(6). https://doi.org/10.3390/geriatrics9060144

Böhler, H., Germelmann, C. C., Baier, D., Woratschek, H., Diller, H., & Kirchgeorg, M. (2021). *Marktforschung*. W. Kohlhammer GmbH. https://doi.org/10.17433/978-3-17-032249-3

Boucsein, W. (2012). *Electrodermal Activity*. Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Brookhuis, K. A., & Waard, D. de (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident; Analysis and Prevention*, *42*(3), 898–903. https://doi.org/10.1016/j.aap.2009.06.001

Caird, J. K., & Horrey, W. J. (2011). Twelve Practical and Useful Questions about Driving Simulation.

Caruelle, D., Gustafsson, A., Shams, P., & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research*, *104*, 146–160. https://doi.org/10.1016/j.jbusres.2019.06.041

Chyung, S. Y. Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement*, *57*(5), 38–48. https://doi.org/10.1002/pfi.21763

Dawes, J. (2002). Five point vs. eleven point scales: Does it make a difference to data characteristics? *Australasian Journal of Market Reserach*(10(1)), 39–47.

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, *130*(3), 355–391. https://doi.org/10.1037/0033-2909.130.3.355

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 97–120. https://doi.org/10.1016/j.trf.2005.04.012

Fors, C., Ahlström, C., & Anund, A. (2013). *Simulator validation with respect to driver sleepiness and subjective experiences : final report of the project SleepEYE II, part 1* (ViP publication : ViP - Virtual Prototyping and Assessment by Simulation 2013-1). Swedish National Road and Transport Research Institute, Human-vehicle-transport system interaction.

Francis, A. L. (2018). The Embodied Theory of Stress: A Constructionist Perspective on the Experience of Stress. *Review of General Psychology*, *22*(4), 398–405. https://doi.org/10.1037/gpr0000164

Ganglbauer, E., Schrammel, J., Schwarz, S., & Tscheligi, M. (2009). Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Healey, J. A., & Picard, R. W [R. W.] (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. https://doi.org/10.1109/TITS.2005.848368

Helton, W. S. (2004). *PsycTESTS Dataset.* https://doi.org/10.1037/t57758-000

Hill, J. D., & Boyle, L. N. (2007). Driver stress as influenced by driving maneuvers and roadway conditions. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(3), 177–186. https://doi.org/10.1016/j.trf.2006.09.002

Hussain, Q., Alhajyaseen, W. K., Pirdavani, A., Reinolsmann, N., Brijs, K., & Brijs, T. (2019). Speed perception and actual speed in a driving simulator and real-world: A validation study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *62*, 637–650. https://doi.org/10.1016/j.trf.2019.02.019

International Organization for Standardization (ISO). *ISO 9241*.

Johnson, M. J., Chahal, T., Stinchcombe, A., Mullen, N., Weaver, B., & Bédard, M. (2011). Physiological responses to simulated and on-road driving. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *81*(3), 203–208. https://doi.org/10.1016/j.ijpsycho.2011.06.012

Kabilmiharbi, N., Kamaliana Khamis, N., & Azila Noh, N. (2022). Commonly Used Assessment Method to Evaluate Mental Workload for Multiple Driving Distractions: A Systematic Review. *Iranian Journal of Public Health*, *51*(3), 482–494. https://doi.org/10.18502/ijph.v51i3.8924

Koohestani, A., Kebria, P. M., Khosravi, A., & Nahavandi, S. (2019). Drivers Awareness Evaluation using Physiological Measurement in a Driving Simulator. In *2019 IEEE International Conference on Industrial Technology (ICIT)* (pp. 859–864). IEEE. https://doi.org/10.1109/ICIT.2019.8755188

Lazarus, R. S. (1990). Theory-Based Stress Measurement. *Psychological Inquiry*, *1*(1), 3–13. https://doi.org/10.1207/s15327965pli0101_1

Leis, O., & Lautenbach, F. (2020). Psychological and physiological stress in non-competitive and competitive esports settings: A systematic review. *Psychology of Sport and Exercise*, *51*, 101738. https://doi.org/10.1016/j.psychsport.2020.101738

Leung, S.-O. (2011). A Comparison of Psychometric Properties and Normality in 4-, 5-, 6-, and 11-Point Likert Scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., & Andreoli, A. (1993). Development of the Perceived Stress Questionnaire: A new tool for psychosomatic research. *Journal of Psychosomatic Research*, *37*(1), 19–32. https://doi.org/10.1016/0022-3999(93)90120-5

Lewis, J. R. (2021). Measuring User Experience With 3, 5, 7, or 11 Points : Does It Matter? *Human Factors*, *63*(6), 999–1011. https://doi.org/10.1177/0018720819881312

Li, J., Zhao, X., Xu, S., Ma, J [Jianming], & Rong, J. (2013). The Study of Driving Simulator Validation for Physiological Signal Measures. *Procedia - Social and*

Behavioral Sciences, *96*, 2572–2583. https://doi.org/10.1016/j.sbspro.2013.08.288

Liebherr, M., Mueller, S. M., Schweig, S., Maas, N., Schramm, D., & Brand, M. (2021). Stress and Simulated Environments: Insights From Physiological Marker. *Frontiers in Virtual Reality*, *2*, Article 618855. https://doi.org/10.3389/frvir.2021.618855

Lin, S.-C., Hsu, C.-H., Talamonti, W., Zhang, Y., Oney, S., Mars, J., & Tang, L. (2018). Adasa. In P. Baudisch, A. Schmidt, & A. Wilson (Eds.), *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (pp. 531–542). ACM. https://doi.org/10.1145/3242587.3242593*

Mansi, S. A., Barone, G., Forzano, C., Pigliautile, I., Ferrara, M., Pisello, A. L., & Arnesano, M. (2021). Measuring human physiological indices for thermal comfort assessment through wearable devices: A review. *Measurement*, *183*, 109872. https://doi.org/10.1016/j.measurement.2021.109872

Mauri, M., Magagnin, V., Cipresso, P., Mainardi, L., Brown, E. N., Cerutti, S., Villamira, M., & Barbieri, R. (2010). Psychophysiological signals associated with affective states. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2010*, 3563–3566. https://doi.org/10.1109/IEMBS.2010.5627465

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Orlovska, J., Novakazi, F., Wickman, C., & Soderberg, R. (2019). Mixed-Method Design for User Behavior Evaluation of Automated Driver Assistance Systems: An Automotive Industry Case. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 1803–1812. https://doi.org/10.1017/dsi.2019.186

Reinhardt, T., Schmahl, C., Wüst, S., & Bohus, M. (2012). Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). *Psychiatry Research*, *198*(1), 106–111. https://doi.org/10.1016/j.psychres.2011.12.009

Selye, H. (1980). *Selye's Guide to Stress Research*. Van Nostrand Reinhold.

Winter, D. J. de, van Leeuwen, P., & Happee, R. (2012). Advantages and Disadvantages of Driving Simulators: A Discussion. *Proceedings of Measuring Behavior*, 47–50.

Witte, M. de, Kooijmans, R., Hermanns, M., van Hooren, S., Biesmans, K., Hermsen, M., Stams, G. J., & Moonen, X. (2021). Self-Report Stress Measures to Assess Stress in Adults With Mild Intellectual Disabilities-A Scoping Review. *Frontiers in Psychology*, *12*, 742566. https://doi.org/10.3389/fpsyg.2021.742566

Winter, D. J. de, van Leeuwen, P., & Happee, R. (2012). Advantages and Disadvantages of Driving Simulators: A Discussion. *Proceedings of Measuring Behavior*, 47–50.

Wynne, R. A., Beanland, V., & Salmon, P. M. (2019). Systematic review of driving simulator validation studies. *Safety Science*, *117*, 138–151. https://doi.org/10.1016/j.ssci.2019.04.004

Yahoodik, S., Tahami, H., Unverricht, J., Yamani, Y., Handley, H., & Thompson, D. (2020). Blink Rate as a Measure of Driver Workload during Simulated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1825–1828. https://doi.org/10.1177/1071181320641439

Yu, Y. J., Yang, Z., Oh, B.-S., Yeo, Y. K., Liu, Q., Huang, G.-B., & Lin, Z. (2016). Investigation on driver stress utilizing ECG signals with on-board navigation systems in use. In *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICARCV.2016.7838780

B.3 Research Paper No. 6: Scent and stress: The role of lavender and perception in simulated driving scenarios

Authors: Czaban, M., Mohr, S.V., Riedl, J., Wengler, S. (2025)

Citation: Czaban, M., Mohr, S. V., Riedl, J., & Wengler, S. (2025). Scent and Stress: The Role of Lavender and Perception in Simulated Driving Scenarios. In: Davis, F.D., Riedl, R., vom Brocke, J., Léger, PM., Randolph, A.B., Müller-Putz, G.R. (eds) Information Systems and Neuroscience. NeuroIS 2025. Lecture Notes in Information Systems and Organisation, Cham

URL-Preprint: https://www.neurois.org/wp-content/uploads/2025/05/NeuroIS-Retreat-2025-Preprint-Proceedings.pdf

Abstract: Stress increases the risk of road accidents by impairing driving performance. Although lavender is known for its calming effects, it remains unclear whether its use can reduce both cognitive (self-reported) and physiological stress in driving situations.

In a simulated driving scenario, participants were randomly assigned to an experimental group exposed to lavender or to a control group. Physiological responses were measured via skin conductance response (SCR) and heart rate (HR), while on subjective level the NASA-TLX and a single-item self-report measure were attached.

Contrary to expectations, lavender exposure generally elevated both physio-logical and self-reported stress levels. However, conscious perception of the scent moderated this effect, with participants who were aware of the lavender reporting significantly lower subjective stress.

These findings suggest that the effectiveness of lavender depends on cognitive awareness, offering novel insights into olfactory interventions in high-stress environments.

Keywords: driving simulation · stress measurement · olfactory stimulation · lavender scent · cognitive load

1 Introduction

Over 90% of traffic accidents are attributable to human error (Singh, 2015), with elevated stress levels playing a significant role (Brookhuis & Waard, 2010; Magaña et al., 2020). According to the Yerkes-Dodson law (Yerkes & Dodson, 1908), moderate stress (eustress) enhances performance, whereas high stress, known as distress, impairs cognitive and motor functions, thereby increasing accident risk (Beanland et al., 2013; Pluut et al., 2022).

Stress can be understood as a psychological and biological/physiological phenomenon (Riedl, 2012), and manifests physiologically, for example, through increased heart rate and altered skin conductance (Andreassi, 2010).

Sensory stimuli, particularly scents, influence psychophysiology (Li et al., 2024): While peppermint has been shown to have a cognitively stimulating effect (Raudenbush et al., 2009), studies indicate that lavender has stress-reducing properties (Ludvigson & Rottman, 1989; Moss et al., 2023). This raises the research question of whether the targeted use of scent in critical driving situations can reduce drivers' stress levels

(Castiello et al., 2006). The effect of scents in the driving context remains insufficiently explored (X. Jiang et al., 2023; Moss et al., 2023).

Driving simulators provide an alternative to examine critical scenarios without endangering participants (Galante et al., 2018; Pawar & Velaga, 2020). To capture stress responses more holistically, recent research in the NeuroIS field has emphasized the complementary use of physiological and psychological measures (Dimoka et al., 2012), as this combination can improve the explanation and prediction of (techno)stress (Tams et al., 2014).

These theoretical considerations lead to the central research question of this study: Can the targeted use of lavender scent reduce physiological and cognitive stress during critical driving situations, and does conscious perception of the scent moderate this effect?

In our study, one group was systematically exposed to lavender scent during the driving task without being informed. The control group drove without scent exposure. Stress levels were measured during the test, and at the end of the test, participants were asked whether they had perceived the scent.

This approach follows calls in IS research to combine behavioral and physiological data in order to better understand the dynamic interaction between person and environment, also referred to as “measurement pluralism” (Fischer & Riedl, 2017). Furthermore the usage of physiological measurements allows the provision of real-time information on user’s stress state (vom Brocke et al., 2020).

2 Theoretical background and hypothesis development

2.1 Stress measurement

Stress arises when there is an imbalance between individual capabilities and situational demand (Cannon, 1929; Koolhaas et al., 2011; Zhou et al., 2022). Depending on the extent, a distinction is made between eustress (positive) and distress (negative) (Lazarus, 1966; Selye, 1950). Various physiological and subjective methods exist to measure stress responses (Witte et al., 2021). Among the most frequently studied physiological measurement methods are Galvanic Skin Response (GSR) and Electrocardiogram (ECG) (Caruelle et al., 2019; Giannakakis et al., 2022; Sharma & Gedeon, 2012) which can capture emotional and cognitive states (Riedl et al., 2010).

GSR measures skin conductance, which is influenced by the activity of eccrine sweat glands and is exclusively controlled by the sympathetic nervous system. Stress induced activation is reflected in short-term changes in conductance (skin conductance response, SCR) or an increased average skin conductance level over time (skin conductance level, SCL) (Andreassi, 2010; Boucsein, 2012).

ECG records the electrical activity of the heart, allowing for the analysis of heart rate (HR) and heart rate variability (HRV). An increased HR reflects heightened sympathetic activation, whereas a reduced HRV correlates with decreased parasympathetic regulation and an elevated stress level (Berntson et al., 2008; Riedl & Léger, 2016).

In stress research, physiological measurements are combined with subjective, questionnaire-based data to achieve more reliable results (Becker et al., 2023)

because physiological tools can provide reliable data which are difficult or impossible to record through traditional tools as e.g. self-reports and can capture unconscious processes with direct responses from the human body (Dimoka et al., 2012; Riedl et al., 2010). Furthermore, the validity of research findings can be improved by combining two or more methods (Riedl et al., 2010). The combination of physiological data and self-reported data is common in NeuroIS research to examine systems (vom Brocke et al., 2020).

A commonly used questionnaire in this context is the NASA Task Load Index (Hart & Staveland, 1988), which measures mental workload, a factor correlated with stress (Hines Duncliffe et al., 2018). Additionally, self-assessments using single-item measurements can be employed (Arza et al., 2019). Various authors have applied these indicators in the context of real or simulated driving (GSR e.g.: Daviaux et al., 2020; Healey & Picard, 2005; Lanata et al., 2015; ECG e.g: Darzi et al., 2018; Kerautret et al., 2023; Zhou et al., 2022; NASA-TLX e.g.: Foy & Chapman, 2018; Sugiono et al., 2018; Yahoodik et al., 2020; Single Items e.g.: Dogan et al., 2019; Lazaro et al., 2022; Lee & Chung, 2017).

2.2 The effect of scent

Scents influence emotions, cognitive processes, and behavior. They can activate memories (Lopis et al., 2023) enhance mood (Rachel S. Herz, 2009; Rachel S. Herz et al., 2004) and modulate cognitive functions (Deivanayagame et al., 2020; Ilm-berger et al., 2001; Michael et al., 2003). Scent molecules are absorbed with each breath and directly reach cortical regions (Royet et al., 2003). Unlike visual or auditory stimuli, scents act directly on the limbic system, explaining their unconscious effects and measurable physiological responses (Alaoui-Ismaïli et al., 1997; Bensafi et al., 2002; R. S. Herz & Engen, 1996; Nomura et al., 2016; Torii et al., 1988).

Accordingly, the effect of the scent is expected to manifest independently of instrumental means-end relationships, as conceptualized in expectancy-based models of motivation (Vroom, 1964; Wigfield & Eccles, 2000).

Due to their link to the autonomic nervous system, scents can trigger various re-actions. Pleasant scents affect both mood (Dmitrenko et al., 2020; Jeon et al., 2014; Roschk & Hosseinpour, 2020) and arousal levels (Joussain et al., 2014; Tisserand, 1988; Warm et al., 1991). While peppermint has a stimulating effect, vanilla and lavender are considered calming (Buchbauer et al., 1991; Ghavami et al., 2022; Luca & Botelho, 2021; Moss et al., 2003; Mustafa et al., 2016).

Several empirical studies have found that scents can positively influence driving behavior, for example, by enhancing attention (Raudenbush et al., 2009), reducing drowsiness (X. Jiang et al., 2023; X. Jiang et al., 2024; Yoshida et al., 2011), decreasing anger, and improving well-being (Dmitrenko et al., 2020; Moss et al., 2023). Some studies report a sedative physiological effect of lavender (Diego et al., 1998; Heuberger et al., 2004; Koulivand et al., 2013; Kuroda et al., 2005)

2.3 Research gap & hypothesis building

Although stress affects driving performance, empirical data on the effect of scent on driver stress are lacking. This study addresses this gap by examining the impact of

lavender scent, which has been associated with a reduction in heart rate (Heuberger et al., 2004), decreased sympathetic activation (Koulivand et al., 2013) and lower subjective stress levels (Lehrner et al., 2005; Moss et al., 2023). We therefore expect differences between experimental groups with and without scent exposure. During a simulated drive, participants are confronted at defined time intervals with five potentially stress-inducing driving situations (e.g., “a child unexpectedly runs onto the road”). Although alternative effects of lavender (e.g., stress-enhancing effects) cannot be entirely ruled out, we formulate a directed hypothesis based on prior empirical findings suggesting an anxiolytic effect. This approach follows the principles of hypothesis-driven experimental research and allows for a clear test of theoretical predictions.

We postulate:

H1: The controlled use of lavender scent during a driving task reduces measurable stress levels in participants, in the form of a decrease in H1.1: physiological stress indicators and H1.2: self-report stress indicators.

In addition to testing the direct effects of lavender scent, this study examines the moderating role of conscious scent perception. According to expectancy theory and cognitive appraisal models, the conscious perception and interpretation of a stimulus can shape its emotional and physiological impact (Kirsch, 1997; Lazarus & Folk-man, 1984). We therefore hypothesize:

H2: The conscious perception of the lavender scent does not moderate its effect on stress levels (R. S. Herz & Engen, 1996; Nomura et al., 2016),

H2.1: in physiological stress indicators; H2.2: in self-report stress indicators.

3 Method

3.1 Experimental design

Participants The study follows a between-subjects design with 26 participants randomly assigned to two groups. One group was exposed to lavender scent, while the control group was not subjected to any scent exposure. The sample consists of 14 women (53.8%) and 12 men (46.2%) with an average age of 25.8 years (SD = 7.84; range: 19–61). Regarding place of residence, 34.6% identify as rural residents, 46.2% as residents of small and medium-sized towns, and 19.2% as city dwellers. While the sample size of 26 participants is relatively small, it provides preliminary insights into the effects of lavender scent on stress responses in driving contexts. Future research with larger sample sizes is needed to validate these findings and improve generalizability. For the analysis we divided the sample into three groups: No scent, scent with perception and scent with no perception. There were no statistically significant differences between the groups with respect to age, gender, or place of residence (age: $F(2, 23) = 1.118$, $p = .344$; gender: $\chi^2(2) = 0.63$, $p = .731$; residence: $\chi^2(4) = 0.69$, $p = .952$).

Physiological Measurement The GSR data were recorded using a Shimmer 3 GSR+ device (Exosomatic, direct current; Boucsein et al., 2012). Electrodes were placed on the palm. Skin conductance response (SCR) was measured as peaks per minute.

Heart rate (HR) in beats per minute (bpm) was recorded via ECG using a Polar H10 chest strap sensor.

Furthermore, additional indicators such as skin conductance level and heart rate variability were collected. However, previous studies conducted by our group suggest that, in particular, SCR and HR tend to cluster together as a single indicator of physiological reaction (Czaban et al., 20XXb).

Cognitive Measurement To assess cognitive stress perception, we used the NASA-TLX (Hart, 2006; Hart & Staveland, 1988) as well as a single-item measurement (stress), in which participants were asked: "How much stress did you experience during the entire drive?". It is important to note that the NASA-TLX is designed to assess cognitive workload. Although cognitive workload and stress represent conceptually distinct constructs, prior research has shown that they are often positively correlated (Alsuraykh et al., 2019).

All questions were recorded on a decimal scale (0–10) to enhance intuitive understanding (Lewis, 2021), increase data variance (Dawes, 2002), ensure normal distribution (Leung, 2011) and enable the application of parametric test (Chyung et al., 2018).

Perception At the end of the test, participants were asked dichotomously whether they had perceived the scent by questioning: ”Did you notice a scent during the experiment?”. It should be noted that the survey was conducted in the presence of the test administrators, allowing participants to openly discuss any notable observations. At no point were the perception of other scents or potential confounding variables raised, suggesting that the participants either perceived the test stimulus (lavender scent) or no scent at all.

Additionally, participants were asked about the type of scent they perceived, how pleasant they found it, how familiar the scent was to them, and how intense they perceived it to be. However, these aspects are not discussed further in the manuscript, as they were not part of our research question.

The cold nebulization scent diffuser was set to an intensity level that, based on prior pilot studies, was perceived as pleasant by participants and ensured that at least half of them detected the scent.

3.2 Materials, driving task and data processing

The experiment was conducted using a medium-fidelity driving simulator (Wynne et al., 2019). **Aroma Conditioning:** In the test group, lavender scent was dispersed during the experiment using an “AromaStreamer 450” (Reima Air Concept). **Procedure:** After a preliminary survey, measurement devices were attached. To reduce simulator sickness, the experiment began with an adaptation phase (Hoffmann et al., 2003), followed by a 1.5-minute baseline recording. The drive lasted an average of 7.5 minutes and included five critical events designed to induce and control stress situations (see Table 1). The critical driving scenarios used in this study were developed for a previous study by our research group. Both an expert rating conducted to select the scenarios and user data indicated that the situations were discriminative with respect to the level of stress they induced (Czaban et al., 20XXa). A scenario-

specific analysis was not conducted in the present study, as the focus was on the overall effect of scent exposure, which did not vary across the different driving scenarios. However, the deliberate inclusion of driving situations with varying levels of user demand ensures that the observed effects of scent exposure cannot be attributed to a methodological artifact resulting from the arbitrary selection of a single scenario.

The total experiment duration was approximately 40 minutes per participant.

Table 1. Overview of the critical driving situations

Order	Situation 1	Situation 2	Situation 3	Situation 4	Situation 5
Event	Child runs on the road	Driving over speedbumps	U-turn	Driving over a pothole	Car taking the right of way
Picture					
Feedback	Person screaming	Shaking of driver's seat	-	Shaking of driver's seat	Honking of the car
Mean Stresslevel Rating (0-10)	7.08	3.92	5.72	2.78	7.88

Cognitive workload was assessed once at the end of the test for all 26 participants. Since no technical difficulties (e.g., sensor Bluetooth disconnection) were encountered during data collection, the dataset was complete and no participants had to be excluded from the analysis. Due to the five critical driving situations per person, a total of 130 physiological single episodes could be analyzed (unpivoting).

Data analysis (SPSS 29) was conducted using Principal Component Analysis (PCA), Levene's test, and t-tests.

4 Results

The use of individual indicators often leads to inconsistent and heterogeneous results (Arza et al., 2019), which is why composite indicators can be used to enhance the robustness and interpretability of the findings. We calculated a mean index from the NASA-TLX items (Cronbach's α = .761), where higher values indicate higher cognitive workload.

To improve the stability of single measurements, we computed more reliable overall indicators using PCA (Czaban et al., 20XXb). NASA-TLX and stress loaded onto one factor, while SCR and HR formed another. These two factors explained 79.59% of the variance of the original items.

We derived a combined indicator, Cognitive Reaction (CR), from the unweighted mean values of NASA-TLX and stress, resulting in a range of 0.92–7.50 with a mean of M = 4.34.

Since the physiological variables SCR and HR have different value ranges (SCR: 4–20.13, M = 11.35; HR: 60.4–131.68, M = 90.98), HR values were adjusted by dividing by 8.01 to match the mean of SCR (for methodology, see (Czaban et al., 20XXb)). The resulting Physiological Reaction (PR) indicator had a range of 5.97–17.84, with a mean of M = 11.35.

For further analysis, our dataset includes three groups: "no scent exposure" (A), "scent exposure without perception" (B1), and "scent exposure with perception" (B2). The Levene's test yielded significance values of p = .579 for PR and p = .131 for CR, indicating homogeneity of variance across groups. Table 2 presents the mean values of PR and CR for the three groups.

Table 2. Means of physiological reaction and cognitive reaction (with/without perception)

		PR	CR
A NoScent		10.9	4.3
B Scent		11.9	4.3
	B1 ScentNoPerception	12.0	6.0
	B2 ScentPerception	11.7	3.3

Taking into account whether the scent was perceived (B2) or not (B1), the physiological stress indicators remain largely unchanged: B1 exhibits significantly higher PR than A (T = -2.139, p = .035), whereas B2 does not (T = -1.769, p = .080). B1 shows the highest absolute PR value, but the difference between B1 and B2 is not statistically significant.

For cognitive stress indicators, B1 scores 1.70 scale points higher than A, though the difference is not significant due to the small sample size (T = -1.538, p = .144). When the scent is consciously perceived (B2), CR is one scale point lower than A (T = -1.119, p = .277) and 2.7 scale points lower than B1, a statistically significant difference (T = 3.062, p = .011).

In light of our findings, we conclude that Hypotheses H1.1, H1.2, and H2.2 are not supported, whereas Hypothesis H2.1 can be accepted.

5 Discussion

Twenty-six participants completed a driving simulation with five critical events. The study investigated whether scent exposure reduced physiological (H1.1) and cognitive stress reactions (H1.2), and whether stress responses differed depending on whether the scent was consciously perceived (H2).

Regarding Physiological Reaction (PR), participants without scent exposure showed significantly lower values, leading to a rejection of H1.1. For Cognitive Reaction (CR), no significant differences were found based on scent exposure, thus H1.2 is not supported. However, CR was noticeably, though not significantly, lower when the scent was consciously perceived, which provides indirect support for H1.2.

H2.1 is supported, as there was no significant difference in PR between the groups with perceived and unperceived scent exposure. In contrast, H2.2 is contradicted, as participants who consciously perceived the scent showed a significantly lower CR. This suggests that conscious perception acts as a key moderating variable.

One possible explanation for these findings is that the significant reduction in cognitive stress under conscious scent perception is due to a cognitively mediated modulation of stress processing. This is comparable to the Hawthorne effect (Adair, 1984), where the awareness of an intervention influences participants' behavior. The conscious recognition of the lavender scent may have triggered a positive coping process, as participants interpreted the scent as an intentional stress-reducing measure.

This interpretation can also be linked to Expectancy Theory (Vroom, 1964), which posits that subjective expectations influence both behavior and physiological responses. If participants consciously perceive lavender—typically associated with relaxation—they may expect a calming effect, which in turn facilitates such a response. This aligns with placebo mechanisms (e.g., Benedetti, 2014), suggesting that conscious scent perception alone may be sufficient to trigger regulatory responses, regardless of any direct physiological effect.

The observed increase in stress during unconscious scent exposure might point to a mismatch between sensory stimulation and cognitive appraisal. Previous studies have shown that unexpected or subliminal olfactory stimuli can increase alertness (M. Jiang et al., 2024). Other possible explanations include scent aversion, novelty effects, or individual differences in olfactory sensitivity—variables not systematically measured in this study. As prior research (Rachel S. Herz et al., 2004) indicates that preference and familiarity with scents modulate both emotional and physiological responses, future studies should more thoroughly assess these individual characteristics.

Our findings stand in contrast to earlier research reporting a generally calming effect of lavender scent (Luca & Botelho, 2021). Potential reasons for this discrepancy may include variations in experimental design, interindividual differences in stimulus processing, or expectancy/placebo-related effects (Howard & Hughes, 2008; Masaoka et al., 2013). Additionally, differences in scent intensity and duration may have contributed to these divergent outcomes. Research suggests that higher intensities are often associated with lower pleasantness ratings (Doty et al., 1978; Henion, 1971), while continuous exposure can lead to rapid olfactory adaptation, diminishing perceptual and physiological responses over time (Mignot et al., 2022).

In the present study, lavender was administered in pulsed intervals via a professional diffuser (Reima AromaStreamer 450), allowing for moderate, sustained intensity and reduced adaptation effects (Croy et al., 2013; Nomura et al., 2016). This controlled delivery method may partly explain the differential effects compared to studies using continuous or unregulated exposure.

In summary, the data suggest that scent exposure may increase stress when the scent is not consciously perceived, possibly due to implicit arousal effects rather than a relaxation response. Cognitive stress reactions appear to be more strongly affected than physiological responses, although not all findings reached statistical significance. Notably, when the scent was consciously perceived, cognitive stress was significantly lower, underscoring the importance of perception as a moderating factor.

6 Limitations

Our findings should be interpreted as exploratory due to the limited sample size and must be validated in subsequent studies with larger and more diverse samples. This limitation may have reduced statistical power, increasing the likelihood of Type II errors, suggesting that smaller, yet potentially meaningful effects may have gone undetected.

The influence of different scents and scent intensities on physiological and cognitive stress responses warrants systematic investigation. Individual scent perception can vary considerably; incorporating a neutral control or placebo scent condition would aid in distinguishing psychological expectation effects from actual scent-related outcomes.

Given that individual physiological variability can influence stress reactivity, baseline correction should be incorporated in future experimental designs. This was not feasible in the present study due to incomplete data collection during the baseline measurement.

Our investigation was limited to short-term effects. Longitudinal research is needed to determine whether the observed outcomes persist, diminish, or intensify with repeated or prolonged scent exposure.

Subsequent studies should also consider individual olfactory characteristics, such as general olfactory sensitivity and personal scent preferences or aversions, as these factors may modulate stress responses. Additionally, examining the role of cognitive appraisal processes in a hypothesis-driven manner may help explain the observed divergence between physiological and cognitive effects of scent exposure.

Finally, the impact of different scent delivery methods and intensity levels should be systematically compared to assess their respective effects on stress responses.

7 Conclusion

This study investigated how exposure to lavender scent influences physiological and cognitive stress responses in a simulated driving environment. The results indicate that lavender scent does not inherently reduce stress. A significant reduction in cognitive stress was observed only when the scent was consciously perceived. In contrast, unconscious exposure was associated with a potential increase in stress levels, possibly due to an arousal effect. Physiological responses were less affected overall than cognitive reactions.

These findings support theoretical frameworks such as Expectancy Theory and placebo mechanisms, while contradicting earlier research that attributed a generally calming effect to lavender. Notably, conscious perception emerged as a critical moderating variable in the effectiveness of olfactory interventions.

From a practical standpoint, scent-based interventions—such as those used in vehicles or high-stress work environments—should be designed to ensure that the scent is consciously perceived, as unconscious exposure may elicit unintended stress responses. Nevertheless, physiological indicators may offer potential for adaptive systems (vom Brocke et al., 2020) that respond in real-time to individual scent preferences and perception, thereby tailoring olfactory environments more effectively.

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*(2), 334–345. https://doi.org/10.1037/0021-9010.69.2.334

Alaoui-Ismaïli, O., Robin, O., Rada, H., Dittmar, A., & Vernet-Maury, E. (1997). Basic emotions evoked by odorants: Comparison between autonomic responses and self-evaluation. *Physiology & Behavior*, *62*(4), 713–720. https://doi.org/10.1016/S0031-9384(97)90016-0

Alsuraykh, N. H., Wilson, M. L., Tennent, P., & Sharples, S. (2019). How Stress and Mental Workload are Connected. In O. Mayora, S. Forti, J. Meyer, & L. Mamykina (Eds.), *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare* (pp. 371–376). ACM. https://doi.org/10.1145/3329189.3329235

Andreassi, J. L. (2010). *Psychophysiology*. Psychology Press. https://doi.org/10.4324/9780203880340

Arza, A., Garzón-Rey, J. M., Lázaro, J., Gil, E., Lopez-Anton, R., La Camara, C. de, Laguna, P., Bailon, R., & Aguiló, J. (2019). Measuring acute stress response through physiological signals: Towards a quantitative assessment of stress. *Medical & Biological Engineering & Computing*, *57*(1), 271–287. https://doi.org/10.1007/s11517-018-1879-z

Beanland, V., Fitzharris, M., Young, K. L., & Lenné, M. G. (2013). Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study. *Accident Analysis & Prevention*, *54*, 99–107. https://doi.org/10.1016/j.aap.2012.12.043

Becker, S., Spinath, B., Ditzen, B., & Dörfler, T. (2023). Psychological Stress = Physiological Stress? *Journal of Psychophysiology*, *37*(1), 12–24. https://doi.org/10.1027/0269-8803/a000301

Benedetti, F. (2014). Placebo effects: From the neurobiological paradigm to translational implications. *Neuron*, *84*(3), 623–637. https://doi.org/10.1016/j.neuron.2014.10.023

Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M., & Holley, A. (2002). Influence of affective and cognitive judgments on autonomic parameters during inhalation of pleasant and unpleasant odors in humans. *Neuroscience Letters*, *319*(3), 162–166. https://doi.org/10.1016/S0304-3940(01)02572-1

Berntson, G. G., Norman, G. J., Hawkley, L. C., & Cacioppo, J. T. (2008). Cardiac autonomic balance versus cardiac regulatory capacity. *Psychophysiology*, *45*(4), 643–652. https://doi.org/10.1111/j.1469-8986.2008.00652.x

Boucsein, W. (2012). *Electrodermal Activity*. Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Brookhuis, K. A., & Waard, D. de (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident; Analysis and Prevention*, *42*(3), 898–903. https://doi.org/10.1016/j.aap.2009.06.001

Buchbauer, G [G.], Jirovetz, L., Jäger, W., Dietrich, H., & Plank, C. (1991). Aromatherapy: Evidence for sedative effects of the essential oil of lavender after inhalation. *Zeitschrift Fur Naturforschung. C, Journal of Biosciences*, *46*(11-12), 1067–1072. https://doi.org/10.1515/znc-1991-11-1223

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, *9*(3), 399–431. https://doi.org/10.1152/physrev.1929.9.3.399

Caruelle, D., Gustafsson, A., Shams, P., & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research*, *104*, 146–160. https://doi.org/10.1016/j.jbusres.2019.06.041

Castiello, U., Zucco, G. M., Parma, V., Ansuini, C., & Tirindelli, R. (2006). Cross-modal interactions between olfaction and vision when grasping. *Chemical Senses*, *31*(7), 665–671. https://doi.org/10.1093/chemse/bjl007

Chyung, S. Y. Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement*, *57*(5), 38–48. https://doi.org/10.1002/pfi.21763

Croy, I., Maboshe, W., & Hummel, T [T.] (2013). Habituation effects of pleasant and unpleasant odors. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *88*(1), 104–108. https://doi.org/10.1016/j.ijpsycho.2013.02.005

Czaban, M., Riedl, J., & Wengler, S. (20XXa). Physiological and Cognitive Stress Responses in Driving Simulators: Investigating the Influence of Situational and Personality Factors. *Manuscript Submitted for Publication*.

Czaban, M., Riedl, J., & Wengler, S. (20XXb). Single Measurements versus Composite Indicators for User Experience Research. *Manuscript Submitted for Publication*.

Darzi, A., Gaweesh, S. M., Ahmed, M. M., & Novak, D. (2018). Identifying the Causes of Drivers' Hazardous States Using Driver Characteristics, Vehicle Kinematics, and Physiological Measurements. *Frontiers in Neuroscience*, *12*, 568. https://doi.org/10.3389/fnins.2018.00568

Daviaux, Y., Bonhomme, E., Ivers, H., Sevin, É. de, Micoulaud-Franchi, J.-A., Bioulac, S., Morin, C. M., Philip, P., & Altena, E. (2020). Event-Related Electrodermal Response to Stress: Results From a Realistic Driving Simulator Scenario. *Human Factors*, *62*(1), 138–151. https://doi.org/10.1177/0018720819842779

Dawes, J. (2002). Five point vs. eleven point scales: Does it make a difference to data characteristics. *Australasian Journal of Market Research*, *10*(1).

Deivanayagame, B., Kumar, A. V. S., Maruthy, K., & Kareem, S. (2020). Effect of Peppermint Aroma on Short Term Memory and Cognition in Healthy Volunteers. *International Journal of Physiology*. https://www.semanticscholar.org/paper/Effect-of-Peppermint-Aroma-on-Short-Term-Memory-and-Deivanayagame-Kumar/e779a40d86e56ee68015b85d97b1f1b8c644bf57

Diego, M. A., Jones, N. A., Field, T., Hernandez-Reif, M., Schanberg, S., Kuhn, C., McAdam, V., Galamaga, R., & Galamaga, M. (1998). Aromatherapy positively affects mood, EEG patterns of alertness and math computations. *International*

Journal of Neuroscience, *96*(3-4), 217–224. https://doi.org/10.3109/00207459808986469

Dimoka, Davis, Gupta, Pavlou, Banker, Dennis, Ischebeck, Müller-Putz, Benbasat, Gefen, Kenning, Riedl, vom Brocke, & Weber (2012). On the Use of Neurophysiological Tools in IS Research: Developing a Research Agenda for NeuroIS. *MIS Quarterly*, *36*(3), 679. https://doi.org/10.2307/41703475

Dmitrenko, D., Maggioni, E., Brianza, G., Holthausen, B. E., Walker, B. N., & Obrist, M. (2020). CARoma Therapy: Pleasant Scents Promote Safer Driving, Better Mood, and Improved Well-Being in Angry Drivers. *CHI '20: CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376176

Dogan, D., Bogosyan, S., & Acarman, T. (2019). Evaluation of driver stress level with survey, galvanic skin response sensor data, and force-sensing resistor data. *Advances in Mechanical Engineering*, *11*(12), Article 1687814019891555. https://doi.org/10.1177/1687814019891555

Doty, R. L., Orndorff, M. M., Leyden, J., & Kligman, A. (1978). Communication of gender from human axillary odors: Relationship to perceived intensity and hedonicity. *Behavioral Biology*, *23*(3), 373–380. https://doi.org/10.1016/S0091-6773(78)91393-7

Fischer, T., & Riedl, R. (2017). Technostress Research: A Nurturing Ground for Measurement Pluralism? *Communications of the Association for Information Systems*, *40*, 375–401. https://doi.org/10.17705/1CAIS.04017

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*, 90–99. https://doi.org/10.1016/j.apergo.2018.06.006

Galante, F., Bracco, F., Chiorri, C., Pariota, L., Biggero, L., & Bifulco, G. N. (2018). Validity of Mental Workload Measures in a Driving Simulation Environment. *Journal of Advanced Transportation*, *2018*, 1–11. https://doi.org/10.1155/2018/5679151

Ghavami, T., Kazeminia, M., & Rajati, F. (2022). The effect of lavender on stress in individuals: A systematic review and meta-analysis. *Complementary Therapies in Medicine*, *68*, 102832. https://doi.org/10.1016/j.ctim.2022.102832

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, *13*(1), 440–460. https://doi.org/10.1109/TAFFC.2019.2927337

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology : Human Mental Workload* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey, J. A., & Picard, R. W. (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. https://doi.org/10.1109/TITS.2005.848368

Henion, K. E. (1971). Odor pleasantness and intensity: A single dimension? *Journal of Experimental Psychology*, *90*(2), 275–279. https://doi.org/10.1037/h0031549

Herz, R. S [R. S.], & Engen, T. (1996). Odor memory: Review and analysis. *Psychonomic Bulletin & Review*, *3*(3), 300–313. https://doi.org/10.3758/BF03210754

Herz, R. S [Rachel S.] (2009). Aromatherapy facts and fictions: A scientific analysis of olfactory effects on mood, physiology and behavior. *The International Journal of Neuroscience*, *119*(2), 263–290. https://doi.org/10.1080/00207450802333953

Herz, R. S [Rachel S.], Schankler, C., & Beland, S. (2004). Olfaction, Emotion and Associative Learning: Effects on Motivated Behavior. *Motivation and Emotion*, *28*(4), 363–383. https://doi.org/10.1007/s11031-004-2389-x

Heuberger, E [Eva], Redhammer, S., & Buchbauer, G [Gerhard] (2004). Transdermal absorption of (-)-linalool induces autonomic deactivation but has no impact on ratings of well-being in humans. *Neuropsychopharmacology*, *29*(10), 1925–1932. https://doi.org/10.1038/sj.npp.1300521

Hines Duncliffe, T., D'Angelo, B., Brock, M., Fraser, C., Austin, N., Lamarra, J., Pusateri, M., Livingston, L., & Batt, A. (2018). The Effects of Stress on the Driving Abilities of Paramedic Students. *EMS World*, *47*, 76.

Hoffmann, S., Kruger, H.-P., & Buld, S. (2003). Avoidance of simulator sickness by training the adaptation to the driving simulation. *VDI BERICHTE*, *1745*, 385–406.

Appendix C

Table 6. Additional papers and publications

Authors (Year)	Title	Journal/Publisher	Status
Riedl, Joachim; Wengler, Stefan; Czaban,Marcin; Steudtel, Simon (2023)	Sexism in Advertisements – A Cross-Cultural Analysis	Marketing Science & Inspirations	published
Riedl, Joachim; Wengler, Stefan; Czaban, Marcin; Mohr, Sarah Victoria; Steudtel, Simon (2024)	Studies on the Human-Machine-Interface in Advanced Driver Assistance Systems towards Autonomous Driving	University of Applied Sciences Hof	published
Wengler, Stefan; Riedl, Joachim; Bichler-Riedl, Wolfgang; Czaban, Marcin; Mohr, Sarah Victoria (2024)	Hypothetical Constructs of Consumer behavior as predictors of pro-environmental behavior – An empirical study based on smartphones	Marketing Science & Inspirations	published
Riedl, Joachim; Wengler, Stefan; Czaban, Marcin; Mohr, Sarah Victoria; Steudtel, Simon (2025)	Studies on vehicle usability	University of Applied Sciences Hof	published
Czarnecki, Christian; Sultanow, Eldar; Sebrak, Sebastian; Gronau, Norbert; Teichmann, Malte; Ritterbusch, Georg David; Mohr, Sarah Victoria; Auman, Matthias; Czaban, Marcin; Wengler, Stefan (2026)	Ideengenerierung mit KI – Anwendungsfälle als Treiber für Innovation	Springer Essential	In publishing
Czaban, Marcin; Riedl, Joachim; Wengler, Stefan (20xx)	Detecting Psychophysiological and cognitive stress in critical driving simulator scenarios	Tbd	tbd
Wengler, Stefan; Riedl, Joachim; Czaban, Marcin (20xx)	Generation and adoption of innovations: conceptual and exploratory insights in the automobile industry from a multi-stage marketing perspective	Journal of Business and Industrial Marketing	Under Review
Wengler, Stefan; Czaban, Marcin; Riedl, Joachim (20xx)	Key account management in fragmented business market value chains: conceptual and exploratory insights from a multi-stage marketing and customer-perceived value perspective	Journal of Business and Industrial Marketing	Under Review

References

Abdel-Aty, M., & Ding, S. (2024). A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nature Communications*, *15*(1), 4931. https://doi.org/10.1038/s41467-024-48526-4

Ahlström, C., Bolling, A., Sörensen, G., Eriksson, O., & Andersson, A. (2012). *Validating speed and road surface realism in VTI driving simulator III*. Statens väg-och transportforskningsinstitut.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Albers, D., Radlmayr, J., Loew, A., Hergeth, S., Naujoks, F., Keinath, A., & Bengler, K. (2020). Usability Evaluation—Advances in Experimental Design in the Context of Automated Driving Human–Machine Interfaces. *Information*, *11*(5), 240. https://doi.org/10.3390/info11050240

Alsuraykh, N. H., Wilson, M. L., Tennent, P., & Sharples, S [Sarah] (2019). How Stress and Mental Workload are Connected. In O. Mayora, S. Forti, J. Meyer, & L. Mamykina (Eds.), *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare* (pp. 371–376). ACM. https://doi.org/10.1145/3329189.3329235

Apraiz Iriarte, A., Lasa Erle, G., & Mazmela Extabe, M. (2021). Evaluating User Experience with physiological monitoring: A systematic literature review. *DYNA NEW TECHNOLOGIES*, *8*(1), [20 p.]-[20 p.]. https://doi.org/10.6036/NT10072

Arza, A., Garzón-Rey, J. M., Lázaro, J., Gil, E., Lopez-Anton, R., La Camara, C. de, Laguna, P., Bailon, R., & Aguiló, J. (2019). Measuring acute stress response through physiological signals: Towards a quantitative assessment of stress. *Medical & Biological Engineering & Computing*, *57*(1), 271–287. https://doi.org/10.1007/s11517-018-1879-z

Baier, D., Karasenko, A., & Rese, A. (2025). Measuring technology acceptance over time using transfer models based on online customer reviews. *Journal of Retailing and Consumer Services*, *85*, 104278. https://doi.org/10.1016/j.jretconser.2025.104278

Bansal, P., Kockelman, K. M., & Singh, A. (2016). Assessing public opinions of and interest in new vehicle technologies: An Austin perspective. *Transportation Research Part C: Emerging Technologies*, *67*, 1–14. https://doi.org/10.1016/j.trc.2016.01.019

Barker, L., Polson, J., & Dupont, P. (1978). *Driver Screening Simulator Evaluation Program.*

Barré, R., Brunel, G., Barthet, P., & Laurencin-Dalicieux, S. (2017). The visual analogue scale: An easy and reliable way of assessing perceived stress. *Quality in Primary Health Care*, *1*(1), 1–5.

Baum, A. (1990). Stress, intrusive imagery, and chronic distress. *Health Psychology*, *9*(6), 653.

Becker, S., Spinath, B., Ditzen, B., & Dörfler, T. (2023). Psychological Stress = Physiological Stress? *Journal of Psychophysiology*, *37*(1), 12–24. https://doi.org/10.1027/0269-8803/a000301

Bella, F. (2008). Driving simulator for speed research on two-lane rural roads. *Accident; Analysis and Prevention*, *40*(3), 1078–1087. https://doi.org/10.1016/j.aap.2007.10.015

Bella, F., Calvi, A., & D'Amico, F. (2014). Analysis of driver speeds under night driving conditions using a driving simulator. *Journal of Safety Research*, *49*, 45–52. https://doi.org/10.1016/j.jsr.2014.02.007

Blaauw, G. J. (1982). Driving Experience and Task Demands in Simulator and Instrumented Car: A Validation Study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(4), 473–486. https://doi.org/10.1177/001872088202400408

Blana, E. (1996). Driving simulator validation studies: a literature review.

Blana, E. (2000). *The behavioural validation of driving simulators as research tools: a case study based on the Leeds Driving Simulator* [, University of Leeds]. BibTeX.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of Training: A Meta-Analytic Review. *Journal of Management*, *36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Blut, M., Chong, A. Y. L., Tsiga, Z., & Venkatesh, V. (2022). Meta-analysis of the unified theory of acceptance and use of technology (UTAUT): challenging its validity and charting a research agenda in the red ocean. In Symposium conducted at the meeting of Association for Information Systems.

Boer, E. R. (2000). Experiencing the same road twice: A driver centered comparison between simulation and reality. In *Proceeding of Driving Simulation Conference DSC 2000*.

Boucsein, W. (2012). *Electrodermal Activity*. Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Branzi, V., Domenichini, L., & La Torre, F. (2017). Drivers' speed behaviour in real and simulated urban roads – A validation study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *49*, 1–17. https://doi.org/10.1016/j.trf.2017.06.001

Brookhuis, K. A., & Waard, D. de (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident; Analysis and Prevention*, *42*(3), 898–903. https://doi.org/10.1016/j.aap.2009.06.001

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2017). *Handbook of Psychophysiology*. Cambridge University Press. https://doi.org/10.1017/9781107415782

Cai, L., Yuen, K. F., & Wang, X. (2023). Explore public acceptance of autonomous buses: An integrated model of UTAUT, TTF and trust. *Travel Behaviour and Society*, *31*, 120–130. https://doi.org/10.1016/j.tbs.2022.11.010

Caird, J. K., & Horrey, W. J. (2016). A review of novice and teen driver distraction. *Handbook of Teen and Novice Drivers*, 189–210.

Cannon, W. B. (1939). The wisdom of the body.

Cannon, W. B., & Rosenberg, C. E. (1932). Homeostasis. *The Wisdom of the Body*, 263–286.

Carroll, M., Rebensky, S., Chaparro Osman, M., & Deaton, J. (2023). Justification for Use of Simulation. In D. A. Vincenzi, M. Moloua, P. A. Hancock, J. A. Pharmer, & J. C. Ferraro (Eds.), *Human Factors in Simulation and Training* (pp. 65–90). CRC Press. https://doi.org/10.1201/9781003401360-2

Carter, C. J., & Laya, O. (1998). Driver's visual search in a field situation and in a driving simulator. *Vision in Vehicles*, *6*, 21–31.

Caruelle, D., Gustafsson, A., Shams, P., & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research*, *104*, 146–160. https://doi.org/10.1016/j.jbusres.2019.06.041

Chen, C.-F. (2019). Factors affecting the decision to use autonomous shuttle services: Evidence from a scooter-dominant urban context. *Transportation Research Part F: Traffic Psychology and Behaviour*, *67*, 195–204. https://doi.org/10.1016/j.trf.2019.10.016

Chen, L., Zhao, Y., Ye, P., Zhang, J., & Zou, J. (2017). Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, *85*, 279–291. https://doi.org/10.1016/j.eswa.2017.01.040

Choi, J. K., & Ji, Y. G. (2015). Investigating the Importance of Trust on Adopting an Autonomous Vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692–702. https://doi.org/10.1080/10447318.2015.1070549

Chrousos, G. P. (1992). The Concepts of Stress and Stress System Disorders. *JAMA*, *267*(9), 1244. https://doi.org/10.1001/jama.1992.03480090092034

Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature Reviews. Endocrinology*, *5*(7), 374–381. https://doi.org/10.1038/nrendo.2009.106

Cohen, S., Gianaros, P. J., & Manuck, S. B. (2016). A Stage Model of Stress and Disease. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(4), 456–463. https://doi.org/10.1177/1745691616646305

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, *24*(4), 385. https://doi.org/10.2307/2136404

Crosswell, A. D., & Lockwood, K. G. (2020). Best practices for stress measurement: How to measure psychological stress in health research. *Health Psychology Open*, *7*(2), 2055102920933072. https://doi.org/10.1177/2055102920933072

Czaban, M., & Himmels, C. (2025). Investigating simulator validity by using physiological and cognitive stress indicators. *Transportation Research Part F: Traffic Psychology and Behaviour*, *114*, 831–851. https://doi.org/10.1016/j.trf.2025.07.006

Daviaux, Y., Bonhomme, E., Ivers, H., Sevin, É. de, Micoulaud-Franchi, J.-A., Bioulac, S., Morin, C. M., Philip, P., & Altena, E. (2020). Event-Related Electrodermal Response to Stress: Results From a Realistic Driving Simulator Scenario. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*(1), 138–151. https://doi.org/10.1177/0018720819842779

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* [, Massachusetts Institute of Technology]. BibTeX.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Davis, F. D., & Granić, A. (2024). *The Technology Acceptance Model*. Springer International Publishing. https://doi.org/10.1007/978-3-030-45274-2

Dehghani, A., Salaar, H., Srinivasan, S. P., Zhou, L., Arbeiter, G., Lindner, A., & Patino-Studencki, L. (2025). Enhancing Availability of Autonomous Shuttle Services: A Conceptual Approach toward Challenges and Opportunities. *SAE International Journal of Connected and Automated Vehicles*, *8*(3). https://doi.org/10.4271/12-08-03-0023

Diels, C., Reed, N., & Robbins, R. (2011). Behavioural Validation of the TRL Driving Simulator DigiCar: Phase 1-Speed Choice.

Dimoka, Davis, Gupta, Pavlou, Banker, Dennis, Ischebeck, Müller-Putz, Benbasat, Gefen, Kenning, Riedl, vom Brocke, & Weber (2012). On the Use of Neurophysiological Tools in IS Research: Developing a Research Agenda for NeuroIS. *MIS Quarterly*, *36*(3), 679. https://doi.org/10.2307/41703475

Donkor, R. A., Burnett, G. A., & Sharples, S [S.] (2014). Measuring the emotional validity of driving simulators. *Advances in Transportation Studies*.

Drosdol, J., & Panik, F. (1985). The Daimler-Benz driving simulator a tool for vehicle development. *SAE Transactions*, 981–997.

Dużmańska, N., Strojny, P., & Strojny, A. (2018). Can Simulator Sickness Be Avoided? A Review on Temporal Aspects of Simulator Sickness. *Frontiers in Psychology*, *9*, 2132. https://doi.org/10.3389/fpsyg.2018.02132

Eden, G., Nanchen, B., Ramseyer, R., & Evéquoz, F. (2017). On the Road with an Autonomous Passenger Shuttle. In G. Mark, S. Fussell, C. Lampe, m. schraefel, J. P. Hourcade, C. Appert, & D. Wigdor (Eds.), *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1569–1576). ACM. https://doi.org/10.1145/3027063.3053126

Engen, T. (2008). Use and validation of driving simulators.

Engert, V., Merla, A., Grant, J. A., Cardone, D., Tusche, A., & Singer, T. (2014). Exploring the use of thermal infrared imaging in human stress research. *PloS One*, *9*(3), e90782. https://doi.org/10.1371/journal.pone.0090782

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: a perceived risk facets perspective. *International Journal of Human-Computer Studies*, *59*(4), 451–474. https://doi.org/10.1016/S1071-5819(03)00111-3

Feuerstein, M., Labbé, E. E., & Kuczmierczyk, A. R. (2013). *Health psychology: A psychobiological perspective*. Springer Science & Business Media.

Fischer, M., Labusch, A., Bellmann, T., & Seehof, C. (2015). A task-oriented catalogue of criteria for driving simulator evaluation. In *Proceedings of the Driving Simulation Conference 2015*.

Fishbein, M., Ajzen, I., & Belief, A. (1975). *Intention and Behavior: An introduction to theory and research*. Addison-Wesley, Reading, MA.

Fors, C., Ahlström, C., & Anund, A. (2013). *Simulator validation with respect to driver sleepiness and subjective experiences : final report of the project SleepEYE II, part 1* (ViP publication : ViP - Virtual Prototyping and Assessment by Simulation 2013-1). Swedish National Road and Transport Research Institute, Human-vehicle-transport system interaction.

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*, 90–99. https://doi.org/10.1016/j.apergo.2018.06.006

Galante, F., Bracco, F., Chiorri, C., Pariota, L., Biggero, L., & Bifulco, G. N. (2018). Validity of Mental Workload Measures in a Driving Simulation Environment. *Journal of Advanced Transportation*, *2018*, 1–11. https://doi.org/10.1155/2018/5679151

Gefen, Karahanna, & Straub (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, *27*(1), 51. https://doi.org/10.2307/30036519

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, *13*(1), 440–460. https://doi.org/10.1109/TAFFC.2019.2927337

Godley, S. T., Triggs, T. J., & Fildes, B. N. (2002). Driving simulator validation for speed research. *Accident; Analysis and Prevention*, *34*(5), 589–600. https://doi.org/10.1016/S0001-4575(01)00056-2

Golbabaei, F., Yigitcanlar, T., Paz, A., & Bunker, J. (2022). Understanding Autonomous Shuttle Adoption Intention: Predictive Power of Pre-Trial Perceptions and Attitudes. *Sensors (Basel, Switzerland)*, *22*(23). https://doi.org/10.3390/s22239193

Gummesson, E. (2008). Extending the service-dominant logic: from customer centricity to balanced centricity. *Journal of the Academy of Marketing Science*, *36*(1), 15–17. https://doi.org/10.1007/s11747-007-0065-x

Hall, J. E., & Hall, M. E. (2020). *Guyton and Hall Textbook of Medical Physiology E-Book: Guyton and Hall Textbook of Medical Physiology E-Book*. Elsevier Health Sciences.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology. Human Mental Workload* (Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey, J. A., & Picard, R. W. (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. https://doi.org/10.1109/TITS.2005.848368

Hebb, D. O. (1955). Drives and the C.N.S. (conceptual nervous system). *Psychological Review*, *62*(4), 243–254. https://doi.org/10.1037/h0041823

Helton, W. S. (2004). *PsycTESTS Dataset.* https://doi.org/10.1037/t57758-000

Herrenkind, B., Brendel, A. B., Nastjuk, I., Greve, M., & Kolbe, L. M. (2019). Investigating end-user acceptance of autonomous electric buses to accelerate diffusion. *Transportation Research Part D: Transport and Environment*, *74*, 255–276. https://doi.org/10.1016/j.trd.2019.08.003

Himmels, C. (2025). A Use-Case Driven Approach to Establishing Driving Simulator Validity/Author Chantal Himmels, M. Sc.

Himmels, C., Venrooij, J., Parduzi, A., Peller, M., & Riener, A. (2024). The bigger the better? Investigating the effects of driving simulator fidelity on driving behavior and perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *101*, 250–266. https://doi.org/10.1016/j.trf.2024.01.007

Hussain, Q., Alhajyaseen, W. K., Pirdavani, A., Reinolsmann, N., Brijs, K., & Brijs, T. (2019). Speed perception and actual speed in a driving simulator and real-world: A validation study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *62*, 637–650. https://doi.org/10.1016/j.trf.2019.02.019

Kabilmiharbi, N., Kamaliana Khamis, N., & Azila Noh, N. (2022). Commonly Used Assessment Method to Evaluate Mental Workload for Multiple Driving Distractions: A Systematic Review. *Iranian Journal of Public Health*, *51*(3), 482–494. https://doi.org/10.18502/ijph.v51i3.8924

Kapser, S., & Abdelrahman, M. (2020). Acceptance of autonomous delivery vehicles for last-mile delivery in Germany – Extending UTAUT2 with risk perceptions. *Transportation Research Part C: Emerging Technologies*, *111*, 210–225. https://doi.org/10.1016/j.trc.2019.12.016

Kaptein, N., Theeuwes, J., & van der Horst, R. (1996). Driving Simulator Validity: Some Considerations. *Transportation Research Record: Journal of the Transportation Research Board*, *1550*, 30–36. https://doi.org/10.3141/1550-05

Kelly, D., Kantor, P. B., Morse, E. L., Scholtz, J., & Sun, Y. (2009). Questionnaires for eliciting evaluation data from users of interactive question answering systems. *Natural Language Engineering*, *15*(1), 119–141. https://doi.org/10.1017/S1351324908004932

Kleinaltenkamp, M., Eggert, A., Kashyap, V., & Ulaga, W. (2022). Rethinking customer-perceived value in business markets from an organizational perspective. *Journal of Inter-Organizational Relationships*, *28*(1-2), 1–18. https://doi.org/10.1080/26943980.2022.2129545

Klüver, M. (2016). *Can we trust driving simulator studies: The behavioral validity of the Daimler AG driving simulators* [, Johannes-Gutenberg-Universität Mainz]. BibTeX.

Korkmaz, H., Fidanoglu, A., Ozcelik, S., & Okumus, A. (2022). User Acceptance of Autonomous Public Transport Systems (APTS): Extended UTAUT2 Model. *Journal of Public Transportation*, *24*(100013). https://doi.org/10.5038/2375-0901.23.1.5

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*(3), 261–273. https://doi.org/10.1111/j.1469-8986.1993.tb03352.x

Lauer, A. R. (1960). The psychology of driving: Factors of traffic enforcement.

Lazarus, R. S. (1966). Psychological stress and the coping process.

Lazarus, R. S. (1990). Theory-Based Stress Measurement. *Psychological Inquiry*, *1*(1), 3–13. https://doi.org/10.1207/s15327965pli0101_1

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Leis, O., & Lautenbach, F. (2020). Psychological and physiological stress in non-competitive and competitive esports settings: A systematic review. *Psychology of Sport and Exercise*, *51*, 101738. https://doi.org/10.1016/j.psychsport.2020.101738

Li, J., Zhao, X., Xu, S., Ma, J., & Rong, J. (2013). The Study of Driving Simulator Validation for Physiological Signal Measures. *Procedia - Social and Behavioral Sciences*, *96*, 2572–2583. https://doi.org/10.1016/j.sbspro.2013.08.288

Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? In *Proceedings of the 17th Australia conference on computer-human interaction: Citizens online: Considerations for today and the future*.

Liu, D., Yu, J., Macchiarella, N. D., & Vincenzi, D. A. (2023). Simulation Fidelity. In D. A. Vincenzi, M. Moloua, P. A. Hancock, J. A. Pharmer, & J. C. Ferraro (Eds.), *Human Factors in Simulation and Training* (pp. 91–108). CRC Press. https://doi.org/10.1201/9781003401360-3

Lobjois, R., Faure, V., Désiré, L., & Benguigui, N. (2021). Behavioral and workload measures in real and simulated driving: Do they tell us the same thing about the validity of driving simulation? *Safety Science*, *134*, 105046. https://doi.org/10.1016/j.ssci.2020.105046

Lohani, M., Payne, B. R., & Strayer, D. L. (2019). A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. *Frontiers in Human Neuroscience*, *13*, 57. https://doi.org/10.3389/fnhum.2019.00057

Madigan, R., Louw, T., Dziennus, M., Graindorge, T., Ortega, E., Graindorge, M., & Merat, N. (2016). Acceptance of Automated Road Transport Systems (ARTS): An Adaptation of the UTAUT Model. *Transportation Research Procedia*, *14*, 2217–2226. https://doi.org/10.1016/j.trpro.2016.05.237

Madigan, R., Louw, T., Wilbrink, M., Schieben, A., & Merat, N. (2017). What influences the decision to use automated public transport? Using UTAUT to understand public acceptance of automated road transport systems. *Transportation Research Part F: Traffic Psychology and Behaviour*, *50*, 55–64. https://doi.org/10.1016/j.trf.2017.07.007

Mahmud, S., Shen, H., Foutz, Y. N. Z., & Anton, J. (2022). Reinforcement Learning Based Route And Stop Planning For Autonomous Vehicle Shuttle Service. In *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)* (pp. 668–674). IEEE. https://doi.org/10.1109/MASS56207.2022.00098

Mauri, M., Magagnin, V., Cipresso, P., Mainardi, L., Brown, E. N., Cerutti, S., Villamira, M., & Barbieri, R. (2010). Psychophysiological signals associated with affective states. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2010*, 3563–3566. https://doi.org/10.1109/IEMBS.2010.5627465

Menon, N. (2017). *Autonomous Vehicles: an Empirical Assessment of Consumers' Perceptions, Intended Adoption, and Impacts on Household Vehicle Ownership*. University of South Florida.

Milleville-Pennel, I., & Charron, C. (2015). Driving for Real or on a Fixed-Base Simulator: Is It so Different? An Explorative Study. *Presence: Teleoperators and Virtual Environments*, *24*(1), 74–91. https://doi.org/10.1162/PRES_a_00216

Mira Bonnardel, S., Antonialli, F., & Attias, D. (2020). Autonomous Vehicles toward a Revolution in Collective Transport. In S. Ersoy & T. Waqar (Eds.), *Autonomous Vehicle and Smart Traffic.* IntechOpen. https://doi.org/10.5772/intechopen.89941

Mohajer, N., Abdi, H., Nelson, K., & Nahavandi, S. (2015). Vehicle motion simulators, a key step towards road vehicle dynamics improvement. *Vehicle System Dynamics*, *53*(8), 1204–1226. https://doi.org/10.1080/00423114.2015.1039551

Mueller, J. A. (2015). *Driving in a simulator versus on-road: The effect of increased mental effort while driving on real roads and a driving simulator*. Montana State University.

Mullen, N., Charlton, J., Devlin, A., & Bedard, M. (2011). Simulator validity: Behaviours observed on the simulator and on the road. In *Handbook of driving simulation for engineering, medicine and psychology* (pp. 1–18). CRC Press.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Nordhoff, S., Malmsten, V., van Arem, B., Liu, P., & Happee, R [Riender] (2021). A structural equation modeling approach for the acceptance of driverless automated shuttles based on constructs from the Unified Theory of Acceptance and Use of Technology and the Diffusion of Innovation Theory. *Transportation Research Part F: Traffic Psychology and Behaviour*, *78*, 58–73. https://doi.org/10.1016/j.trf.2021.01.001

Nordhoff, S., van Arem, B., Merat, N., Madigan, R., Ruhrort, L., Knie, A., & Happee, R [Riender] (2017). User acceptance of driverless shuttles running in an open and mixed traffic environment. In *12th ITS European Congress.* Symposium conducted at the meeting of Strasbourg, France.

Nordhoff, S., Winter, J. de [Joost], Madigan, R., Merat, N., van Arem, B., & Happee, R [Riender] (2018). User acceptance of automated shuttles in Berlin-Schöneberg: A questionnaire study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 843–854. https://doi.org/10.1016/j.trf.2018.06.024

Nurdianto, Singgih, M. L., & Gunarta, I. K. (2024). Building Composite Performance Index for Broadband Internet Customer Experience. In *2024 IEEE International Symposium on Consumer Technology (ISCT)* (pp. 669–675). IEEE. https://doi.org/10.1109/ISCT62336.2024.10791161

Othman, K. (2023). Public attitude towards autonomous vehicles before and after crashes: A detailed analysis based on the demographic characteristics. *Cogent Engineering*, *10*(1), Article 2156063. https://doi.org/10.1080/23311916.2022.2156063

Parduzi, A. (2021). *Bewertung der Validität von Fahrsimulatoren anhand vibro-akustischer Fahrzeugschwingungen* [, Dissertation, Berlin, Technische Universität Berlin, 2021]. BibTeX.

Pawar, N. M., & Velaga, N. R. (2020). Modelling the influence of time pressure on reaction time of drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *72*, 1–22. https://doi.org/10.1016/j.trf.2020.04.017

Pawar, N. M., Velaga, N. R., & Sharmila, R. B. (2022). Exploring behavioral validity of driving simulator under time pressure driving conditions of professional drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *89*, 29–52. https://doi.org/10.1016/j.trf.2022.06.004

Pinel, J. P. J., & Barnes, S. J. (2021). *Biopsychology*. Pearson Higher Ed.

Qu, W., Zhang, Q., Zhao, W., Zhang, K., & Ge, Y. (2016). Validation of the Driver Stress Inventory in China: Relationship with dangerous driving behaviors. *Accident; Analysis and Prevention*, *87*, 50–58. https://doi.org/10.1016/j.aap.2015.11.019

Reimer, B., D'Ambrosio, L. A., Coughlin, J. E., Kafrissen, M. E., & Biederman, J. (2006). Using self-reported data to assess the validity of driving simulation data. *Behavior Research Methods*, *38*(2), 314–324. https://doi.org/10.3758/BF03192783

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: A field study and simulation validation. *Ergonomics*, *54*(10), 932–942. https://doi.org/10.1080/00140139.2011.604431

Rejali, S., Aghabayk, K., Mohammadi, A., & Shiwakoti, N. (2024). Evaluating public a priori acceptance of autonomous modular transit using an extended unified theory of acceptance and use of technology model. *Journal of Public Transportation*, *26*, 100081. https://doi.org/10.1016/j.jpubtr.2024.100081

Ren, P., Barreto, A., Gao, Y., & Adjouadi, M. (2013). Affective Assessment by Digital Processing of the Pupil Diameter. *IEEE Transactions on Affective Computing*, *4*(1), 2–14. https://doi.org/10.1109/T-AFFC.2012.25

Rendon-Velez, E., van Leeuwen, P. M. ., Happee, R [R.], Horváth, I., van der Vegte, W. F., & Winter, J. de [J.C.F.] (2016). The effects of time pressure on driver performance and physiological activity: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *41*, 150–169. https://doi.org/10.1016/j.trf.2016.06.013

Rese, A., Baier, D., Geyer-Schulz, A., & Schreiber, S. (2017). How augmented reality apps are accepted by consumers: A comparative analysis using scales and opinions. *Technological Forecasting and Social Change*, *124*, 306–319. https://doi.org/10.1016/j.techfore.2016.10.010

Rese, A., Schreiber, S., & Baier, D. (2014). Technology acceptance modeling of augmented reality at the point of sale: Can surveys be replaced by an analysis of online reviews? *Journal of Retailing and Consumer Services*, *21*(5), 869–876. https://doi.org/10.1016/j.jretconser.2014.02.011

Riedl, J., Wengler, S., Czaban, M., Mohr, S. V., & Steudtel, S. (2024). *Studies on the Human-Machine-Interface in Advanced Driver Assistance Systems towards Autonomous Driving*. Hochschule Hof. https://doi.org/10.57944/1051-147

Ringgold, V., Shields, G. S., Hauck, F., Kurz, M., Schindler-Gmelch, L., Abel, L., Richer, R., Eskofier, B. M., & Rohleder, N. (2024). The Short Stress State Questionnaire in German (SSSQ-G). *European Journal of Health Psychology*, *31*(4), 189–200. https://doi.org/10.1027/2512-8442/a000160

Robinson, A. M. (2018). Let's Talk about Stress: History of Stress Research. *Review of General Psychology*, *22*(3), 334–342. https://doi.org/10.1037/gpr0000137

Roos, A.-L., Goetz, T., Voracek, M., Krannich, M., Bieg, M., Jarrell, A., & Pekrun, R. (2021). Test Anxiety and Physiological Arousal: A Systematic Review and Meta-Analysis. *Educational Psychology Review*, *33*(2), 579–618. https://doi.org/10.1007/s10648-020-09543-z

Rowden, P., Matthews, G., Watson, B., & Biggs, H. (2011). The relative impact of work-related stress, life stress and driving environment stress on driving outcomes. *Accident; Analysis and Prevention*, *43*(4), 1332–1340. https://doi.org/10.1016/j.aap.2011.02.004

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods.

Applied Psychology, *53*(1), 61–86. https://doi.org/10.1111/j.1464-0597.2004.00161.x

Salonen, A., & Haavisto, N. (2019). Towards Autonomous Transportation. Passengers' Experiences, Perceptions and Feelings in a Driverless Shuttle Bus in Finland. *Sustainability*, *11*(3), 588. https://doi.org/10.3390/su11030588

Sapolsky, R. M. (2004). *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping*. Holt paperbacks.

Schreiber, S. (2020). Augmented-Reality-Anwendungen im Handel. In S. Schreiber (Ed.), *Forschungsgruppe Konsum und Verhalten. Die Akzeptanz von Augmented-Reality-Anwendungen im Handel* (pp. 11–57). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-29163-1_2

Selye, H. (1950). Stress and the general adaptation syndrome. *British Medical Journal*, *1*(4667), 1383–1392. https://doi.org/10.1136/bmj.1.4667.1383

Selye, H. (1976). Forty years of stress research: principal remaining problems and misconceptions. *Canadian Medical Association Journal*, *115*(1), 53.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *14*(2), 410–417. https://doi.org/10.1109/TITB.2009.2036164

Sheperdson, D. (2025). *U.S. traffic deaths fell 3.8% in 2024, lowest number since 2020*. https://www.reuters.com/world/us/us-traffic-deaths-fell-38-2024-lowest-number-since-2020-2025-04-08/

Singh, S. (2018). Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. *Traffic Safety Facts Crash Stats. Report No. DOT HS 812 506*.

Statistisches Bundesamt. (2025). *Anzahl der Getöteten bei Straßenverkehrsunfällen in Deutschland in den Jahren 1950 bis 2024*. https://de.statista.com/statistik/daten/studie/185/umfrage/todesfaelle-im-strassenverkehr/

Stoma, M., Dudziak, A., Caban, J., & Droździel, P. (2021). The Future of Autonomous Vehicles in the Opinion of Automotive Market Users. *Energies*, *14*(16), 4777. https://doi.org/10.3390/en14164777

Tams, S., Hill, K., Guinea, A., Thatcher, J., & Grover, V. (2014). NeuroIS—Alternative or Complement to Existing Methods? Illustrating the Holistic Effects of Neuroscience and Self-Reported Data in the Context of Technostress Research. *Journal of the Association for Information Systems*, *15*(10), 723–753. https://doi.org/10.17705/1jais.00374

Terumitsu, H., Tetsuo, Y., & Tsuyoshi, T. (2007). Development of the driving simulation system MOVIC-T4 and its validation using field driving data. *Tsinghua Science and Technology*, *12*(2), 141–150. https://doi.org/10.1016/S1007-0214(07)70021-4

Törnros, J. (1998). Driving behavior in a real and a simulated road tunnel--a validation study. *Accident; Analysis and Prevention*, *30*(4), 497–503. https://doi.org/10.1016/S0001-4575(97)00099-7

Venkatesh, Morris, & Davis (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425. https://doi.org/10.2307/30036540

Venkatesh, Thong, & Xu (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157. https://doi.org/10.2307/41410412

Vienne, F., Caro, S., Désiré, L., Auberlet, J.-M., Rosey, F., & Dumont, E. (2014). Driving simulator: an innovative tool to test new road infrastructures. In *TRA-Transport Research Arena*.

Wang, N., Pei, Y., & Wang, Y.-J. (2022). Antecedents in Determining Users' Acceptance of Electric Shuttle Bus Services. *Mathematics*, *10*(16), 2896. https://doi.org/10.3390/math10162896

Wang, Y., Mehler, B., Reimer, B., Lammers, V., D'Ambrosio, L. A., & Coughlin, J. F. (2010). The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing. *Ergonomics*, *53*(3), 404–420. https://doi.org/10.1080/00140130903464358

Winter, J. de [Joost], van Leeuwen, P. M., Happee, R [Riender], & others (2012). Advantages and disadvantages of driving simulators: A discussion. In *Proceedings of measuring behavior.* Symposium conducted at the meeting of Utrecht: sn.

Witte, M. de, Kooijmans, R., Hermanns, M., van Hooren, S., Biesmans, K., Hermsen, M., Stams, G. J., & Moonen, X. (2021). Self-Report Stress Measures to Assess Stress in Adults With Mild Intellectual Disabilities-A Scoping Review. *Frontiers in Psychology*, *12*, 742566. https://doi.org/10.3389/fpsyg.2021.742566

Wynne, R. A., Beanland, V., & Salmon, P. M. (2019). Systematic review of driving simulator validation studies. *Safety Science*, *117*, 138–151. https://doi.org/10.1016/j.ssci.2019.04.004

Xue, H., Previati, G., Gobbi, M., & Mastinu, G. (2023). Research and Development on Noise, Vibration, and Harshness of Road Vehicles Using Driving Simulators—A Review. *SAE International Journal of Vehicle Dynamics, Stability, and NVH*, *7*(4). https://doi.org/10.4271/10-07-04-0035

Yamaguchi, M., & Sakakima, J. (2007). Evaluation of driver stress in a motor-vehicle driving simulator using a biochemical marker. *The Journal of International Medical Research*, *35*(1), 91–100. https://doi.org/10.1177/147323000703500109

Zhong, S., Fu, X., Lu, W., Tang, F., & Lu, Y. (2022). An Expressway Driving Stress Prediction Model Based on Vehicle, Road and Environment Features. *IEEE Access*, *10*, 57212–57226. https://doi.org/10.1109/ACCESS.2022.3165570

Zöller, I. M. (2015). Analyse des Einflusses ausgewählter Gestaltungsparameter einer Fahrsimulation auf die Fahrerverhaltensvalidität.