Effect of motion mismatches on ratings of motion incongruence and simulator sickness in urban driving simulations

Kolff, Maurice; Himmels, Chantal; Venrooij, Joost; Parduzi, Arben; Pool, Daan M.; Riener, Andreas; Mulder, Max

DOI
10.1016/j.trf.2025.103370

Publication date
2025

Document Version
Final published version

Published in
Transportation Research Part F: Traffic Psychology and Behaviour

Citation (APA)
Kolff, M., Himmels, C., Venrooij, J., Parduzi, A., Pool, D. M., Riener, A., & Mulder, M. (2025). Effect of motion mismatches on ratings of motion incongruence and simulator sickness in urban driving simulations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *115*, Article 103370. https://doi.org/10.1016/j.trf.2025.103370

Transportation Research Part F: Psychology and Behaviour 115 (2025) 103370

Contents lists available at ScienceDirect

Transportation Research Part F: Psychology and Behaviour

journal homepage: www.elsevier.com/locate/trf

Effect of motion mismatches on ratings of motion incongruence and simulator sickness in urban driving simulations

Maurice Kolff [a,*], Chantal Himmels [b,c,d], Joost Venrooij [b], Arben Parduzi [b], Daan M. Pool [a], Andreas Riener [c], Max Mulder [a]

[a] Control & Simulation, Faculty of Aerospace Engineering, Delft University of Technology, HS, Delft, 2629, The Netherlands
[b] BMW Group Research & Development, Knorrstraße 147, Munich, 80788, Germany
[c] Human-Computer Interaction Group, Technische Hochschule Ingolstadt (THI), Esplanade 10, Ingolstadt, 85049, Germany
[d] Johannes Kepler University Linz, Altenberger Str. 69, Linz, 4040, Austria

ARTICLE INFO

Keywords:
Simulator sickness
Driving simulation
Motion incongruences
Urban driving
Continuous ratings

ABSTRACT

This paper investigates the effects of motion mismatches on simulator sickness and subjective ratings of the motion. In an open-loop driving simulator experiment, participants were driven through a recorded urban drive twelve times, in which mismatches were induced by manipulating the following three aspects in motion cueing: (i) mismatches in specific vehicle axes, (ii) mismatch types (scaling, missing, and false cues), and (iii) inconsistent scaling between different motion axes. Subjects ($N = 52$) reported simulator sickness post-hoc (after each drive), as well as continuously during each drive, a first in simulator sickness research. Furthermore, subjective post-hoc motion incongruence ratings on the quality of the motion were extracted. Results show that longitudinal motion mismatches lead to the most simulator sickness and the highest ratings, followed by mismatches in lateral motion, then yaw rate. False cues induce the most sickness, followed by missing and then scaled motion. Inconsistent scaling between the axes has no significant effect. The continuous sickness ratings support that the occurrence and severity of simulator sickness are indeed related to mismatches in simulator motion of specific maneuvers. This paper contributes to an improved understanding of the relationship between simulator motion and sickness, allowing for more targeted motion cueing strategies to prevent and reduce sickness in driving simulators. These strategies may include the appropriate selection of the simulator, the motion cueing, and the sample of participants, following the presented results.

1. Introduction

SMULATOR SICKNESS poses a crucial issue in driving simulator when conducting user studies (Caird and Horrey, 2011; De Winter et al., 2012). Although often considered a form of motion sickness (Johnson, 2007), simulator sickness can also occur without actual physical movement, e.g., in static driving simulation. Symptoms of simulator sickness include dizziness, headache, sweating, stomach awareness, and even nausea (Reason and Brand, 1975), making it an unpleasant experience. Apart from leading to a decreased sense of presence (Almallah et al., 2021), simulator sickness can cause subjects dropping out of experiments, resulting in incomplete data sets, requiring larger samples and increasing the costs of studies. It also negatively affects the validity of simulator studies, as subjects may behave systematically different when experiencing symptoms of simulator sickness (Cobb et al., 1999; Igoshina et al., 2022). For these reasons, simulator sickness is generally to be avoided and driving simulations that are expected to induce a high degree

* Corresponding author.
E-mail address: m.j.c.kolff@tudelft.nl (M. Kolff).

https://doi.org/10.1016/j.trf.2025.103370
Received 3 October 2024; Received in revised form 9 September 2025; Accepted 9 September 2025
Available online 18 September 2025

of sickness are ideally identified beforehand. These predictions would enable the testing of Motion Cueing Algorithms (MCAs) and simulators *without* the need for test drivers, saving time and money. For this, predictive models, or at least descriptive knowledge, of the quality of the motion and the induced level of simulator sickness would be of great use.

The occurrence of simulator sickness in driving simulators has been shown to depend on a variety of experiment variables, such as the participant age (Keshavarz et al., 2018), as well as the driving scenario. For example, both Klüver et al. (2015) and Mourant and Thattacherry (2000) found simulator sickness to be more severe when simulating driving on rural roads compared to highway driving. Similar results were obtained by Himmels et al. (2022), who additionally found that urban scenarios induce even more sickness than rural scenarios. Real-life urban driving itself may already cause a high degree of *motion sickness* (Irmak et al., 2021; Salter et al., 2019; Turner and Griffin, 2000). An additional issue for urban *simulations* is that the dynamic maneuvers involved, such as sharp corners and frequent acceleration/deceleration, come with strong visual stimuli, while the corresponding physical motion is difficult to reproduce on the motion platform. This can induce mismatches between expected and perceived motion. If a driver notices these mismatches, the motion is defined as *incongruent* (Cleij et al., 2018). Subjective ratings on motion incongruences are valuable information to obtain, as they can be used to identify the most critical mismatches in the simulator motion and through mitigating these mismatches the simulation realism can be increased. Although it does not necessarily occur to all drivers, one of the main theories in motion sickness research is that simulator sickness is the result of sensory conflicts (Bos, 2011; Reason, 1978), and can arise in the presence of incongruent motion. If so, the question arising is then *which* mismatches induce incongruences and simulator sickness the most, such that these can be systematically avoided. With knowledge on both phenomena, studies inducing motion incongruences and/or a high degree of simulator sickness could then be identified before-hand, such that the most appropriate simulator, MCA, and subject sample size can be selected.

In literature, apart from investigating the effect of scenario choice, most investigations focus on the benefits of particular motion (sub)systems, as these designs often have the specific potential to reduce mismatches in a certain simulator Degree of Freedom (DoF). For small, 3-DoF systems, Parduzi (2021) found no difference in simulator sickness compared to a static simulator. Zöller (2015) even found an *increase* in sickness in a 3-DoF driving simulator, which was attributed to the limited motion space. More recent work showed beneficial effects of using larger hexapod motion system with Y-drives (although small) (Klüver et al., 2015) or even XY-drives (Himmels et al., 2022). Although these studies provide insight in general qualities of various motion systems, the surplus in available motion space does not inherently lead to smaller mismatches. For example, simulators combined with redundant rail systems, such as the BMW Group's Sapphire Space simulator, offer large potentials to drive urban scenarios with larger motion scaling factors. However, they also have the potential to do *more wrong*, as false cues that might be negligible on a small system, also become enlarged using higher scaling factors. Clearly, the top priority of any motion cueing system must be to minimize all cueing errors. However, one could argue that the better the motion cueing can potentially be, it may be more important to focus on mitigating what the simulator could be doing wrong instead of further optimizing what it is doing right.

Crucial for the prediction of simulator sickness would be to have a more refined understanding of *how* the mismatches between expected and simulator motion actually contribute to the development of sickness symptoms. General scenario descriptions such as "rural", "highway", or "urban" are not very informative about the tested route. To transfer (often expensively) gained knowledge to scenarios that were not tested before, one must be able to attribute the emergence of simulator sickness to much more specific properties *during* the scenario, such as mismatches in inertial motion while driving through curves, accelerating/decelerating, or along roundabouts. Current sickness models lack any distinction between the different types of errors (such as scaled, missing, and false cues) that may occur in such maneuvers: all different types of errors are treated as equally sickening (Irmak et al., 2023). Furthermore, these models do not include the relative importance of the vehicle axes, such as longitudinal, lateral, and yaw motion. Recent work Kolff et al. (2024) found that in the evaluation of the Perceived Motion Incongruence (PMI) (i.e., a driver's opinion on the (in)congruence of the simulator motion), mostly the accurate reproduction of the lateral and longitudinal specific force channels are important (in this order), with an additional, minor, role of the yaw rate. If simulator sickness is caused by inertial motion mismatches, this might suggest a similar division of the vehicle motion channels on simulator sickness. Currently, most research deals with the topic at hand by acquiring sickness evaluations after a drive ("post-hoc") (Himmels et al., 2022; Mourant et al., 2007). Such evaluations are not detailed enough to link to specific mismatches during the drive. Continuous sickness ratings have been used previously to overcome these issues (Irmak et al., 2021; Qiu et al., 2023). However, a systematic investigation on the importance of mismatches in different vehicle axes as contributors to simulator sickness does not yet exist.

This paper aims at improving the understanding and the predictive capabilities regarding simulator sickness and motion incongruence ratings based on objective mismatches of the simulator motion. In a dedicated driving simulator experiment, 52 participants were driven as passengers ("open-loop") through an urban simulation on BMW's largest driving simulator "Sapphire Space". Twelve variations of motion cueing were tested, inducing scaled, missing, and false cue motion in the vehicle's three main axes: longitudinal (f_x), lateral (f_y), and yaw (ω_z). Subjects provided a Misery Scale (MISC) rating (Reuten et al., 2021) on their level of simulator sickness, as well as a subjective rating on the (in)congruence of the motion after each drive ("post-hoc"). Subjects also identified their current level of simulator sickness continuously *during* each drive using a rating knob, resulting in a continuous measurement, similar to continuous ratings of motion incongruences (Cleij et al., 2018) but a first in simulator sickness research. This allows relating sickness increments to specific events in the simulation, such as different driving maneuvers. An investigation of the importance of (1) the three main axes, (2) the criticality of scaling, missing, and false cue motion, and (3) the importance of axis inconsistency for simulator sickness is performed. The main contribution of this paper lies in providing a better understanding of how motion mismatches cause simulator sickness and motion incongruence to increase.

The paper is structured as follows. Section 2 introduces the considered methods. Results are described in Section 3. This is followed by a discussion in Section 4 and conclusions in Section 5.

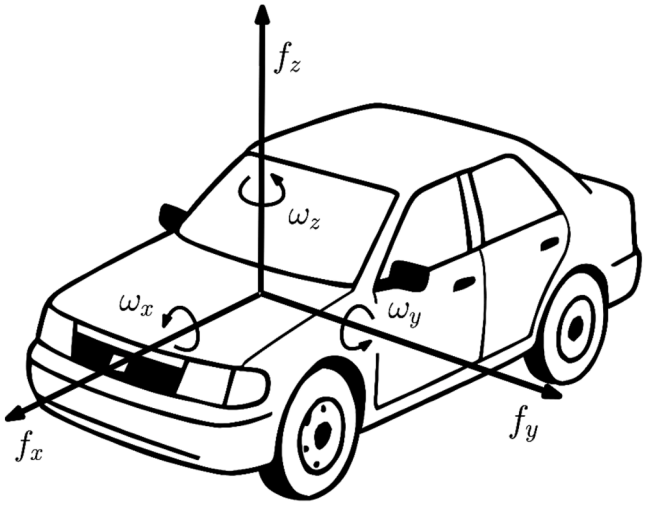


Fig. 1. Specific forces and rotational rates of a car, indicating the right-handed reference system used. The car illustration was generated using DALL-E (OpenAI).

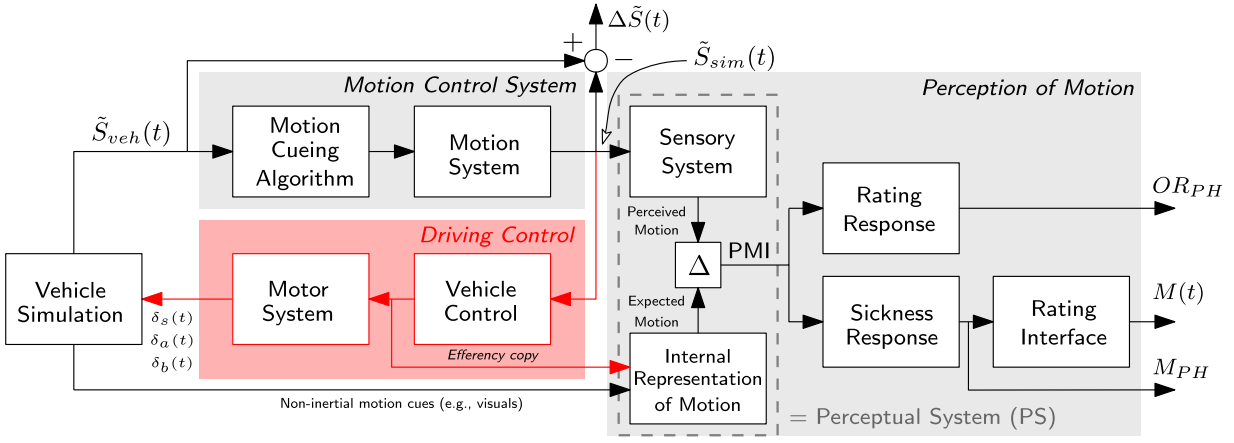


Fig. 2. Block diagram of the human rating of PMI and simulator sickness, through an overall post-hoc MIR (OR_{PH}) and MISC (continuously: $M(t)$, post-hoc: M_{PH}), respectively. The part "Driving Control" (red) is only present in closed-loop driving, and not active in the present experiment. Figure adapted from Kolff et al. (2025).

2. Methods

2.1. Perceived motion incongruence

During the simulation, the vehicle model produces the reference motion that is to be reproduced by the simulator. This motion consists of the three specific forces f_x (longitudinal), f_y (lateral), and f_z (vertical). Furthermore, there are three rotational rates ω_x (roll), ω_y (pitch), and ω_z (yaw), as shown in Fig. 1. Collectively, these motions are described by $\tilde{S}_{veh}(t)$, as shown in Fig. 2, and used as the reference to the *Motion Control System*. This block consists of two parts, the Motion Cueing Algorithm (MCA) and the Motion System. The latter is the physical component of the simulator that produces the actual inertial motion. It is driven by the MCA, which, using knowledge of the configuration and dimensions of the motion system, converts the vehicle motion to simulator motion that can be reproduced by the motion system. Therefore, using the input signals $\tilde{S}_{veh}(t)$, the *Motion Control System* gives the output $\tilde{S}_{sim}(t)$. Differences between the vehicle reference and simulator motion signals are then the objective mismatches, i.e., $\Delta\tilde{S}(t) = \tilde{S}_{veh}(t) - \tilde{S}_{sim}(t)$.

The blocks indicated by *Perception of Motion* then describe how humans perceive the simulator motion. They can notice a difference (Δ) between the physical motion they perceive in the simulator and the motion they would expect to feel from the simulated vehicle. This is known as the PMI (Cleij et al., 2018; Kolff et al., 2024), see Fig. 2. As the driver does not exactly know what the vehicle motion would feel like in a particular situation (i.e., the signal S_{veh} is unavailable to them), they must use their *internal representation* (Stassen et al., 1990) of the vehicle motion based on previous experience and the non-motion cues (e.g., visuals) provided in the simulation.

Individual variations can arise due to different familiarity or experience with the simulated vehicle (Kolff et al., 2024). Note that thus both the expected motion and the vehicle model motion $\tilde{S}_{veh}(t)$ are approximated versions of the real vehicle's motion S_{veh}.

Although ideally the PMI would be measured directly, this is not possible, as it is "internal" to the driver. Instead, a Motion Incongruence Rating (MIR) was asked from the drivers (Cleij et al., 2018). This is a subjective rating that represents the PMI. The PMI and MIR can differ, as the latter can be affected by the *rating response* (Fig. 2). This can be affected by a variety of factors concerning the experiment (Kolff et al., 2024), such as the participant's understanding of the rating method, as well as the (participants response to the) type of rating method used. In the present case, the rating aims to summarize the PMI across the whole drive. Such *overall ratings*, denoted OR_{PH} (see Fig. 2), are known to correlate well with the worst considered maneuver during a drive (Cleij et al., 2018; Kolff et al., 2025, 2024). The subscript $_{PH}$ denotes that the rating was acquired at the end of each drive, i.e., *post-hoc*. In the current experiment, the rating values were on a semantic differential scale, ranging between values of 0 (fully congruent) and 10 (highly incongruent), with steps of 1, to remain close to the work of Kolff et al. (2025, 2024).

Additionally, in closed-loop control (when driving the car oneself), the red path in Fig. 2 would be active, denoted by *Driving Control*. In closed-loop control, an *efference copy* (Kolff et al., 2025; Mulder et al., 2022) of the intended control actions is used to form the expected motion. The expected motion is then not only based on what drivers expect to feel from the non-inertial motion cues, but also what they expect to feel as a result of their own intended control actions of the steering wheel, accelerator pedal, and braking pedal deflections ($\delta_s(t)$, $\delta_a(t)$, and $\delta_b(t)$ respectively). In "open-loop" driving (being driven as passengers, as in the present experiment), the lack of driving control implies that no efference copy is present. This means that the internal representation of motion, and in turn the expected motion, is only fed by the non-inertial motion cues. Kolff et al. (2025) has shown that closed-loop and open-loop driving are rated equivalently in an urban scenario.

2.2. MIR rating model

A rating model is used to predict the ratings based on objective mismatch signals. The latter are defined as the differences in inertial motion (specific forces and rotational rates) between the vehicle motion $\widetilde{S}_{veh,m}(t)$ and the simulator motion $\widetilde{S}_{sim,m}(t)$, i. e., $\Delta\widetilde{S}_m(t)$, with $\widetilde{P}_m(t) = |\Delta\widetilde{S}_m(t)|$. Here, m represents the mismatch type (specific force or rotational rate) and direction, e.g., $m \in [f_x,\ f_y,\ \ldots,\ \omega_z]$. $\widetilde{P}_m(t)$ represents the modeled PMI, which may differ from the real PMI of the driver, as defined in Fig. 2.

In Kolff et al. (2024) a linear model was proposed that predicts the continuous rating of the *average* participant. Its structure consists of a first-order low-pass filter transfer function $H_m(j\omega)$ between the absolute mismatch signal $\widetilde{P}_m(t)$ and a modeled rating signal $\tilde{R}(t)$:

$$\hat{\tilde{R}}(j\omega) = \left(\frac{\omega_c}{j\omega + \omega_c}\right) \sum_m K_{\widetilde{P}_m} \hat{\widetilde{P}}_m(j\omega), \tag{1}$$

with the low-pass filter's cut-off frequency ω_c and the gains of the several mismatch channels $K_{\widetilde{P}_m}$. The $\hat{(\cdot)}$-terms indicate the Fourier transforms. The low-pass filter represents the participants' lagged response to the mismatches. In Kolff et al. (2024) it was shown that the continuous ratings of an Oracle MCA condition as measured in that study could be largely explained when considering the longitudinal specific force mismatch $\widetilde{P}_{f_x}$, lateral specific force mismatch $\widetilde{P}_{f_y}$, as well as the yaw rate mismatch $\widetilde{P}_{\omega_z}$ (i.e., $m \in [f_x,\ f_y,\ \omega_z]$), with the parameters: $\omega_c = 0.52$ rad/s, $K_{f_x} = 0.62$, $K_{f_y} = 1.11$, and $K_{\omega_z} = 1.08$. This model, denoted "p-ORC" was fitted on rating data of the same Oracle algorithm as in the present chapter. It is used instead of the more general "p-ALL", which was fitted on a combination of Oracle and Classical Washout Algorithm (CWA) rating data, as that model does not incorporate a yaw rate mismatch component. Whereas the rating model of Kolff et al. (2024) was developed without deliberate missing or false cue motion, the goal of the present rating model application is to study is validity under such error types.

Kolff et al. (2024) also investigated the relationship between continuous and post-hoc rating signals, from which it was found that the latter is well described by the most incongruent point in the simulation (i.e., the point with the highest continuous rating). Furthermore, the explicit rating relationship $OR = 2.0 + 0.8 \cdot \max[R(t)]$ was found. Thus, using simulations of the continuous rating signals, a prediction of the overall post-hoc can be made.

2.3. Simulator sickness

The second quantity evaluated by the subjects was their own perceived level of simulator sickness. One of the main theories in motion sickness research considers sensory-expectancy conflict as the prime cause for sickness (Bos, 2011; Reason, 1978). Here, the conflict refers to a mismatch between the sensed sensory signals and the expected sensory signals (Irmak et al., 2023). Considering the similarity with PMI here, it is likely that in the context of (driving) simulation, simulator sickness might *result* from PMI. However, simulator sickness is a phenomenon that does not necessarily occur to everyone. Not only can individuals have different susceptibility to motion and/or simulator sickness (some get sick easily, some never), expectation of the motion has also been shown to affect motion sickness (Kuiper et al., 2020). Therefore, individuals with a lot of driving experience might get sick earlier than those without, as they might notice the motion to be incongruent while their less-experienced colleagues may not. The "Sickness Response" block in Fig. 2 represents this individual response.

For measuring the participants' level of simulator sickness, the Misery Scale (MISC) (Bos et al., 2005; Wertheim et al., 2001) (or: Motion Illness Symptoms Classification Scale) is often used. This 11-point symptom-based questionnaire ranges from 0 (no sickness) to 10 (vomiting), see Table 1. Its intermediate values represent different levels of motion sickness, including dizziness, headache, sweat, or stomach awareness (Reason and Brand, 1975). It relies on the fact that stronger sickness symptoms (nausea, vomiting) are

Table 1
The Misery Scale (MISC), adapted from Bos et al. (2005).

Symptom		Score
No problems		0
Slight discomfort but no specific symptoms		1
Dizziness, warm, headache, stomach awareness, sweating, etc.	vague	2
	some	3
	medium	4
	severe	5
Nausea	some	6
	medium	7
	severe	8
Vomiting		10

usually preceded by lighter symptoms (headache, sweating, etc.) (Reason and Brand, 1975). Reuten et al. (2021) found that there is a particular order in which sickness symptoms develop, confirming the progressiveness of the symptom-based MISC. MISC ratings were further found to be highly correlated to unpleasantness ratings (Reuten et al., 2021; de Winkel et al., 2022). The MISC is a more refined and single-item alternative to, for example, the sixteen-item Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993), which is too extensive for real-time use. The MISC is easy to understand and can be used with little training (Bos et al., 2010).

The MISC-value is usually given verbally after each drive, i.e., post-hoc, as will also be done here. This results in values of M_{PH}, as shown in Fig. 2. Another possibility is to ask for MISC values during a drive after a given interval, like every minute (Diels et al., 2023; Hogerbrug et al., 2020) or every 30 or 40 s (Irmak et al., 2022, 2021). A benefit of the latter approach is the higher temporal resolution, as sickness symptoms might build up and/or disappear *during* a drive, which cannot be measured using post-hoc ratings. A drawback, however, is that repeatedly asking subjects to rate their well-being, reminds them to think about how they feel, which may affect the MISC values. Furthermore, even a 30-s resolution might be insufficient to relate specific maneuvers and sickness, especially in very dynamic urban scenarios.

In the present experiment a novel approach to obtaining MISC ratings during the drive was tested. Here, participants could provide the MISC rating at *any* time during the drive, through a rating interface (See Fig. 2), and were instructed to *only* change their rating when they noticed a change in their well-being. As a novelty in simulator sickness research, but similar to its use in general motion sickness research (Qiu et al., 2023), the resulting 'real-time' rating $M(t)$, yields a continuous rating of simulator sickness, rather than measuring on fixed intervals. Through obtaining the MISC as soon as it changes, it should become more clear which maneuvers or parts of a drive affect simulator sickness. While this approach potentially reduces the impact of measurement interference mentioned above, a potential risk is that, in due time, some participants may forget to change the rating.

2.4. Sample

In total, 52 subjects participated in the experiment (43 men, 9 women). All participants were in the possession of a European car driver's license B. The average age was $M = 41$ yrs ($SD = 8$ yrs). 43 participants had experience with simulator driving. The experiment was terminated if a participant reached a post-hoc MISC value of six in two successive drives (*some nausea*), or a value of seven or higher in a single drive (*medium nausea*; for an explanation of MISC ratings, see Table 1. This was the case for four participants, all males. In two additional cases the experiment could not be completed due to technical issues. A total of 46 (52 – 6) complete data sets were obtained. The six incomplete data sets were excluded from analysis.

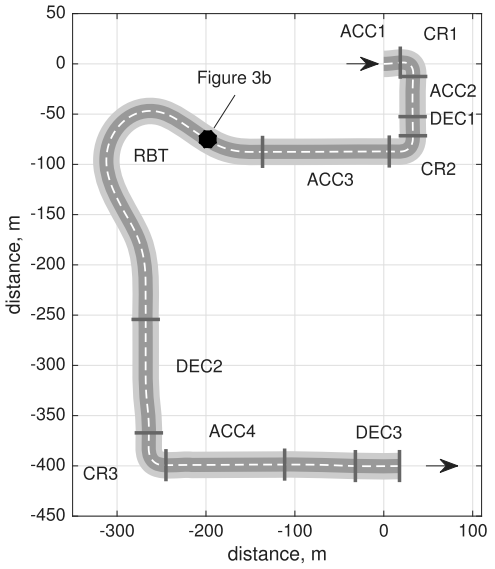
All participants were employees of the BMW Group and participated in the study on a voluntary basis during their regular working hours. Participants were not specifically selected for their own estimated sensitivity to simulator sickness, to obtain a representative group of participants. They were informed about the purpose of the study. The participation could be ended at any time on the participants' initiative without consequences. The experiment was approved following BMW's internal ethics review procedures. All participants provided informed consent.

2.5. Scenario

A single recorded drive through an urban environment was used, see Fig. 3a. This scenario was based on the driven route used in Kolff et al. (2024), although a new recording was used and the route was shortened to last 133.8 s. It consisted of several maneuvers typical for urban driving: accelerations ('ACC#'), decelerations ('DEC#'), three 90° corners ('CR#'), and a roundabout ('RBT'). The highest occurring speed in the recording was 50 km/h. Traffic was not simulated, i.e., there were no other road users.

2.6. Apparatus

The experiment was performed on the BMW Group's Sapphire Space simulator, see Fig. 4a. The motion system has nine DoFs, consisting of three subsystems: a 19.14 m × 15.70 m XY-drive, a 1.15 m stroke hexapod and a 360° yaw-drive. The simulator is fully



(a) Top-down view of the driven route, containing corners (CR#), accelerations (AC#) decelerations (DC#), and a roundabout (RBT#).

(b) The Misery Scale (MISC) value given by the participants as visible in the simulated scene, consisting of a coloured bar and the numeric value (currently set to 7).

Fig. 3. The scenario.

(a) The Sapphire Space simulator (Image: BMW Group)

(b) The iDrive rating knob, used for the extraction of the continuous Misery Scale (MISC) values, indicated by the white box.

Fig. 4. The experiment set-up.

enclosed by a dome, in which a BMW X5 series (G05) vehicle mockup was placed. The steering wheel rotated corresponding to the simulated drive.

The iDrive navigation knob on the center console (Fig. 4b) was used as the rating interface by the drivers to provide the continuous MISC rating (denoted as $M(t)$). Inside the dome, visuals were rendered using Unreal Engine and displayed using 12 Norxe P1 projectors, resulting in a full 360° projection in the dome. The current MISC rating value in the form of a 'rating bar' was visible in the central field of view in a type of head-up display. The size and color of this rating bar changed (see screenshot in Fig. 3b) from rating 0 (short, white) to rating 10 (long, red), to make the rating method more intuitive for drivers to use. The current numerical MISC value of the participants was also displayed on the rating bar. The full MISC including verbal anchors was furthermore displayed on the vehicle Central Information Display (CID) throughout the drives. The velocity of the vehicle was visible on the tachometer on the dashboard and in a head-up display alike screen projection, together with the driving direction (arrows).

Table 2
Scaling factors of the vehicle data in each direction for each condition. Bold values highlight the manipulated motion channels.

	Baseline	Scaling				Missing				False		
Axis	A	B	C	D	E	F	G	H	I	J	K	L
long. (f_x)	0.8	**0.4**	0.8	0.8	**0.4**	**0.0**	0.8	0.8	**0.0**	**−0.8**	0.8	0.8
lat. (f_y)	0.8	0.8	**0.4**	0.8	**0.4**	0.8	**0.0**	0.8	**0.0**	0.8	**−0.8**	0.8
yaw (ω_z)	0.8	0.8	0.8	**0.4**	**0.4**	0.8	0.8	**0.0**	**0.0**	0.8	0.8	**−0.8**

2.7. Independent variables

The experiment tested twelve variations of motion cueing using the optimization-based MCA described in Ellensohn et al. (2019). Here, the simulator motion is optimized beforehand by minimizing the difference between the reference drive and simulator motion along the whole drive. It is referred to as *Oracle* motion cueing due to its knowledge of all future states (only possible when using a pre-recorded drive) and represents the best-possible motion cueing for a given simulator and recorded drive (Kolff et al., 2022). The combination of Oracle motion cueing and the Sapphire Space's large and dynamic workspace allows for motion cueing with high scaling factors (Kolff et al., 2022). For the considered urban scenario, it was iteratively found that values of 0.8 were the highest possible scaling factors in all directions that still fit in the workspace of the simulator. A higher scaling factor would have resulted in the simulator reaching its limits. The corresponding motion was used as the baseline condition 'A'. The other eleven experimental conditions (labeled 'B' to 'L') were variations of the baseline condition 'A', but then containing scaled, missing, and false cue motion in the three main vehicle directions (longitudinal, lateral, and yaw). These three motion were chosen as these typically dominate the motion of vehicles in urban environments (Ellensohn, 2020), compared to the relatively small roll, pitch, and vertical motions.

To investigate the dependency on these specific vehicle axes, the specific forces and rotational rates that act on the simulator cabin must be manipulated. These are the outputs of the Oracle algorithm. Because of the presence of the simulator's yaw-drive, a simulator motion in a certain direction in its inertial frame does not necessarily equate to a motion in the same direction as acting on the simulator cabin. For example, a longitudinal simulator motion under a 90° yaw-drive angle will result in a lateral motion acting on the simulator cabin (Kolff et al., 2023). Thus, to manipulate the forces acting on the cabin, it is not possible to turn off a specific simulator axis. Rather, the non-linear simulator motion must be optimized using Oracle to produce the manipulated motion as its output. Therefore, Oracle was run twelve times separately, once for each condition, using scaling factors on the vehicle reference data (0.4 for scaled, 0 for missing, and −0.8 for false cue motion) in the relevant directions, see Table 2. This thus allows for investigating the effect of scaled, missing, and false cue motion in the three main vehicle DoFs. Note that this also includes the overall scaling factor of 0.8. Although Oracle would also be able to scale down the motion itself as a result of its optimization, this 'pre-scaling' of the vehicle reference input was used to ensure that the motion channels that were *not* manipulated in a specific condition were always scaled down using the same factor. Otherwise, depending on the condition, Oracle could have either less or more workspace available in the other directions, resulting in different scaling factors between the conditions in the non-manipulated axes. The results of these optimizations are shown in Fig. 5. In two conditions ('E' and 'I'), the motion was manipulated in all three axes, such that the axis manipulations are consistent with each other. The scaling factors for the baseline motion (0.8) and scaled motion (0.4) were chosen as they correspond to the upper and lower scaling factor values considered to be acceptable by participants in driving simulation (Berthoz et al., 2013). Note that the various conditions were constructed based on the combinations of the profiles presented in Fig. 5. For example, the longitudinal false cue condition "J" contains the false cue (red) profile from Fig. 5a, and the baseline (grey) profiles from Fig. 5b–f.

Note that only the scaling factors of the f_x, f_y, and ω_z motion channels were varied. The vertical (f_z) channel, as well as the rotational roll (ω_x) and pitch (ω_y) rate channels was exactly the same in all conditions, even in condition 'I'. This resulted in a fairer comparison with the conditions in which the motion was dominant and active. Considering the terrain flatness in the scenario and the corresponding small roll and pitch rotational rates, the magnitudes of these signals were very small (grey lines in 5e-d). In all conditions, Oracle applied tilt-coordination, in which a rotation of the simulator cabin is used to generate a sustained specific force through the gravity vector (Stratulat et al., 2011). For this to be also perceived as a pure sustained acceleration by the participants, the associated rotational motion must be below the perceptual threshold. Therefore, the Oracle was constrained to keep the rotational rates ω_x and ω_y below <3 deg/s (Reymond and Kemeny, 2000). As a rotation around the z-axis does not result in a sustained specific force (Kolff et al., 2023) and cannot be used for tilt-coordination, the ω_z channel was not constrained.

2.8. Procedures

The participants first completed a pre-questionnaire including demographic questions as well as questions related to simulator sickness experience. The latter was done using the Motion Sickness History Questionnaire (MSHQ) (Griffin and Howarth, 2000). Participants were then led to the simulator and given a safety briefing. Each subject experienced all twelve cueing variants, resulting in twelve rides of 133.8 s each (total driving time: about 27 min per subject). The order of the conditions was counter-balanced between participants using the Latin-square method to minimize order effects between the conditions. With 46 participants and 12 conditions, the Latin square was unfinished.

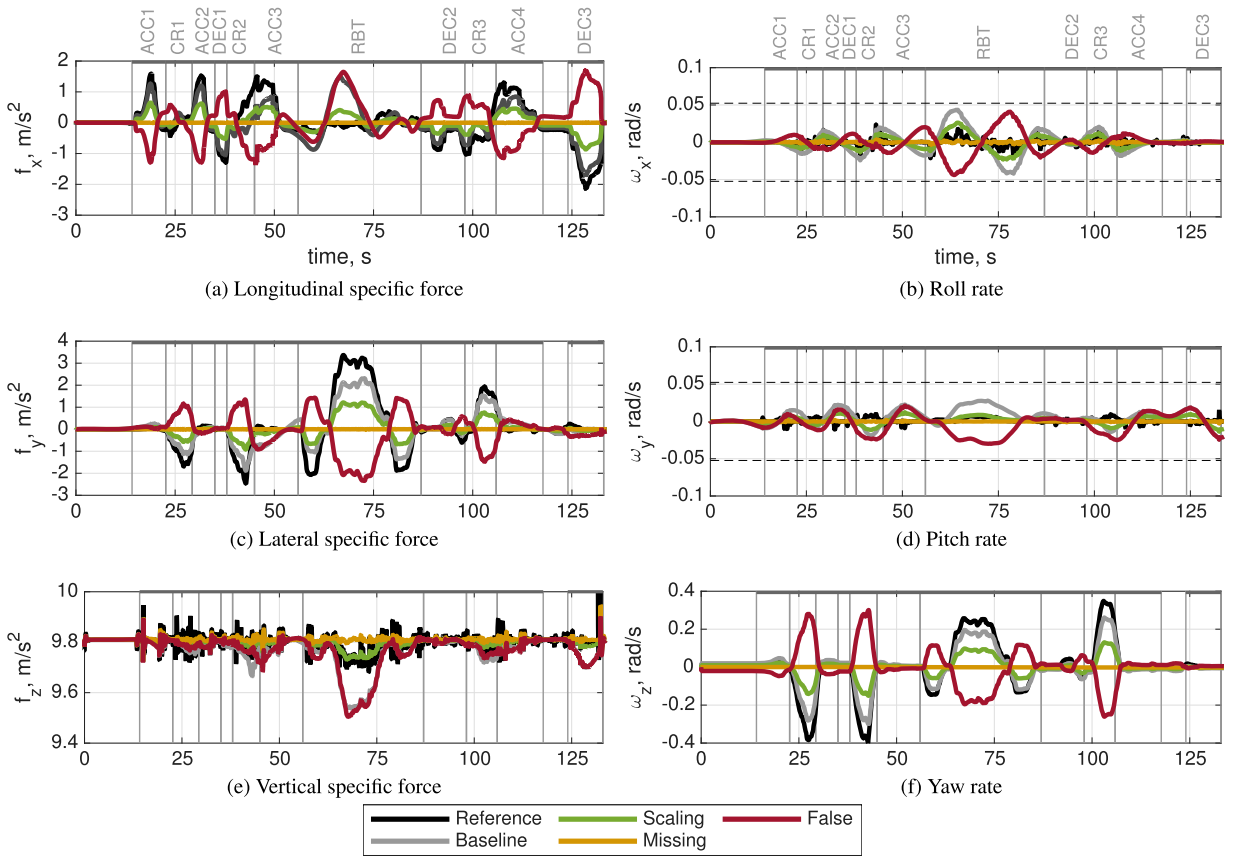


Fig. 5. Motion mismatch profiles of the various conditions. In this figure the black lines represent the vehicle data, grey lines the baseline simulator output (0.8), green lines the scaled output (0.4), yellow lines the missing output (0.0), and red lines the false direction output (−0.8).

During the drives, the participants' task was to continuously evaluate their MISC-level using the rating knob. After each drive (*post-hoc*), participants verbally answered a questionnaire on their MISC level at that moment, as well as their subjective rating on the overall quality of the motion (MIR). The questionnaire further included a question related to the perceived match of motion and visuals, a question on perceived realism, and a question on the sense of presence. Any question wordings and scales, including the MISC, were displayed continuously on the central display inside the vehicle for reference. Only MISC and MIR are further evaluated in the present chapter.

After the sixth drive, a five-minute break was held, in which participants were also allowed to step out of the simulator. This break was included to allow for some relief for the participants' eyes and from the rating tasks, as well as to reduce the level of simulator sickness to make sure that more participants could finish the whole experiment. Due to this break, it is possible that the carry-over effect of simulator sickness that is present between all conditions is not present between the conditions directly before and after the break, thus affecting the simulator sickness results. However, due to the different order in which the conditions were presented to the participants, the condition after which this break occurred differed per participant, which should thus cancel out over all participants.

2.9. Hypotheses

The two dependent measures in the experiment, MIR and MISC, are clearly related. However, their exact relation is unknown and likely to be nonlinear, as high MIR ratings may not necessarily lead to high MISC ratings (i.e., motion can be mismatched, but not sickening). Furthermore, some participants can be expected to be quite capable of indicating incongruent motion while for them this incongruent motion does not lead to simulator sickness.

The following hypotheses are tested with regard to motion incongruences (MIR):

- **H1:** False cues (conditions J-K-L) receive the highest MIR values, followed by missing cues (conditions F-G-H-I), and then the scaled motion cues (conditions B-C-D-E).
- **H2:** Mismatches in the f_y (lateral), f_x (longitudinal), and ω_z (yaw) channels lead to an increase in MIR, in this order of severity *(confirmation of Kolff et al., 2024)*.

The following hypotheses are tested with regard to simulator sickness (MISC):

Table 3
Effects considered in the model. The fixed effect is nested within the levels of the random effects.

	effect type	levels
cueing condition	fixed	A-L (see Table 2)
axis inconsistency	random	yes, no
axis	random	f_x, f_y, ω_z
error type	random	scaling, missing, false

Table 4
Cueing condition as nested in 'axis inconsistency', 'axis', and 'error type'. Empty cells (·) indicate that no level was assigned to the cueing condition, resulting in a different number of observations depending on the random effects included in each tested model.

condition	axis inconsistency	axis	error type
A (0.8 0.8 0.8)	no	·	·
B (**0.4** 0.8 0.8)	yes	Long.	scaling
C (0.8 **0.4** 0.8)	yes	Lat.	scaling
D (0.8 0.8 **0.4**)	yes	Yaw	scaling
E (**0.4 0.4 0.4**)	no	·	scaling
F (**0.0** 0.8 0.8)	yes	Long.	missing
G (0.8 **0.0** 0.8)	yes	Lat.	missing
H (0.8 0.8 **0.0**)	yes	Yaw	missing
I (**0.0 0.0 0.0**)	no	·	missing
J (**−0.8** 0.8 0.8)	yes	Long.	false
K (0.8 **−0.8** 0.8)	yes	Lat.	false
L (0.8 0.8 **−0.8**)	yes	Yaw	false

H3: False cues (conditions J-K-L) receive the highest MISC values, followed by missing cues (conditions F-G-H-I), and then the scaled motion cues (conditions B-C-D-E).

H4: Mismatches in the f_y (lateral), f_x (longitudinal), and ω_z (yaw) channels lead to an increase in MISC, in this order of severity.

3. Results

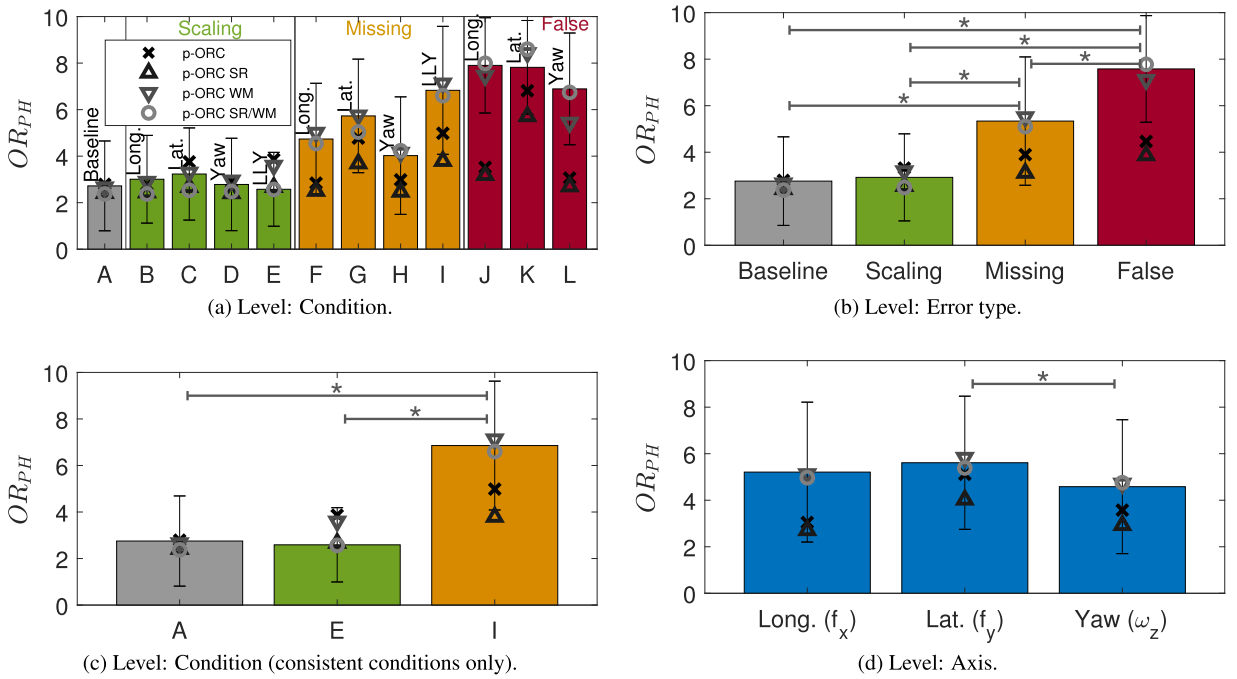
The experiment had an incomplete factorial design with two varied factors: the 'axis' (longitudinal, lateral, or yaw) in which the mismatch was induced, and the 'error type' (scaling, missing, false). The levels of one factor were not always present in all levels of another factor (e.g., condition 'A' does not have a manipulation in 'axis'). This makes the use of a typical two-way Analysis of Variance (ANOVA) for statistical analysis of the data infeasible. For the post-hoc ratings, linear models were therefore built with several fixed and random predictors: 'condition' (fixed) nested within the random factors 'axis', and 'error type' (Table 3). Then, it was determined which of these predictors explain a relevant share of variance in post-hoc MIR.

Although not part of the hypotheses, the variable 'axis inconsistency' was also considered, to exploratively investigate whether there are effects of manipulating one axis in isolation in contrast to manipulating all axes at the same time. Therefore, 'axis inconsistency', 'axis', and 'error type' are considered as (higher level) random effects in the linear model, in which the 'cueing condition' (considered as fixed effect) is nested, see Table 3. The levels of the random effects were assigned to the cueing condition levels as described in Table 4.

3.1. Post-hoc MIR ratings

The post-hoc overall MIR ratings of the twelve conditions are illustrated in Fig. 6. Here, for the predictors 'subject' and 'condition', the Intraclass Correlation Coefficient (ICC) and R^2 values are displayed. Residual maximum likelihood was applied as method for mixed models, as this is more suitable than maximum likelihood when interested in the relevance of random effects. We do not report significance tests, as likelihood ratio test*p*values have been found to be conservative (Pinheiro and Bates, 2000). Further, step-wise testing has been criticized in the past as a misuse of hypothesis testing (Whittingham et al., 2006). Instead, the data are analyzed based on Akaike's Information Criterium (AIC). Here, the measure of fit is calculated based on deviance, while penalizing complexity (Bolker et al., 2009). While the AIC avoids step-wise procedures, it may still over fit, and thus suffer from the same issue like the likelihood ratio test's *p*. Differences in AIC values of 2 or more are considered relevant (Burnham et al., 1998).

The random effects were first added to the model independently, to see whether they can be considered relevant at all (Field et al., 2012). This step shows only a small decrease in AIC when including 'axis inconsistency' and will therefore not be considered further. A better model fit is obtained when including 'axis' and 'error type' as random factors. Especially the 'axis' provides a large decrease in AIC and is thus the most critical predictor to include in the model. Including both axis and error type (model 5 in Table 5) results



(a) Level: Condition.

(b) Level: Error type.

(c) Level: Condition (consistent conditions only).

(d) Level: Axis.

Fig. 6. Post-hoc Motion Incongruence Rating (MIR) values and model predictions for the considered levels. Green, yellow, and red indicate scaling, missing, and false cue errors, respectively. Grey (condition 'A') is the baseline. For each condition in Fig. 6a, the manipulated Degree of Freedom (DoF) is indicated (longitudinal, lateral, or yaw). LLY indicates manipulations in longitudinal, lateral, and yaw directions. The star symbols indicate significant differences.

Table 5
Model performance comparison with varying predictors, dependent measure: post-hoc Motion Incongruence Rating (MIR). The values of Intraclass Correlation Coefficient (ICC) and R^2 cannot be provided for the random effects models. Model 5 (bold) performs best in terms of the Akaike's Information Criterium (AIC).

model #	predictors	ICC (adj.)	ICC (unadj.)	R^2 (conditional)	R^2 (marginal)	AIC
0	subject	.08	.08	.08	.00	2976.93
1	condition; subject	.24	.13	.59	.45	2575.28
2	condition, subject, axis inconsistency					2574.31
3	condition, subject, error type					2566.34
4	condition, subject, axis					1934.49
5	**condition, subject, error type, axis**					**1925.96**

in the lowest AIC value and is therefore the best model. It performs better than model 1 (fixed-effect only) and the models including only one random effect (models 3 and 4), confirming that there are indeed effects of axis and error type on the MIR, see Fig. 6a.

Fig. 6b shows the comparison of the error type (levels tested: baseline, scaling, missing, false). The baseline cueing condition A was included as a fourth level. The MIRs were lowest for baseline and scaling, followed by missing and then false cues in ascending order, confirming Hypothesis H1. The effects were subjected to Bonferroni-corrected pairwise post-hoc tests to determine which factor levels of the random effects actually differ. These tests indicated no difference between baseline (A) and scaling, but significant differences across all other conditions. As the number of observations and conditions with each level considered in the post-hoc test vary, the consistent cueing conditions A (0.8 0.8 0.8), E (0.4 0.4 0.4), and I (0.0 0.0 0.0) were also considered separately, see Fig. 6c. This comparison confirms that there was no significant difference between the baseline (A) and scaled motion (E), but both of them being advantageous over missing motion (I). As there was no condition with full, consistent false cue motion (−0.8 −0.8 −0.8), a similar comparison for false cue motion was not possible.

Finally, Fig. 6d shows the MIRs grouped per axis. The lateral specific force manipulations result in the highest MIR (i.e., worst ratings), followed by the longitudinal specific force and then the yaw rotational rate. This order is the same as described in Kolff et al. (2024) and supports Hypothesis H2. However, the post-hoc tests show a significant difference only between the lateral and the yaw axis. This indicates that a mismatch in the yaw rate (ω_z) affects the MIRs less negatively than a mismatch in the lateral specific force (f_y) does. Thus, decreasing the lateral specific force mismatches are more important for reducing MIRs than the yaw rate mismatches. There were no significant differences across the mismatches of f_x and f_y or f_x and ω_z. Thus, Hypothesis H2 can only be partially confirmed with statistical significance.

Table 6
Gains and cost function values of the four evaluated MIR rating models. The gains of the models p-ORC/p-ORC SR come from Kolff et al. (2024).

model	ref. scaling	K_{f_x}	K_{f_y}	K_{ω_z}	K_{mc}	K_{fc}	$\sum(OR_{PH} - \widetilde{OR}_{PH})^2/12$
p-ORC	1.0	0.62	1.11	1.08	1.0	1.0	3.82
p-ORC SR	0.6	0.62	1.11	1.08	1.0	1.0	5.55
p-ORC WM	1.0	1.12	0.73	3.49	2.0	2.0	0.33
p-ORC SR/WM	**0.6**	**1.05**	**0.65**	**4.51**	**3.0**	**3.0**	**0.20**

Table 7
Model performance comparison with varying predictors, dependent measure: post-hoc Misery Scale (MISC). Intraclass Correlation Coefficient (ICC) and R^2 values cannot be provided for the random effects models. Model 5 (bold) performs best in terms of the Akaike's Information Criterium (AIC).

model #	predictors	ICC (adj.)	ICC (unadj.)	R^2 (conditional)	R^2 (marginal)	AIC
0	subject	.53	.53	.53	.00	1568.49
1	condition, subject	.58	.55	.61	.06	1527.93
2	condition, subject, axis inconsistency					1529.93
3	condition, subject, error type					1497.96
4	condition, subject, axis					1205.59
5	**condition, subject, error type, axis**					**1179.26**

3.2. MIR rating predictions

The p-ORC model of Kolff et al. (2024) is now applied to check its validity under the various motion manipulations. Fig. 6a shows the model (crosses), predicting the overall ratings per condition. The model overestimates ratings of the scaling conditions and underestimates those of the missing and false cue conditions. Furthermore, more clearly visible in Fig. 6d, a different balance in the three axes is present than what the rating model predicts. Based on the presented findings of the experiments' measured MIRs, two improvements to the rating model can be proposed.

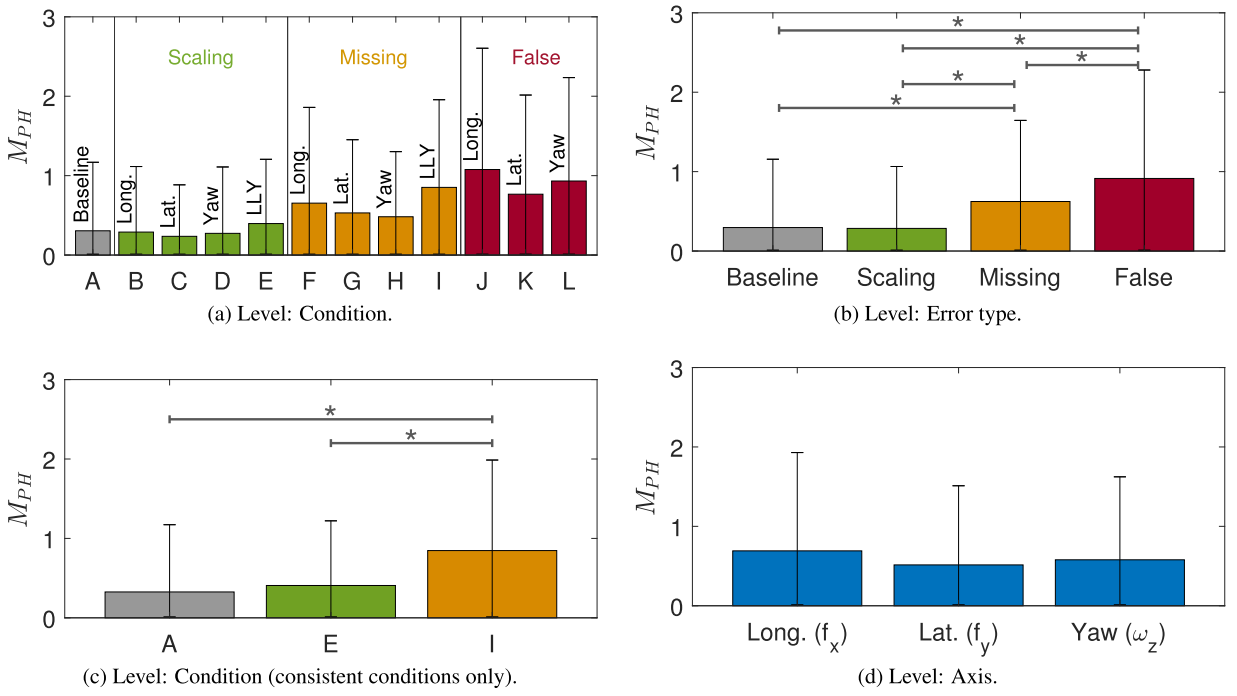
First, considering the finding that there is no significant difference between the baseline (scaling of 0.8) and scaling motion (scaling of 0.4), which is in line with the preferred scaling range of $0.4 - 0.8$ found by Berthoz et al. (2013), an adaptation to the input of the rating model can be made. Instead of taking the full vehicle reference in the calculation of $\tilde{S}_m(t)$ in (1), a reference scaling of $0.6\tilde{S}_m(t)$ is applied. This results in the new model "p-ORC SR", with SR indicating "Scaled Reference". This model is indeed better able to predict the scaled conditions (as indicated by the triangles, also visible in Fig. 6c and b), but underpredicts the error type and axis levels.

The second improvement applied to the model was therefore to consider additional gains on the missing and false cue conditions, i.e., K_{mc} and K_{fc}. The rationale is that humans might rate these types of motion disproportionally worse than that the linear rating model predicts (Cleij, 2020). These gains were only applied on the mismatch signals in which the missing or false cue motion was active. The gains of the three axes are also adjusted, i.e., K_{f_x}, K_{f_y}, and K_{ω_z} to find updated values compared to Kolff et al. (2024). Together, the gains are optimized to minimize the average quadratic difference between the measuring and predicted post-hoc overall ratings, i.e., $\sum(OR_{PH} - \widetilde{OR}_{PH})^2/N_{cond}$, with $N_{cond} = 12$ the number of conditions. The optimization was performed using a Nelder-Mead algorithm in Matlab. As this procedure does not guarantee finding a global optimum (Kolff et al., 2024), the procedure was repeated 50 times with randomly generated initial values of the five gains. To simplify the procedure, the missing and false cue gains K_{mc} and K_{fc} were only allowed to attain more physically interpretable values, i.e., rounded to half values, e.g., $K_{mc} \in [0.0,\ 0.5,\ 1.0,\ \ldots\ 4.0]$.

The resulting weighted mismatches (WM) model also includes the SR logic and is therefore denoted as p-ORC SR/WM. This model, with updated weights, indeed performs much better in predicting the overall post-hoc ratings, as visible by the circles in Fig. 6a. Table 6 shows the parameters of the various models as well as the fit quality, confirming that the model p-ORC SR/WM strongly outperforms both the p-ORC and p-ORC SR models. The weights for the missing and false cue motions, $K_{mc} = 3.0$ and $K_{fc} = 3.0$, respectively, were best found to measure the rating data. This indicates that these error types were rated three times higher than the linear rating model p-ORC predicts. The model also outperforms a variant without the SR logic, i.e., fitted to the full reference motion (p-ORC WM).

3.3. Post-hoc MISC ratings

The post-hoc MISC data were subjected to the same analysis steps as the post-hoc MIR ratings. An analysis on the AIC of the various model predictors (Table 7) shows that, similar to the post-hoc overall MIR analysis, the model fit did not improve significantly when adding 'axis inconsistency' as random effect. The fit did improve when adding 'error type' and 'axis' independently compared to the fixed effect-only model. Similar to the post-hoc overall MIR predictors, the largest AIC reduction of a single predictor is achieved using the 'axis'. Furthermore similar, the random intercept model 5, including axis and error type, has the lowest AIC value and thus performed best. This indicates effects of both error type and axis on simulator sickness.



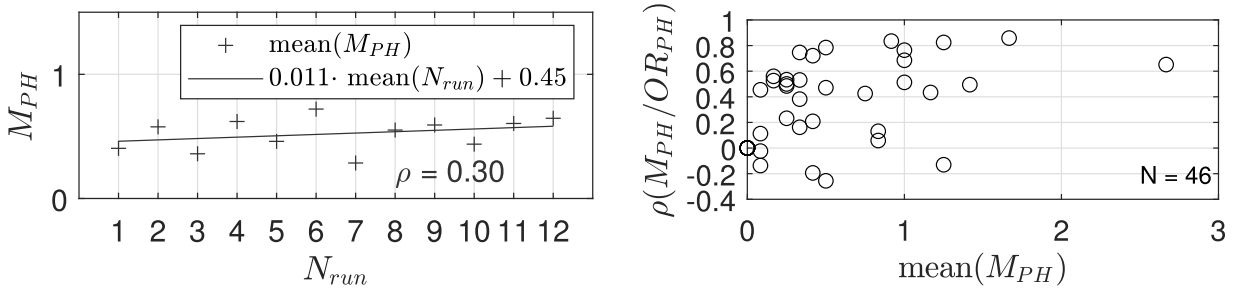
(a) Level: Condition.

(b) Level: Error type.

(c) Level: Condition (consistent conditions only).

(d) Level: Axis.

Fig. 7. Post-hoc Misery Scale (MISC) values for the considered levels. Green, yellow, and red lines indicate scaling, missing, and false cue errors, respectively. Grey line (condition 'A') is the baseline. For each condition in Fig. 6a, the manipulated Degree of Freedom (DoF) is indicated (longitudinal, lateral, or yaw). LLY indicates manipulations in longitudinal, lateral, and yaw directions. The star symbols indicate significant differences.



(a) The post-hoc Misery Scale (MISC) averaged over participants for consecutive runs with a linear regression. The small slope (0.011) shows that sickness build-up during the experiment was low, indicating little carry-over between runs.

(b) The mean post-hoc Misery Scale (MISC) given versus the Pearson correlation between the post-hoc MISC and Motion Incongruence Rating (MIR) ratings per subject.

Fig. 8. Relations of the post-hoc Misery Scale (MISC) ratings.

Fig. 7a shows the post-hoc MISC distributions per condition. Note that here it is clear that, when considering the whole group, the participants did not get very sick, considering $M_{PH} < 3$. The distributions grouped per error type are shown in Fig. 7b. Here, the groups differ significantly except for baseline and scaling (similar to the MIR results), with baseline and scaling inducing the lowest ratings of simulator sickness, followed by missing and false. Conditions A (0.8 0.8 0.8), E (0.4 0.4 0.4), and I (0.0 0.0 0.0) were again also considered separately, confirming that there was indeed no significant difference between the baseline and scaled motion conditions, but both the baseline and scaled motion conditions being advantageous over missing motion (see Fig. 7c); Hypothesis H3 can therefore be confirmed. Finally, Fig. 7d shows the conditions grouped by axis. Here, the longitudinal specific force motion induces most sickness, followed by the yaw rate motion, and then the lateral specific force motion. The post-hoc tests did not yield any significant results, indicating that effects were too small (and hence: practically less relevant) to be detected by the corrected post-hoc tests (Fig. 7d); Although the nested model analysis from Table 7 points at the existence of an effect, Hypothesis H4 cannot be confirmed.

One possible concern of the experiment was the occurrence of carry-over effects between the runs. Fig. 8a shows the increase in M_{PH} over the twelve runs during the experiment, averaged over the participants. Note that, due to the different order of the

experimental conditions presented to the participants, the actual underlying tested condition in each run differs per participant. The straight line represents a linear regression to the data points with a slope of 0.011. This implies that over the twelve runs, there is a slight build-up in MISC, resulting in an overall MISC increase of $12 \cdot 0.011 = 0.13$. This low value shows that participants barely got more sick *throughout* the experiment, showing that the carry-over effects between the conditions were indeed limited. However, the break that was taken between runs 6 and 7 is visible, as this holds the largest decrease in M_{PH} between any two runs.

3.4. Relating post-hoc MIR and MISC ratings

Considering that the extraction of the post-hoc MIR and MISC values occurred at the same point in time, it is possible that participants' answer for both are directly correlated. Therefore, Fig. 8b shows, for each subject, the Pearson correlation between the post-hoc MISC and MIR values as a function of the average post-hoc MISC value. The figure shows that for subjects that do not get sick (i.e., low MISC_{PH}, the correlation values show a large range between -0.4 and 1.0. For most subjects (83 %) the correlation is low (< 0.6), indicating that they answered the questions on MIR and MISC in different ways. However, for subjects with higher MISC values, these values correlate better to the MIR, up to a correlation of 1.0, whereas low correlations do not occur anymore. For these subjects, how sick they feel corresponds to how they disapprove the motion. This might indicate that the MIR can indeed be affected by the participants' current level of simulator sickness. The participants with a negative correlation gave lower MIR values for conditions where they indicated to get more sick (higher MISC), although this only happened to five participants and with low correlation values (< -0.2), which might not be of practical significance.

3.5. Continuous MISC ratings

The continuous MISC ratings obtained *during* the simulations are shown in Fig. 9. Note that the subfigure names correspond to the conditions ('A' to 'L'). The lines represent the mean values over all participants; the spreads indicate the 95 % confidence intervals. The maneuvers as defined in Fig. 3a are indicated as a reference. Note that the values are generally low (< 2), as they also include the participants that rated their simulator sickness with 0 all the time.

The colors green, yellow, and red indicate the scaling, missing, and false cue conditions, respectively, similar to Figs. 6 and 7. In all figures, the coloured circles indicate the post-hoc MISC values given after each drive, averaged over all participants. As explained in Section 2.8, participants were instructed to let the post-hoc MISC value represent the level of simulator sickness at that point in time, i.e., after the run. Therefore, this should be similar to the last continuous rating given, although there is some time (approximately ten seconds) between their extraction. Generally, these values indeed correspond well; in 76 % of the cases the same value was given (over all conditions), which shows the validity of the continuous MISC rating method. Note that the continuous rating value was not reset to 0 at the start of each run. Therefore, at the start of each run it starts at the last logged value of the previous run. As is visible in most conditions (such as Fig. 9K), participants here decrease the $M(t)$ value. This is likely caused by the small 'breaks' between the conditions, lasting around 30 s, allowing for some relief and resulting in participants decreasing the rating right at the start of the next run.

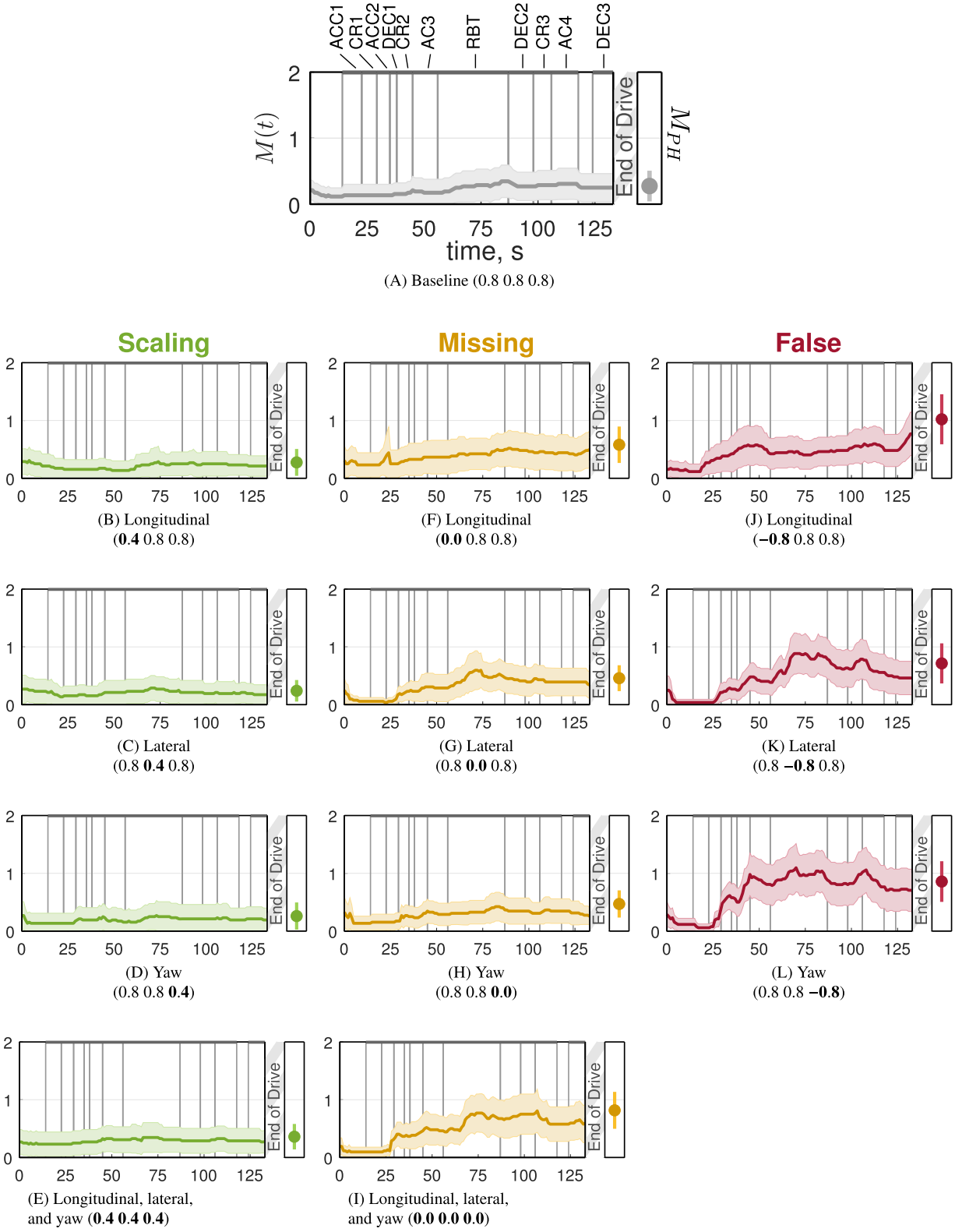
Considering the difficulty of statistically analyzing differences in continuous ratings, the analysis on the differences between the conditions is only provided qualitatively. Comparing the conditions without axis inconsistency (Fig. 9A and E) shows that whereas lower MCA gains do not result in an increase in simulator sickness, even though mismatches are present, providing *no* motion in the three main axes *does*, see Fig. 9I. The ratings are even similar in magnitude to the conditions with false cue motion (9J-L). For the scaled conditions with axis inconsistencies (Fig. 9B-D), i.e., conditions were only one axis was scaled down to 0.4, no clear increase in sickness is visible over time. Comparing scaled and missing motion (with axis inconsistencies, Fig. 9B-D vs. Fig. 9F-H), the MISC-values *do* increase over time. This shows that providing a form of scaled motion, even in case of inconsistencies, has a beneficial effect on simulator sickness compared to when a motion channel is not active at all. Finally, for the false cue conditions (Fig. 9J-L), the MISC values are the highest. Here, it is also visible that the MISC-values increase as a function of specific maneuvers. For the condition 'J', with false cue motion in the longitudinal channel, the reported MISC values especially increase after strong longitudinal false cues, for example at 'DEC3', the final braking maneuver. Similarly, conditions 'K' and 'L' increase most after corner maneuvers, such as the roundabout ('RBT').

4. Discussion

The presented study investigated the effects of motion mismatches on motion incongruence ratings as well as simulator sickness. Scaled, missing, and false motions were systematically applied to the longitudinal (f_x), lateral (f_y), and yaw (ω_z) channels. Participants rated their level of simulator sickness using the MISC (Reuten et al., 2021; Wertheim et al., 2001), both continuously during the drive as well as directly after (post-hoc). They also rated the incongruence of the motion, using a post-hoc MIR. They also rated their level of simulator sickness using the MISC (Reuten et al., 2021; Wertheim et al., 2001), both continuously during the drive as well as post-hoc.

4.1. Post-hoc MIR ratings

The post-hoc MIR ratings show strong evidence of an effect of error type, with false cues inducing the highest MIRs, followed by missing and scaled cues, confirming Hypothesis H1 with regard to post-hoc ratings. There was no significant difference across



(A) Baseline (0.8 0.8 0.8)

(B) Longitudinal (**0.4** 0.8 0.8)

(F) Longitudinal (**0.0** 0.8 0.8)

(J) Longitudinal (**−0.8** 0.8 0.8)

(C) Lateral (0.8 **0.4** 0.8)

(G) Lateral (0.8 **0.0** 0.8)

(K) Lateral (0.8 **−0.8** 0.8)

(D) Yaw (0.8 0.8 **0.4**)

(H) Yaw (0.8 0.8 **0.0**)

(L) Yaw (0.8 0.8 **−0.8**)

(E) Longitudinal, lateral, and yaw (**0.4 0.4 0.4**)

(I) Longitudinal, lateral, and yaw (**0.0 0.0 0.0**)

Fig. 9. Average continuous Misery Scale (MISC) ratings as a function of time (left) and Post-Hoc (PH) MISC ratings (right) for each condition. The shaded areas represent the confidence intervals. The circles are the average post-hoc MISC values, with the vertical bars representing the confidence intervals.

baseline and scaled motion. There was evidence of an effect of the axis. The lateral specific force (f_y) mismatch was indeed found to be the most critical for the MIRs, followed by longitudinal specific force (f_x), then yaw rate (ω_z) mismatches. This is in line with the findings of Kolff et al. (2024). Significant differences were only found between ω_z and f_y. Concluding, these findings only partially conform Hypothesis H2.

4.2. MIR rating predictions

Applying the model of Kolff et al. (2024) resulted in overall values that underestimated rating of the missing and false cue conditions. The model that best predicts the data uses a combination of vehicle reference input scaling of 0.6 and applying additional weights of 3 to the missing and false cue motion. This thus results in an updated model structure, compared to (1):

$$\hat{\tilde{R}}(j\omega) = \left(\frac{\omega_c}{j\omega + \omega_c}\right) \sum_m K_{MFC} K_{\tilde{P}_m} \left(\left| 0.6\hat{\tilde{S}}_{veh,m}(j\omega) - \hat{\tilde{S}}_{sim,m}(j\omega) \right| \right), \tag{2}$$

with K_{MFC} the gain for the missing and false cue motion. Furthermore, the weights of the three mismatch channels were retuned through a fitting procedure to accurately predict the ratings of the twelve conditions tested in the experiment. Due to the "extreme" conditions, where full false cue motion is not representative of realistic driving simulator experiments, the found mismatch weights might be less representative than the parameters found in Kolff et al. (2024), which was further validated in Kolff et al. (2025). Further validation through future work is required to test whether the updated model may indeed still be valid for realistic motion profiles such as presented in Kolff et al. (2024) and Kolff et al. (2025).

4.3. Post-hoc MISC ratings

The analysis of the post-hoc MISC ratings indicated no relevant effect of axis mismatches. Regarding effects of different error types, false cues had the most negative effect on rated simulator sickness, followed by missing cues, and then scaled motion, confirming Hypothesis H3. Ratings for scaled motion did not significantly differ from those obtained in baseline motion. Although the model comparisons suggest an effect of the axis, the pairwise post-hoc tests were all non-significant, such that Hypothesis H4 cannot be accepted. Longer exposure times or more distinctions between the conditions would likely have been beneficial to find stronger differences in the post-hoc MISC ratings.

The confirmed Hypothesis H3 stated that false cues are worse than missing cues, and these are worse than scaling errors. Similar to the confirmed Hypothesis H1, both hypotheses can be interpreted differently, namely that false cues are not just worse than missing cues, but also *disproportionately* worse. The fact that false cues are rated worse than missing cues might explained by their larger objective mismatch. Hence, it is recommended to further investigate the roles of scaling, missing, and false cues in the case of equal mismatch magnitudes, as this could reveal whether the *type* of error induces relatively high or little simulator sickness and/or MIRs.

4.4. Continuous MISC ratings

Continuous MISC ratings of the level of simulator sickness hlwere also recorded. For the first time in simulator sickness research, participants could change the MISC hlrating at their own initiative These acquired ratings provided results that seem intuitively correct, namely that the ratings increase for maneuvers relevant for the specific condition (corners for lateral and yaw motion, accelerations/decelerations for longitudinal motion). Considering the similarity of the last given values and the post-hoc MISC – with post-hoc ratings being widely accepted as a measure of simulator sickness – this supports that the simulator sickness is a direct result of the motion mismatches, making the continuous rating method seem valid for measuring MISC values. The continuous MISC rating results furthermore allow an investigation how unpleasantness and/or simulator sickness is directly related to the mismatches of specific maneuvers. Reducing these mismatches will then automatically reduce simulator sickness. For experiments focusing on the modelling of motion or simulator sickness, such as Irmak et al. (2023), the continuous rating method thus also provides a useful alternative tool for obtaining high-resolution data on sickness. As a logical next step, it is suggested to investigate whether the acquired continuous MISC ratings can be predicted using a MIR rating model. For example, it is possible that the level of simulator sickness can be represented by an integrated MIR over time. There may be, however, substantial differences between predicting the MIR and the MISC. The MIR serves as an almost instantaneous method of measuring motion incongruences, whereas the MISC measures more slowly occurring effects. Brief, subtle cueing errors are therefore measurably expressed in the MIR, but will be virtually invisible in the MISC. In that case, integrating the MIR over time may thus differ from the MISC. Whereas this goes beyond the scope of the current paper, this could be done, for example, in a study in which both the MIR and the MISC are measured continuously and subsequently compared.

4.5. Implications

The presented quality levels in terms of MIR and MISC provide useful insights for planning future simulator experiments. Missing motion was shown to be the second strongest contributor to simulator sickness and the MIR. This suggests that dynamic simulators are beneficial over static ones for a dynamic urban scenario. No significant difference was found between scaled and baseline motion. This shows that for the present urban simulation, the scaled motion (scaling of 0.4) is just as good as the baseline motion (scaling of 0.8), which was the maximal motion possible in the simulator workspace. This corresponds to the range of acceptable scaling factors

(0.4 – 0.75) found by Berthoz et al. (2013). Even if the baseline motion is possible on a given simulator (Such as BMW's Sapphire Space simulator), working with a lower scaling factor might be beneficial when designing motion cueing for a driving simulator study. Possible inaccuracies in the motion cueing or the vehicle model might be less amplified, making the motion cueing easier to design or tune. Thus, for a scaling of scaling of 0.4, an acceptable state of motion cueing might be reached earlier and/or with less effort while at the same time reaching a similar level of MIR and MISC compared to applying a scaling of 0.8. For smaller simulators, with allowable scaling factors of 0.2 or 0.1, it is suggested that future work investigates whether similar levels of MIR and MISC uphold, or if these levels tend more towards the higher (i.e., worse) levels observed for the missing motion.

In the present experiment, the false cue motion, in which the motion was inverted, was found to be the strongest contributor to simulator sickness. Hence, it is of utmost importance to minimize false cues when conducting simulator experiments. Even though the false cue conditions themselves are not realistic motion conditions for a whole drive, the false cue motion in the maneuvers (partially) do correspond to washout motion. In washout, the simulator is moved back to its neutral position after a maneuver. This phenomenon, common in filter-based algorithms, is applied in motion cueing to ensure that the simulator is "ready" to cue the next maneuver. The present results regarding simulator sickness highlight the importance of methods to reduce washout effects, such as optimization-based methods that not only better exploit the available motion space, but can also keep the washout motion under the perceptual threshold.

Note, however, that the presented inverted motion is a form of false cue motion, whereas also other forms of false cue motion exist. For example, even though the false cue motion was completely inverted in the experiment, it is still derived from realistic car motion (in the present case, the baseline motion). However, false cue motion could also occur completely random. For example, pre-positioning can be employed by MCAs to move the simulator below the perceptual threshold of the humans to increase the available workspace of the simulator. If this motion *is* perceived, for example by faulty MCA tuning, this can create a false cue without any relation to the currently presented vehicle motion. The lower predictability of this motion might result in more sickness (Kuiper et al., 2020). Specific future research in this direction is thus recommended.

With regard to axis effects, the acquired findings are especially helpful when planning the use of a dynamic simulator system with one rail (X or Y), which can usually be used to either redundantly provide f_y or f_x motion. When there is no specific dynamic phenomenon that is of interest, the findings suggest using the simulator with a redundant f_y axis to achieve the lowest (i.e., best) motion incongruences in an urban scenario.

The analysis of the post-hoc MISC evaluations is supported by the continuous MISC ratings. Although the latter were only analyzed descriptively, the severity of the various effects in the continuous ratings corresponds to the post-hoc ratings. Therefore, the continuous rating method can be considered promising with regard to identifying the maneuver-specific causes of sickness. The method can therefore be recommended for use in future investigations, as it reduces the workload compared to extraction MISC-values at a fixed point in time, such as in Hogerbrug et al. (2020), Diels et al. (2023), Irmak et al. (2021). Furthermore, future work can explicitly look into modelling the continuous ratings, similar to models applied to predict MIR (Cleij et al., 2018; Kolff et al., 2024). Considering the strong relationship between the continuous and post-hoc MISC, this would be a useful method to be able to predict post-hoc MISC ratings. This would then allow for identifying simulators, MCA settings, and scenarios that induce a high degree of simulator sickness prior to a driving simulator experiment.

This also highlights another strength of the continuous method and the importance of the post-hoc MISC interpretation: post-hoc MISC ratings correspond to the state of the driver at the end of each drive, but do not necessarily represent sickness experienced *throughout* the drive, which can vary significantly over time. The continuous evaluation method provides the benefit of being useable in closed-loop (participants driving themselves) driving experiments, compared to the interval-based methods (Diels et al., 2023; Hogerbrug et al., 2020), as it can be changed on the participant's own initiative and induces little intervention and/or distractions to any other tasks, such as driving.

A final point of consideration considers the extraction of the post-hoc MIR and MISC values. It was found that the worse subjects feel, the better the post-hoc MIR values correlate to the post-hoc MISC. It is possible that the adapted physical state affects the perception of motion, leading them to consider more motion as incongruent motion. In that case, the MIR might become more of a representation of the MISC and less of the actual quality of the motion. Further research in this direction is recommended.

4.6. Limitations

Participants drove open-loop (i.e., driven as passengers) throughout the whole experiment. Despite the fact that Kolff et al. (2025) showed that similar MIRs can be obtained with closed- and open-loop driving, this does not necessarily apply to MISC ratings. For real vehicles, closed-loop driving is known to induce less motion sickness (Rolnick and Lubow, 1991) compared to open-loop driving. As a consequence, closed-loop driving experiments might induce less simulator sickness than the presented results of this paper suggest. If simulator sickness is generally less present in closed-loop driving than in open-loop driving, this would require a smaller sample size, which would be a beneficial development. As part of the motivation of the present study is to support the experiment design of other, upcoming driving simulator studies, such as simulator selection, it had a focus on understanding relative differences between the tested conditions, which are likely similar in closed-loopdriving.

It should be further noted that the applied experiment design was not fully factorial, as this would have required a cueing condition with inverted motion in all three axes, where too many dropouts were feared. Scaled, missing and false motion may be differently relevant depending on the axis in which the incorrect motion is applied, while interaction effects could not be explicitly tested in the mixed effects models. Tentatively, there is no indication of the existence of such interaction effect (see Fig. 7a). The qualitative data of the MIR, however, do suggest that the negative effects of missing and false motion in the ω_z axis may be most severe.

One limitation of the continuous MISC rating task is its passivity. Some subjects might have simply forgotten to change their continuous MISC rating, as they were not explicitly asked to do so while driving, such that lower values were recorded than actually was the case. This might also explain why in some conditions (mainly the false cue variants), the post-hoc MISC ratings were slightly higher than the continuous ratings at the end of each run. In the ideal case, these ratings would be the same, as they should measure the same simulator sickness phenomenon. Another explanation for their difference might be that simulator sickness symptoms might continue to get worse after each simulation (i.e., between the extraction of both rating methods). In contrast, symptoms of simulator sickness also could have decreased due to a relief from the sickening motion, although the observed higher values of the post-hoc MISC ratings compared to the last measured continuous ratings do not support this hypothesis.

A further point of consideration is that the measured continuous ratings also contain “fast responses”, which can be attributed to feelings of discomfort. This does not have to be a specific limitation of the continuous rating method, as in fact, the same phenomenon can be present in post-hoc MISC ratings. An example of this is visible in Fig. 9J, in which the post-hoc MISC value is dominated by the last maneuver. This also shows a prime benefit of the continuous ratings, as they provide insight in how the post-hoc ratings are potentially ‘biased’ by what occurred at the end of a run. For example, in the analysis of the post-hoc MISC, it was found that manipulations in f_x induce the most simulator sickness. Considering Fig. 9J, this is not necessarily representative of how sickening the whole drive is. For the f_x channel, the post-hoc MISC value is likely largely influenced by the last maneuver, the false cue deceleration cue (‘DEC3’).

Last, the short duration of the drive may have been a real limitation of this study. With one run lasting 134 s, the drive may have been too short to evoke strong symptoms of simulator sickness, with the frequent breaks in which the post-hoc ratings were collected further mitigating symptoms. There were, however, several considerations for choosing shorter drives with moments of recovery in between. It ensured that symptoms did not get too severe, which allowed for more testing and reduced the risk of carry-over effects to subsequent conditions. Indeed, it was shown that these carry-over effects were limited, although truly eliminating these is impossible for a single-session experiment. More incongruent motion might have extended the validity of the findings, but also likely would have increased the more severe and long-lasting sickness symptoms, but potentially would also increase the risk of carry-over effects. Another consideration was to remain close to realistic experiments, which often contain many short drives, for which the experiments are intended to be used. The limited length and lack of carry-over effects might have been a reason that generally low MISC values were obtained. This might indicate that the presented results are valid in this range. More research is required for investigating the validity under more severe simulator sickness symptoms.

5. Conclusion

This chapter investigated the effect of simulator motion mismatches on ratings of simulator sickness and motion incongruence in urban driving simulations. In a driving simulator experiment, subjects were driven as passengers (“open-loop”) and experienced baseline, scaling, missing, and false cue motion mismatches in manipulations of the longitudinal (f_x), lateral (f_y), and yaw (ω_z) directions. The subjects evaluated their level of simulator sickness using the Misery Scale (MISC), both continuously during the drive as well as afterwards (post-hoc). The Motion Incongruence Rating (MIR), their subjective rating on the (in)congruence of the motion, was only extracted post-hoc.

For both the post-hoc MISC and MIR, it can be concluded that false cues produce significantly higher (i.e., worse) ratings than missing cues. The post-hoc MIRs are disproportionally higher ratings than the linear rating model of Kolff et al. (2024) suggest, requiring additional weighting on missing and false cue motion. Furthermore, scaling errors do not result in higher MISC and MIR than the baseline motion. The order found by Kolff et al. (2024) was confirmed with regard to MIRs: the reproduction of the lateral specific force motion is most important for minimizing motion incongruences, followed by the longitudinal specific force, and then the yaw rate. While there were no significant differences of the manipulated axis for simulator sickness, here the order was longitudinal specific force, yaw rate, and then lateral specific force.

The continuous MISC measurements furthermore support that the level of simulator sickness depends on the mismatches of the simulator motion. This also highlights the usefulness of this novel evaluation method to attribute the emergence of simulator sickness to explicit driving maneuvers during a simulation. Through this better understanding of the effect of mismatches on motion incongruence and simulator sickness, both can be more accurately avoided in the development and tuning of future urban driving simulations.

CRediT authorship contribution statement

Maurice Kolff: Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization; **Chantal Himmels:** Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization; **Joost Venrooij:** Writing – review & editing, Validation, Methodology, Conceptualization; **Arben Parduzi:** Writing – review & editing, Validation, Methodology, Conceptualization; **Daan M. Pool:** Writing – review & editing, Validation, Methodology, Conceptualization; **Andreas Riener:** Writing – review & editing, Validation, Methodology, Conceptualization; **Max Mulder:** Writing – review & editing, Validation, Methodology, Conceptualization.

Data availability

The data that has been used is confidential.

Acknowledgement

The first two authors contributed equally to this research. For the rest, the SDC approach was applied for the sequence of authors (sequence determines contribution).

References

Almallah, M., Hussain, Q., Reinolsmann, N., Alhajyaseen, W. K. M., 2021. Driving simulation sickness and the sense of presence: Correlation and contributing factors. *Transportation Research Part F: Traffic Psychology and Behaviour 78*, 180–193. https://doi.org/10.1016/j.trf.2021.02.005

Berthoz, A., Bles, W., Bülthoff, H. H., Correia Grácio, B. J., Feenstra, P., Filliard, N., Hühne, R., Kemeny, A., Mayrhofer, M., Mulder, M., Nusseck, H.-G., Pretto, P., Reymond, G., Schlüsselberger, R., Schwandter, J., Teufel, H. J., Vailleau, B., van Paassen, M. M., Vidal, M., Wentink, M., 2013. Motion scaling for high-performance driving simulators. *IEEE Transactions on Human-Machine Systems 43*(3), 265–276. https://doi.org/10.1109/TSMC.2013.2242885

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., White, J.-S. S., 2009. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution 24*(3), 127–135.

Bos, J., De Vries, S., Emmerik, M., Groen, E., 2010. The effect of internal and external fields of view on visually induced motion sickness. *Applied Ergonomics 41*, 516–521. https://doi.org/10.1016/j.apergo.2009.11.007

Bos, J. E., 2011. Nuancing the relationship between motion sickness and postural stability. *Displays 32*(4), 189–193. https://doi.org/10.1016/j.displa.2010.09.005

Bos, J. E., MacKinnon, S. N., Patterson, A., 2005. Motion sickness symptoms in a ship motion simulator: Effects of inside, outside, and no view. *Aviation, Space, and Environmental Medicine 76*(12), 1111–1118.

Burnham, K. P., Anderson, D. R., Burnham, K. P., Anderson, D. R., 1998. Practical use of the information-theoretic approach. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-2917-7_3

Caird, J. K., Horrey, W. J., 2011. Twelve practical and useful questions about driving simulation. *Handbook of Driving Simulation for Engineering, Medicine, and Psychology 2*, 5.1–5.18

Cleij, D., 2020. Measuring, modelling and minimizing perceived motion incongruence. Phd dissertation. Delft University of Technology. https://doi.org/10.4233/uuid:45fd3f70-2ba6-43fa-a2c4-018967bfdc88

Cleij, D., Venrooij, J., Pretto, P., Pool, D. M., Mulder, M., Bülthoff, H. H., 2018. Continuous subjective rating of perceived motion incongruence during driving simulation. *IEEE Transactions on Human-Machine Systems 48*(1), 17–29. https://doi.org/10.1109/THMS.2017.2717884

Cobb, S. V. G., Nichols, S., Ramsey, A., Wilson, J. R., 1999. Virtual reality-induced symptoms and effects (VRISE). *Presence: Teleoperators & Virtual Environments 8*(2), 169–186. https://doi.org/10.1162/105474699566152

De Winter, J., van Leeuwen, P. M., Happee, R., et al., 2012. Advantages and disadvantages of driving simulators: a discussion. In: *Proceedings of measuring behavior. Vol. 2012*, p. 8th.

Diels, C., Dugenetet, P., Brietzke, A., Pham Xuan, R., 2023. Design strategies to alleviate motion sickness in rear seat passengers - a test track study. In: *Proceedings of the IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. Bilbao, Spain, pp. 5254–5258. https://doi.org/10.1109/ITSC57777.2023.10421968

Ellensohn, F., 2020. Urban motion cueing algorithms - trajectory optimization for driving simulators. Phd dissertation. Technische Universität München.

Ellensohn, F., Venrooij, J., Schwienbacher, M., Rixen, D., 2019. Experimental evaluation of an optimization-based motion cueing algorithm. *Transportation Research Part F Traffic Psychology and Behaviour 62*, 115–125. https://doi.org/10.1016/j.trf.2018.12.004

Field, Z., Miles, J., Field, A., 2012. Discovering statistics using R. Sage Publications, Thousand Oaks, CA.

Griffin, M. J., Howarth, H. V. C., 2000. Motion sickness history questionnaire. Technical Report ISVR Technical Report No. 283. Institute of Sound and Vibration Research, University of Southampton.

Himmels, C., Venrooij, J., Gmünder, M., Riener, A., 2022. The influence of simulator and driving scenario on simulator sickness. In: *Proceedings of the driving simulation conference 2022 Europe*. Strasbourg, France, pp. 29–36.

Hogerbrug, M., Venrooij, J., Pool, D. M., Mulder, M., 2020. Simulator sickness ratings reduce with simulator motion when driven through urban environments. In: *Proceedings of the Driving Simulation Conference 2020 Europe*. Antibes, France, pp. 175–178.

Igoshina, E., Russo, F. A., Shewaga, R., Haycock, B., Keshavarz, B., 2022. The relationship between simulator sickness and driving performance in a high-fidelity simulator. *Transportation Research Part F: Traffic Psychology and Behaviour 89*, 478–487. https://doi.org/10.1016/j.trf.2022.07.015

Irmak, T., Kotian, V., Happee, R., de Winkel, K. N., Pool, D. M., 2022. Amplitude and temporal dynamics of motion sickness. *Frontiers in Systems Neuroscience 16*. https://doi.org/10.3389/fnsys.2022.866503

Irmak, T., Pool, D. M., Happee, R., 2021. Objective and subjective responses to motion sickness: the group and the individual. *Experimental Brain Research 239*(2), 515–531. https://doi.org/10.1007/s00221-020-05986-6

Irmak, T., Pool, D. M., de Winkel, K. N., Happee, R., 2023. Validating models of sensory conflict and perception for motion sickness prediction. *Biological Cybernetics 117*, 185–209. https://doi.org/10.1007/s00422-023-00959-8

Johnson, D. M., 2007. Simulator sickness research summary. Technical Report TR-HFM-121-Part-II. US Army Research Institute for the Behavioral and Social Science Ft. Rucker.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., Lilienthal, M. G., 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology 3*(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

Keshavarz, B., Ramkhalawansingh, R., Haycock, B., Shahab, S., Campos, J. L., 2018. Comparing simulator sickness in younger and older adults during simulated driving under different multisensory conditions. *Transportation Research Part F: Traffic Psychology and Behaviour 54*, 47–62. https://doi.org/10.1016/j.trf.2018.01.007

Klüver, M., Herrigel, C., Preuß, S., Schöner, H.-P., Hecht, H., 2015. Comparing the incidence of simulator sickness in five different driving simulators. In: *Proceedings of the driving simulation conference 2015 Europe*. Tübingen, Germany, p. 87-94.

Kolff, M., Venrooij, J., Arcidiacono, E., Pool, D. M., Mulder, M., 2025. Predicting motion incongruence ratings in closed- and open-loop urban driving simulation. *IEEE Transactions on Intelligent Transportation Systems 26*(1), p. 517–528. https://doi.org/10.1109/TITS.2024.3503496

Kolff, M., Venrooij, J., Schwienbacher, M., Pool, D. M., Mulder, M., 2022. Motion cueing quality comparison of driving simulators using oracle motion cueing. In: *Proceedings of the driving simulation conference 2022 Europe*. Strasbourg, France, p. 111-118.

Kolff, M., Venrooij, J., Schwienbacher, M., Pool, D. M., Mulder, M., 2023. The importance of kinematic configurations for motion control of driving simulators. In: *IEEE 26th international conference on intelligent transportation systems (ITSC)*. Bilbao, Spain, pp. 1000–1006. https://doi.org/10.1109/ITSC57777.2023.10422127

Kolff, M., Venrooij, J., Schwienbacher, M., Pool, D. M., Mulder, M., 2024. Reliability and models of subjective motion incongruence ratings in urban driving simulations. *IEEE Transactions on Human-Machine Systems 54*(6), p. 634–645. https://doi.org/10.1109/THMS.2024.3450831

Kuiper, O. X., Bos, J. E., Schmidt, E. A., Diels, C., Wolter, S., 2020. Knowing what's coming: Unpredictable motion causes more motion sickness. *Human Factors 62*(8), 1339–1348. https://doi.org/10.1177/0018720819876139

Mourant, R. R., Rengarajan, P., Cox, D., Lin, Y., Jaeger, B. K., 2007. The effect of driving environments on simulator sickness. In: *Proceedings of the human factors and ergonomics society annual meeting*, pp. 1232–1236. https://doi.org/10.1177/154193120705101838

Mourant, R. R., Thattacherry, T. R., 2000. Simulator sickness in a virtual environments driving simulator. In: *Proceedings of the human factors and ergonomics society annual meeting*. Los Angeles, CA, pp. 534–537. https://doi.org/10.1177/154193120004400513

Mulder, M., Pool, D. M., van der El, K., van Paassen, R. M. M., 2022. Neuroscience perspectives on adaptive manual control with pursuit displays. *IFAC-PapersOnLine 55*(29), 160–165. https://doi.org/10.1016/j.ifacol.2022.10.249

Parduzi, A., 2021. Bewertung der Validität von Fahrsimulatoren anhand vibro-akustischer Fahrzeugschwingungen. Phd dissertation. Technische Universität Berlin.

Pinheiro, J. C., Bates, D. M., 2000. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56 https://doi.org/10.1007/0-387-22747-4_1

Qiu, Z., Mcgill, M., Pöhlmann, K. M. T., Brewster, S. A., 2023. Manipulating the orientation of planar 2d content in VR as an implicit visual cue for mitigating passenger motion sickness. In: *Proceedings of the 15th International conference on automotive user interfaces and interactive vehicular applications*, pp. 1–10. https://doi.org/10.1145/3580585.3607157

Reason, J. T., 1978. Motion sickness adaptation: a neural mismatch model. *Journal of the Royal Society of Medicine 71*(11), 819–829. https://doi.org/10.1177/014107687807101109

Reason, J. T., Brand, J. J., 1975. Motion sickness. Academic press, Cambridge, MA.

Reuten, A., Nooij, S., Bos, J. E., Smeets, J., 2021. How feelings of unpleasantness develop during the progression of motion sickness symptoms. *Experimental Brain Research 239*(12), 3615–3624. https://doi.org/10.1007/s00221-021-06226-1

Reymond, G., Kemeny, A., 2000. Motion cueing in the renault driving simulator. *Vehicle System Dynamics 34*(4), 249–259. https://doi.org/10.1076/vesd.34.4.249.2059

Rolnick, A., Lubow, R. E., 1991. Why is the driver rarely motion sick? the role of controllability in motion sickness. *Ergonomics 34*(7), 867–879.

Salter, S., Diels, C., Herriotts, P., Kanarachos, S., Thake, D., 2019. Motion sickness in automated vehicles with forward and rearward facing seating orientations. *Applied Ergonomics 78*, 54–61. https://doi.org/10.1016/j.apergo.2019.02.001

Stassen, H. G., Johannsen, G., Moray, N., 1990. Internal representation, internal model, human performance and mental workload. *Automatica 26*(4), 811–820. https://doi.org/10.1016/S1474-6670(17)53877-X

Stratulat, A., Roussarie, V., Vercher, J., Bourdin, C., 2011. Does tilt/translation ratio affect perception of deceleration in driving simulators? *Journal of Vestibular Research: Equilibrium & Orientation 21*(3), 127–139. https://doi.org/10.3233/VES-2011-0399

Turner, M., Griffin, M., 2000. Motion sickness in public road transport: The effect of driver, route and vehicle. *Ergonomics 42*(12), 1646–1664. https://doi.org/10.1080/001401399184730

Wertheim, A. H., Bos, J. E., Krul, A. J., 2001. Predicting motion induced vomiting from subjective misery (MISC) ratings obtained in 12 experimental studies. Technical Report TNO-TM-01-A066. TNO Human Factors Research Institute.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., Freckleton, R. P., 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology 75*(5), 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

de Winkel, K. N., Irmak, T., Kotian, V., Pool, D. M., Happee, R., 2022. Relating individual motion sickness levels to subjective discomfort ratings. *Experimental Brain Research 240*, 1231-1240. https://doi.org/10.1007/s00221-022-06334-6

Zöller, I. M., 2015. Analyse des Einflusses ausgewählter Gestaltungsparameter einer Fahrsimulation auf die Fahrerverhaltensvalidität. Phd dissertation. Technische Universität Berlin.