Reliability and Models of Subjective Motion Incongruence Ratings in Urban Driving Simulations

Kolff, Maurice; Venrooij, Joost; Schwienbacher, Markus; Pool, Daan M.; Mulder, Max

DOI
10.1109/THMS.2024.3450831

Publication date
2024

Document Version
Final published version

Published in
IEEE Transactions on Human-Machine Systems

Citation (APA)
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), 634-645. https://doi.org/10.1109/THMS.2024.3450831

Reliability and Models of Subjective Motion Incongruence Ratings in Urban Driving Simulations

Maurice Kolff, Joost Venrooij, Markus Schwienbacher, Daan M. Pool, *Member, IEEE*, and Max Mulder, *Senior Member, IEEE*

***Abstract*—In moving-base driving simulators, the sensation of the inertial car motion provided by the motion system is controlled by the motion cueing algorithm (MCA). Due to the difficulty of reproducing the inertial motion in urban simulations, accurate prediction tools for subjective evaluation of the simulator's inertial motion are required. In this article, an open-loop driving experiment in an urban scenario is discussed, in which 60 participants evaluated the motion cueing through an overall rating and a continuous rating method. Three MCAs were tested that represent different levels of motion cueing quality. It is investigated under which conditions the continuous rating method provides reliable data in urban scenarios through the estimation of Cronbach's alpha and McDonald's omega. Results show that the *better* the motion cueing is rated, the *lower* the reliability of that rating data is, and the less the continuous rating and overall rating correlate. This suggests that subjective ratings for motion quality are dominated by (moments of) incongruent motion, while congruent motion is less important. Furthermore, through a forward regression approach, it is shown that participants' rating behavior can be described by a first-order low-pass filtered response to the lateral specific force mismatch (66.0%), as well as a similar response to the longitudinal specific force mismatch (34.0%). By this better understanding of the acquired ratings in urban driving simulations, including their reliability and predictability, incongruences can be more accurately targeted and reduced.**

***Index Terms*—Driving simulators, measurement reliability, motion cueing, subjective ratings, urban driving.**

I. Introduction

URBAN driving is an important use-case in driving simulation due to its high importance in vehicle development. Especially for the design of autonomous vehicles, driving in urban environments proves to be one of the most challenging use-cases. Interactions with the surroundings have a higher level of complexity [1] and the likelihood of motion sickness due to the vehicle movements increases [2], [3], [4] compared to other scenarios. Driving simulators offer a unique ability to support the development of vehicle technologies by creating safe and repeatable test conditions [5]. Many driving simulators are equipped with a motion system to recreate the inertial motion of the simulated vehicle as closely as possible through the reproduction of its specific forces and rotational rates. This conversion is performed by the motion cueing algorithm (MCA). Especially for urban driving, with its characteristic sharp curves, roundabouts, and lane changes (strong lateral motion) and frequent decelerations/accelerations (strong longitudinal motion) [6], the workspace-constrained motion system can often not (fully) reproduce the reference motion [7], such that mismatches occur. Not all mismatches are necessarily problematic, however, since some can go unnoticed by the driver [8]. Only when a driver notices a deviation between their expectation of the real vehicle motion and what they actually perceive can the simulator motion be considered *incongruent* [9]. In an urban simulation, the presence of incongruences combined with the strong visual stimuli can induce relatively high simulator sickness levels [10]. Understanding *which* mismatches lead to incongruences is therefore paramount for improving these simulations.

Evaluating the (in)congruence of motion is most commonly based on subjective evaluations obtained from drivers. Such subjective ratings provide a direct measurement of the perceived quality of the presented motion cueing. Thus, they are crucial when design choices in motion cueing have to be made for (upcoming) driving simulator experiments, such as selecting a simulator, motion cueing algorithm, and/or MCA parameters. Several different subjective rating methods exist. For example, it is possible to extract an overall rating that summarizes a single maneuver [9] or a whole drive [11], [12]. A problem with these subjective rating methods is that they can only be obtained when the motion cueing is tested by human test drivers. In practice, it is not realistic to obtain statistically relevant rating data for all possible variations of motion cueings. Furthermore, some novel MCAs in development might not even be testable in a simulator yet. Only with an understanding of the relative importance of the various mismatch channels can attempts to improve the motion cueing be performed with a focus on the most critical mismatches. Thus, *predicting* subjective ratings would be a crucial development, which requires predictive models on

Received 31 December 2022; revised 19 May 2024 and 24 July 2024; accepted 24 August 2024. Date of publication 18 September 2024; date of current version 20 November 2024. This article was recommended by Associate Editor Chen Lv. *(Corresponding author: Maurice Kolff.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by BMW Group.

Maurice Kolff is with the Department of Research and Development, BMW Group, 80807 Munich, Germany, and also with the Faculty of Aerospace Engineering, Delft University of Technology, 2629 HS Delft, Netherlands (e-mail: m.j.c.kolff@tudelft.nl).

Joost Venrooij and Markus Schwienbacher are with the Department of Research and Development, BMW Group, 80807 Munich, Germany (e-mail: joost.venrooij@bmw.de; markus.schwienbacher@bmw.de).

Daan M. Pool and Max Mulder are with the Section Control and Simulation, Delft University of Technology, 2628HS Delft, Netherlands (e-mail: d.m.pool@tudelft.nl; m.mulder@tudelft.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2024.3450831.

Digital Object Identifier 10.1109/THMS.2024.3450831

the expected subjective rating data. However, the subjective rating methods that are generally applied in simulator driving experiments (e.g., maneuver-based and overall ratings) are often not of sufficiently high resolution that they can be used for extracting predictive models. Cleij et al. [9] therefore introduced a *continuous* rating method: while being driven around, drivers continuously give a rating that aims to reflect their impression at each point in time. The method has since been used in [13] (same scenario as [9]), [14], [15], [16] (rural scenarios), [17] (rural-urban scenario), and [18] (used for predictive modeling). These continuous ratings, with their high temporal resolution, *do* allow for modeling how objective motion mismatches relate to perceived motion incongruences. Thus, they are the missing link in predicting motion cueing quality for driving simulator experiments.

Cleij et al. [9] showed that for a scenario with three basic maneuvers (braking/acceleration, cornering, and these combined), participants are generally able to successfully perform the continuous rating task and provide useful data. The latter was investigated by estimating the reliability of the data through the estimation of Cronbach's alpha, although the relationship between the (in)congruence of the motion and the associated reliability has not yet been investigated. Furthermore, Cleij et al.[9] showed that the *worst*-rated segment of the maneuver correlates most with the overall rating of that maneuver. This gives rise to the hypothesis that incongruent motion generally shapes the overall impression of drivers more than congruent motion. Whether this holds for longer drives (containing multiple maneuvers), where the worst-rated maneuver would also correlate most to the overall rating of this complete drive, is unknown, as the overall ratings could be biased through short-term memory effects, such as the serial position effect [19] or the peak-and-end-rule [20]. Finally, Cleij et al.[9] showed that their continuous rating data can be described by a moving average filter of weighted lateral and vertical specific force, as well as roll and yaw rotational rate mismatches terms. However, Ellensohn et al. [21] showed that such moving average dynamics are not sufficient to predict the ratings in a more complex and longer rural scenario. It is, thus, unknown what model structure should be used for realistic urban scenarios, and what relative weightings best describe the data, as this could be different for each scenario. Due to the strong longitudinal motions in urban driving, it can be hypothesized that these motions strongly affect the ratings, in contrast to the findings of Cleij et al. [9].

This article presents four contributions. First, it investigates whether the continuous rating method of [9] yields useful results for a realistic urban driving scenario. Second, it examines whether a general relation exists between the maximum of the continuous ratings in each maneuver and the overall ratings for a long and realistic urban scenario, in contrast to the short scenario described in [9]. Third, the relation between the ratings and their reliability is investigated through the estimation of Cronbach's alpha and McDonald's omega. The latter has shown to provide better estimates of reliability, as Cronbach's alpha is known to underestimate reliability [22]. Fourth, a predictive model is developed, which was fit on the mismatch signals; these signals are selected based on their contribution to the fit.

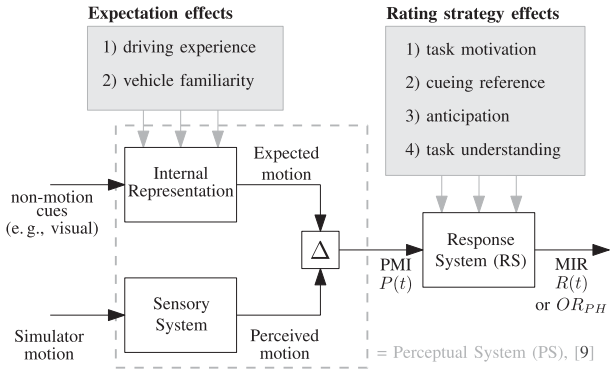

Fig. 1. Block diagram of the rating process. PMI = perceived motion incongruence, MIR = motion incongruence rating. In the present experiment, the latter is extracted using a continuous rating $R(t)$ and an overall rating $\mathrm{OR_{PH}}$.

To support these contributions, this article uses data from a driving simulator experiment in a realistic urban scenario with 60 participants, in which both continuous and overall ratings were recorded [23]. Three MCA settings were tested: 1) a classical washout algorithm *without* tilt-coordination, with large mismatches in the longitudinal and lateral specific forces, expected to provide low motion cueing quality; 2) the same algorithm *with* tilt-coordination, with smaller specific force mismatches (medium quality); 3) an optimization-based algorithm with perfect prediction capabilities, best able to reproduce the specific forces on a simulator (highest quality). This wide range of motion cueing settings allows for a better understanding of the impact of (in)congruent motion on reliability and predictability.

The rest of this article is organized as follows. Section II introduces the rating task, reliability estimates, and the modeling method. The experiment setup is explained in Section III. Results are presented in Section IV, followed by a discussion in Section V. Finally, Section VI concludes this article.

II. Methods

A. Rating Task

In the experiment, participants were driven around passively (referred to as "open-loop"), rather than driving themselves. Their task was to evaluate how well the perceived inertial motion in the simulator matched to what they would expect to feel from the simulated vehicle, i.e., their perceived motion incongruence (PMI) [24]. A block diagram of the human rating process in such tasks is shown in Fig. 1. As participants do not know exactly what the vehicle would feel like in a particular situation, they must form an internal representation [25] of the expected motion based on nonmotion cues (such as the visuals) of the simulation. This internal representation can be affected by the participant's level of experience with the task (driving) and with the vehicle that is simulated. While the simulator motion is identical for all test drivers, the *expected* motion signal can thus be different for each participant. The simulator motion is perceived through the human vestibular and proprioceptive systems, indicated as "sensory system." The internal representation and sensory system combined are indicated in [9] as the "perceptual system (PS)".

The PMI defines a participant's impression of what is (in)congruent, and would be the most useful quantity to measure. It is, however, internal to the human and not directly measurable. Instead, Cleij et al. [9] proposed to measure a subjective motion incongruence rating (MIR) that represents the PMI. The response system (RS) between the PMI and MIR can include the rating strategy, which can vary between drivers, as well as any dynamics of the rating interface. In the case of continuous ratings, the MIR is typically given through a rotary knob that can be adjusted at any time, resulting in a time signal $R(t)$. After each run, an overall rating representing the overall impression is given verbally, yielding a single rating measurement OR_{PH}. The subscript $_{\text{PH}}$ denotes posthoc, as the rating is taken after the completion of the drive. For both methods, the MIR varies between 0 and 10, in steps of 1, where 0 indicates "fully congruent motion" and 10 indicates "highly incongruent motion" [14], [15], [16], [17]. Based on earlier experience with participants, it is expected that, especially for the continuous ratings, the RS can be affected by a number of rating strategy effects.

1) *Task motivation* describes the willingness to focus (on the motion) and actively perform the (rating) task [26], [27].
2) *Cueing reference* refers to what values drivers apply for the given incongruences, which depends on which PMI-level they associate with the maximum (10) MIR score. In [9], participants were shown the full range of the incongruences before the experiment. In the present experiment, they were presented with a false cue in the training sessions to anchor to the highest MIR (10).
3) *Anticipation* can occur when incongruences of upcoming maneuvers are expected based on previous drives or from recognizing that a certain MCA setting is active.
4) *Task understanding* of the participant that only the PMI is to be evaluated, and no other motion-related phenomena (e.g., visual motion, engine sound, or vibrations).

B. Reliability

Recordings of continuous ratings over various conditions yield a collection of rating time signals $R_{cjp}(t)$, with c the condition, j the condition repetition and p the participant. If along one of these elements the average is taken, this element is taken out of the subscript, such that, for example, $R(t)$ represents the rating of the average participant across all repetitions in a given condition.

In the experiment described in the current article, each run lasted 255 s, with continuous rating data being recorded at 100 Hz ($\Delta t = 10$ ms); each recording R_{cjp} contains $N = 25,500$ samples. In psychometric theory, the *total score* is the sum of the run items $X_{cjp} = \sum_t R_{cjp}(t)$, where $\sigma^2_{X_{cp}}$ is the variance of total scores over multiple repetitions. Theoretically, if an infinite number of identical and independent repetitions were performed by a participant, the average of all total scores would result in the *true score*, i.e., the expected value of the rating: $T_{cp} = E[X_{cjp}]$. Each separate test result is bound to end up with a random, stochastic measurement error $E_{cjp} = X_{cjp} - T_{cp}$. Reliability is defined by how much of the test score variance can be explained by the true score variance [22]. As the true score cannot be determined, only estimations of a lower bound of reliability can be made. Here, the most common method (for continuous ratings [9], [17], [21]) is by determining Cronbach's alpha, which represents a reliability value for each participant p [28]

$$\alpha_{cp} = \frac{J}{J-1}\frac{\sum_j \sigma^2_{cjp}}{\sigma^2_{X_{cp}}}. \tag{1}$$

Here, J is the total number of repetitions and σ^2_{cjp} is the variance of the individual samples. The coefficient α is unbounded on the lower side, i.e., $[-\infty < \alpha \leq 1]$, where the upper bound of 1 indicates full reliability. The main assumption in the derivation of Cronbach's alpha is "tau-equivalence" [22], meaning that all repetitions of a single condition share the same true score. Due to this constraining assumption, the use of Cronbach's alpha has been criticized [22] as it can lead to underestimations of reliability. As an alternative, McDonald's omega [22], [29], as introduced in [30], is calculated as

$$\Omega_{cp} = \frac{\left(\sum_j \lambda_{cjp}\right)^2}{\left(\sum_j \lambda_{cjp}\right)^2 + \sum_j \left(1 - \lambda^2_{cjp}\right)} \tag{2}$$

where λ_{cjp} are the factor loadings. McDonald's omega is in the same range as Cronbach's alpha. As a crucial difference, however, McDonald's omega allows the variation of the true scores, i.e., does not require the assumption of tau-equivalence. This provides a more accurate estimation on reliability than Cronbach's alpha. Due to the true score variation, McDonald's omega is always equal to or higher than Cronbach's alpha [22]. The factor loadings λ_{cjp} were determined using `factoran` in MATLAB R2018b, yielding Ω_{cp} using (2).

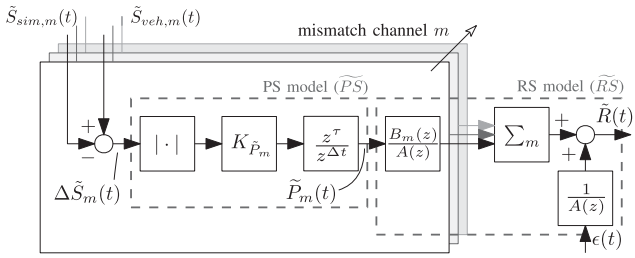

Fig. 2. Proposed human rating model structure in open-loop driving. The rectangle layers represent the various mismatch channels present in the model.

C. Predictive Model

1) Model Selection: To develop a response system model (see Fig. 2), a multiple-input-single-output (MISO) autoregressive exogenous (ARX) model is fitted. Its polynomial relationships $\frac{B_m(z)}{A(z)}$, with the discrete-time complex variable z, represent the linear transfer functions $H_m(z)$ between the measured mismatch signals $\widetilde{P}_m(t)$ (inputs) and a modeled rating signal $\tilde{R}(t)$ (output)

$$\tilde{R}(t) = \frac{1}{A(z)}\epsilon(t) + \sum_m \frac{B_m(z)}{A(z)}\widetilde{P}_m(t) \tag{3}$$

with polynomials of the form

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{n_a} z^{-n_a} \quad (4)$$

$$B_m(z) = b_{m,1} z^{-1} + b_{m,2} z^{-2} + \ldots + b_{m,n_b} z^{-n_b}. \quad (5)$$

Here, m represents the channel of the mismatch, e.g., $m \in [f_x, f_y, \ldots]$; n_a and n_b are the orders of the dynamics and $\epsilon(t)$ is the error term reflecting the noise to the system.

The signals $\tilde{P}_m(t)$ are formed by a model of the perceptual system ($\widetilde{\text{PS}}$), with the mismatches $\Delta\widetilde{S}_m(t)$ between the vehicle motion $\widetilde{S}_{\text{veh},m}(t)$ and the simulator motion $\widetilde{S}_{\text{sim},m}(t)$ as inputs. The absolute value represents that both positive and negative mismatches result in an increase of the rating value. K_m represents the gains of the mismatch signals.

To express the fit quality, the variance-accounted-for (VAF) is determined using $e(t) = R(t) - \tilde{R}(t)$

$$\text{VAF} = \left[1 - \sigma^2_{e(t)}/\sigma^2_{R(t)}\right] \cdot 100\%. \quad (6)$$

The $\sigma^2_{(\cdot)}$-terms indicate the variances. The VAF indicates how much of the variance of the difference between the modeled and measured signals can be explained by the measured signal variance [31]. A value of 100% indicates a perfect fit, whereas it is unbounded on the lower side, i.e., $[-\infty < \text{VAF} \leq 100\%]$.

To only select and include the most influential mismatch signals ranked on their contribution to the model quality of $\widetilde{\text{PS}}$, a forward regression (FR) algorithm is used [32]. At the start of the selection, the mismatch signals in the translational acceleration and jerk, as well as the rotational velocity, acceleration, and jerk are considered as possible candidates. These signals only relate to the mismatches in the inertial motion. Any mismatches in the visual motion channels (i.e., the realism of the visuals) are not explicitly considered, as the prime research motivation lies in understanding incongruences as a function of inertial motion mismatches. However, note that the definition of incongruence considers the difference between the perceived simulator motion and what the participants would *expect* to perceive in the real vehicle. In reality, this expected motion is primarily based on what the participant sees through the visuals of the simulation. Thus, this visual information is *implicitly* incorporated in the perception of motion incongruence.

Starting with an empty model, each mismatch signal is fit separately to the data. The signal that provides the highest VAF is selected. In the second iteration, all other remaining signals are tested in combination with the signal of the first iteration, selecting the second signal for the model. This process is repeated until no term provides at least an increase of 1% VAF. This method allows for testing all mismatch signals, such that only the most influential signals are included in the model, and unnecessary model complexity is avoided.

The time delay in $\widetilde{\text{PS}}$ is modeled by a term $z^{-\tau/\Delta t}$, where τ is the time delay constant. As the ARX-structure cannot estimate a time delay, the FR method is repeated for delay constants ranging between 0 and 2.5 s with steps of 0.05 s, considering the delay of 1.45 s found by [9]. The method is again repeated with orders $N = n_a = n_b$ ranging from 1 upwards until less than a 1% increase in VAF is observed.

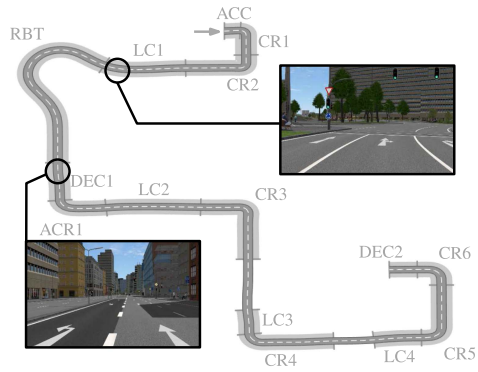

Fig. 3. Top–down view of the driven route with urban maneuvers with acceleration (ACC), corners (CR), decelerations (DEC), lane changes (LC), and a roundabout (RBT).

2) Parametric Model: The FR ARX method delivers transfer functions $H_m(z)$ in the z-domain to the most influential mismatch channels. These are converted to transfer functions $H_m(s)$ using the bilinear transformation. To obtain more flexibility in the model structure and to acquire explicit parameter values, a parametric model is fitted using the same mismatch channels as estimated by the FR method. As in the FR ARX fit, a fixed $A(z)$ term for all mismatch channels is assumed, such that all mismatch channels pass through the same rating response filter. The model is fitted in the time-domain through the minimization of a cost function

$$\arg\min_\Theta J = \sum_t \left[R(t) - \tilde{R}(t|\Theta)\right]^2 \quad (7)$$

where Θ is the parameter set. In contrast to an ARX fit, this method does not guarantee to find the global optimum. Therefore, 50 iterations are performed with uniformly distributed random numbers between 0 and 3 as initial conditions. The parameter set leading to the lowest cost is then selected.

III. Experiment Setup

A. Scenario

In the experiment [23], all participants experienced the same recording of a drive through typical urban maneuvers (see Fig. 3), consisting of lateral/yaw maneuvers [corners (CR), lane changes (LC), and roundabout (RBT)] as well as longitudinal maneuvers [accelerations (AC) and decelerations (DEC)]. As later runs might induce more anticipation effects, the driving direction (left/right arrows) was shown, together with the vehicle velocity.

B. Apparatus

BMW Group's Ruby Space simulator [see Fig. 4(a)] was used, with nine degrees of freedom. It consists of a hexapod on a tripod system, where the latter adds additional workspace in longitudinal, lateral, and yaw directions. BMW's iDrive navigation knob on the center console was used by participants to give the continuous rating [see Fig. 4(b)]. The 240° projection screen showed the visuals and a "rating bar" [9], displaying the

(a) (b)

Fig. 4. Experiment setup (photos adapted from [16]). (a) Ruby Space simulator while moving. (b) Test driver using the rating knob.

current continuous rating value. The size and color of the rating bar changed from 0 (short, white) to 10 (long, red).

It was checked at the beginning of every experiment session (i.e., for each new participant) whether the participant could comfortably and fully rotate through the rating range (0–10) with one hand movement, which all participants were able to do without problems. During the experiment, participants could rest their right arm on the center console. Feedback obtained from the participants showed that they generally found the knob easy and intuitive to operate. No comments or complaints regarding discomfort and/or difficulty operating the rating knob were made. A typical (fast) transition time of the rating recorded in the experiment required 40 ms per rating step. The rating knob was connected to the CAN bus of the simulator, which is synced with the central simulation software, together with the MCA control data and the motion system. This ensured that the recorded rating signals were always synchronized with the motion of the simulator.

C. Independent Variables

Three MCAs were tested, reflecting different levels of (expected) quality. First, a classical washout algorithm [34], [35] was used, where the vehicle motion is distributed over the hexapod (high-frequency) and tripod (low-frequency) channels, and washed-out with second-order high-pass filters. There was no tilt-coordination (NTC), such that large mismatches in the f_x and f_y channels (see Fig. 5) are present. This is expected to result in high ratings (i.e., highly incongruent), as tilt-coordination can be used to improve the cueing of sustained longitudinal and lateral specific forces, as long as the tilting rates are not noticeable [36]. Therefore, it is expected that this condition provides a reference for "low" quality.

Second, a variant of the same classical washout algorithm was used, with active tilt-coordination (CWA). Due to the tilt-coordination, the sustained specific forces in f_x and f_y cause the mismatches to be smaller. The yaw rate remains unaffected, as shown in Fig. 5(f). The tilt-coordination was tuned aggressively such that the roll rate could be noticeable [>3 deg/s [33], see Fig. 5(b)] to obtain a better reproduction of lateral specific force. The aim of this condition is to represent a state-of-the-art algorithm that can potentially be used in real-time simulations. The condition "CWA" is expected to represent "medium" quality due to two inherent limitations: as it uses linear filters, a CWA must always be tuned to account for the worst-case maneuver, limiting the simulator motion in all other maneuvers. Second, as the algorithm uses causal filters, it cannot incorporate future states in the motion cueing.

As the third condition, an optimization-based algorithm was tested [16], where the simulator motion along the complete recorded drive was optimized offline, the *Oracle* (ORC). This algorithm can only be used in open-loop simulations, but allows for the investigation of how the available simulator workspace may be fully exploited. As a result, this condition has the smallest mismatches (see Fig. 5). Therefore, this condition is expected to represent "high" quality. The rotational rates ω_x and ω_y were below the perceptual threshold (<3 deg/s).

D. Participants and Procedures

A total of 60 subjects participated (50 men, 10 women), all employees of the BMW Group with a European car driver's license B ($\mathrm{M} = 22.38$ yrs, $\mathrm{SD} = 10.16$ yrs) and an average yearly driven distance of $\mathrm{M} = 18,833$ km ($\mathrm{SD} = 13,207$ km). The average age was $\mathrm{M} = 40.1$ yrs ($\mathrm{SD} = 10.1$ yrs). 33 participants had previous experience in driving simulators. Participants provided informed consent. The experiment was approved following BMW's internal ethics review procedure.

The experiment started with two training runs, after which participants drove with either CWA, NTC, or ORC. Each condition was repeated three times, yielding nine runs. After every third run, a five-minute break was taken. Ten participants were unable to finish the experiment due to various reasons.

IV. Results

A. Differences in Ratings

For nine participants, the experiment could not be finished (eight due to simulator sickness, one due to technical problems). The data of these participants were discarded. Fig. 6 shows the continuous ratings (left) and overall ratings (right) of the remaining 51 participants. For the former, the lines indicate the mean ratings over the participants and repetitions, $R(t)$; the shaded areas represent the standard deviation. Lower ratings indicate better perceived motion cueing quality. The differences between conditions were reported in detail in [23]. Here, we summarize the main findings. Over the whole length of the drive, the ratings of ORC are always lower than those for the CWA, indicating that ORC was perceived as better. Only the corner maneuvers (CR1-6) are statistically different through the means of the ratings in those maneuvers. The NTC condition, i.e., the CWA condition without tilt-coordination, performs worst, attaining the highest ratings of the whole experiment in the roundabout. Between these latter two conditions, all maneuvers except for the first four (ACC, CR1, CR2, and LC1) are significantly different. Similarly, for the overall ratings, the ORC ($\mu = 2.44$, $\sigma = 1.04$) is rated better than CWA ($\mu = 3.89$, $\sigma = 1.79$). NTC is rated the worst with $\mu = 5.18$ and $\sigma = 2.17$, which are significantly different [23]. The order of the ratings is the same as for the continuous ratings, where the condition with the smallest mismatches in the f_x, f_y, and ω_z channels (ORC, Fig. 5), performs best, followed by CWA and lastly NTC (with the largest mismatches).

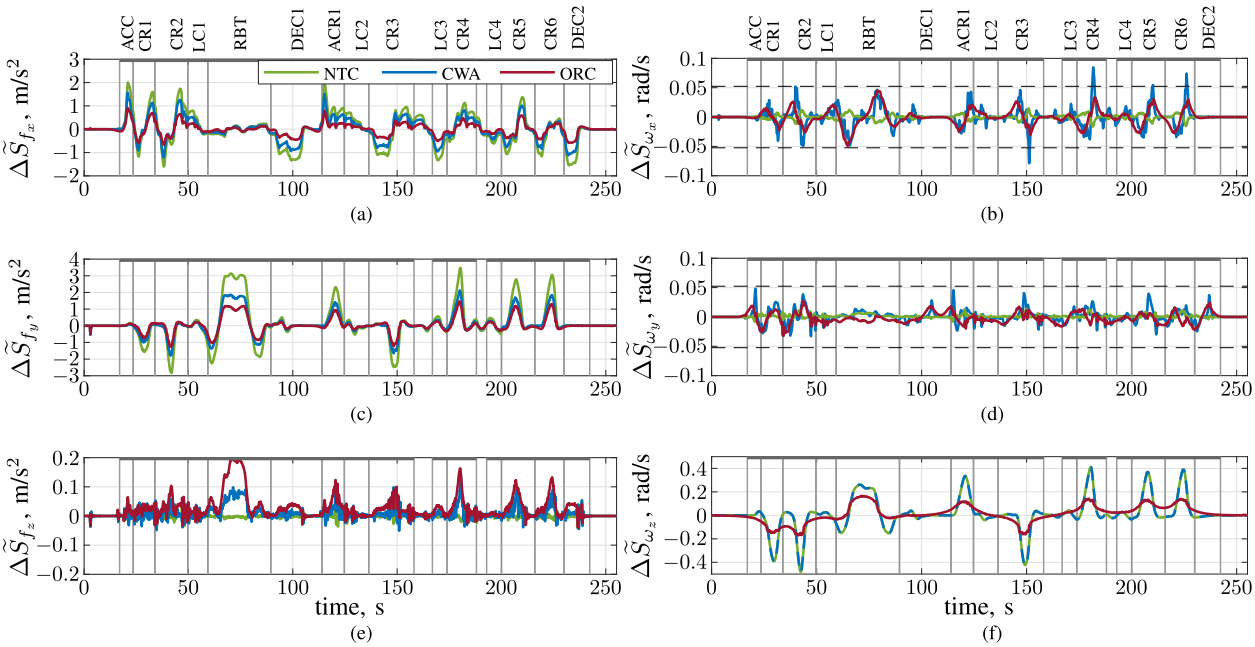

Fig. 5. Mismatches of the three algorithms (NTC, CWA, ORC). Grey vertical lines indicate the maneuvers, with the gray text entries in (a) the urban maneuvers from Fig. 3. The dashed horizontal lines in (b) and (d) indicate the rotational threshold of 3 deg/s [33], relevant for the use of tilt-coordination. (a) Longitudinal specific force. (b) Lateral specific force. (c) Vertical specific force. (d) Body roll rate. (e) Body pitch rate. (f) Body yaw rate.

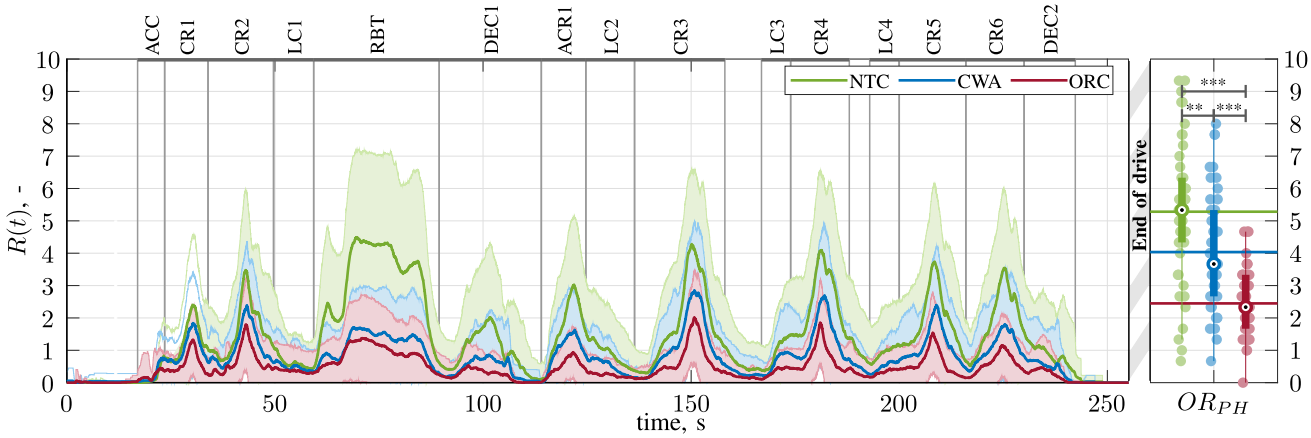

Fig. 6. Left: Averaged MIRs per MCA (as a function of time in seconds) with the standard deviation displayed as shaded areas. Right: Box plots of the three distributions of the overall ratings; their means are indicated by horizontal lines.

B. Correlation Between Continuous and Overall Ratings

To better understand the relation between the continuous ($R(t)$) and overall ratings (OR_{PH}), the Pearson correlation ρ between OR_{PH} and the maximum of $R(t)$ within each maneuver ($\hat{R}_{\text{man}}$) is calculated [see Fig. 7(a), note that the horizontal axis is sorted by the average correlation over the three conditions for increased readability], similar as in [9]. Some maneuvers correlate well (for CR6 and CR3 in NTC, $\rho = 0.88$) with the overall ratings, similar to values as found in [9]. There is a clear difference between the three conditions, where the lower rated (i.e., better) condition ORC also correlates the *least* to its own overall ratings and NTC correlates best. To further investigate the relation between the rating and the correlation, the same values of ρ are plotted as a function of the given rating in Fig. 7(b). A positive linear relationship exists between the CR and its correlation with the overall rating. The maneuver with the highest correlation, CR4, predict the overall ratings through the relationship $\text{OR}_{\text{PH}} = 2.0 + 0.8 \cdot \max[R(t)]$.

C. Reliability Estimates

Fig. 8 shows the estimated reliability coefficients for the continuous MIR data for all participants, split over the three conditions. The average reliabilities of NTC, CWA, and ORC are for Ω: 0.79, 0.68, and 0.65 and for α: 0.74, 0.62, and 0.55, respectively. The reliability values per participant are also shown as a function of the corresponding average rating in that condition (hence, the rating averaged over time and averaged over three runs). The overall trend shows that the higher the ratings,

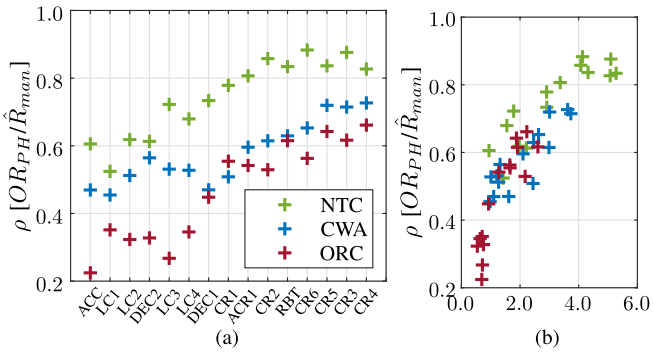

Fig. 7. Pearson correlation coefficients between the overall ratings (OR_{PH}) and the maximum of the continuous ratings within each maneuver ($\hat{R}_{man}$). (a) Per maneuver. (b) Rating power.

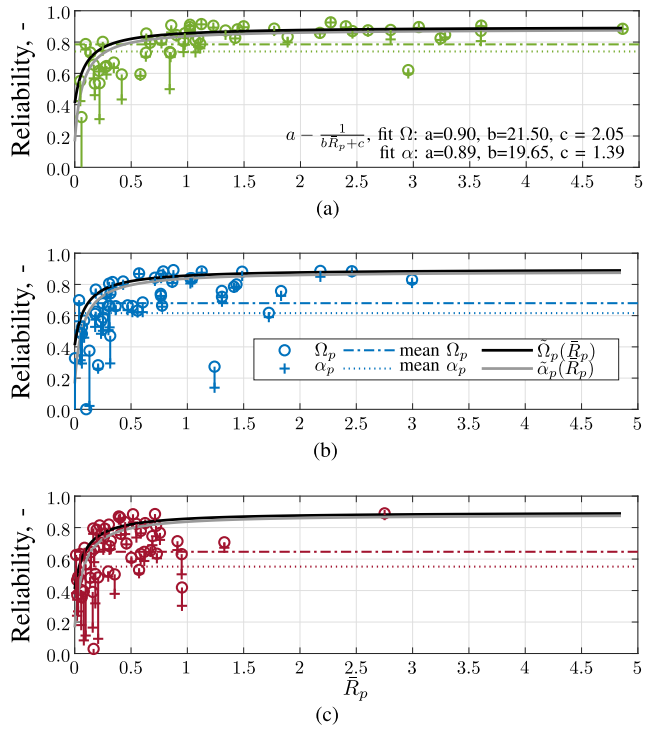

Fig. 8. Reliability coefficients α_p and Ω_p of all subjects per condition, showing that reliability decreases with lower ratings. The legend in Fig. 8(b) also applies to the same elements visible in the other subfigures. (a) NTC. (b) CWA. (c) ORC.

the more reliable the obtained data is. This again confirms our expectation that more incongruent motion results in more reliable data, and vice versa.

The figure contains both reliability metrics α and Ω, where Ω is by definition equal or higher compared to α (see Section II-B). The vertical bars show the difference between both metrics. Differences are prominent (up to 0.3) for participants for whom α is low, in line with predictions by [37]. The spread of the reliability between participants also becomes larger for smaller average ratings. It is thus at more congruent motion where the use of Ω is beneficial, as it provides a significantly higher lower bound of reliability, avoiding the false conclusion that some participants' data are unreliable at this point.

A regression of the form $r = a - 1/(b\bar{R}_p + c)$ is fit to the data, with a, b, and c the fit coefficients and $\bar{R}_p$, the average rating (over time and repetitions) per participant p. This follows the range of both α and Ω, i.e., $[-\infty < r \leq 1]$ and describes the trend of the reliability values. This function allows for *predicting* reliability based on measured ratings.

Reliability is also calculated for the overall ratings. For the continuous ratings, the presented values represent within-subject reliability. This cannot be calculated for the overall ratings, as per subject and per condition, only one data point exists. Therefore, the *between*-subject reliability is calculated, i.e., the reliability of the whole group. The values for Ω are 0.91, 0.89, and 0.73, for conditions NTC, CWA, and ORC, respectively. For the overall ratings, the values of Ω for the between-subject reliability are 0.92, 0.81, and 0.72. These values also indicate a decrease in reliability of the overall ratings, such that the decrease can be considered *inherent* to the difficulty of rating congruent motion, rather than a limitation in the continuous ratings.

D. Model Predictions

1) ARX Forward Regression: Results of the ARX FR method are shown in Fig. 9. Note that the method was applied for the ratings of the three conditions separately (referred to as models a-CWA, a-NTC, and a-ORC), as well as for all conditions grouped together in a single rating signal (a-ALL). The estimated time delay parameter τ was 0 s (a-NTC and a-ALL) and 0.05 s (a-CWA and a-ORC), independent of the model order. For $N = 1$, in all models except a-CWA, the mismatch signal $\widetilde{P}_{f_y}$ (lateral specific force) forms the most important contribution to the model followed by the longitudinal specific force mismatch $\widetilde{P}_{f_x}$. Model a-ORC contains an additional yaw rate term $\widetilde{P}_{\omega_z}$. The model fit on the CWA data (a-CWA) has a different structure: its most important term is the yaw rate mismatch $\widetilde{P}_{\omega_z}$, followed by the longitudinal specific force and yaw acceleration mismatches. Higher orders, as shown in Table I, do not provide a meaningful contribution to the model fits in terms of VAF.

To calculate the relative contributions of the most important terms to the (first-order) models, an influence factor is calculated as

$$I_m = \sum_t \widetilde{P}_m(t) / \sum_t \widetilde{P}(t) \cdot 100\% \tag{8}$$

with $\widetilde{P}_m(t) = K_m|\Delta\widetilde{S}_m(t)|$ (see Fig. 2), and m the mismatch channel. This value represents the relative contributions of the mismatches of the channels, such that the sum of all channels in the model is always 100%. This metric was introduced by [9], and thus, allows for a direct comparison to their reported values. The values are shown in Table II under "ARX FR," showing similar contributions of $\widetilde{P}_{f_y}$ and $\widetilde{P}_{f_x}$, except for a-CWA. In the latter, the $\widetilde{P}_{\omega_z}$ also provides a strong contribution at 72.0%. Note that although $\widetilde{P}_{\alpha_z}$ was included in the model a-CWA, its contribution relative to the other terms is negligible.

When repeating the process for higher orders (i.e., $N = 2$, $N = 3$), the same orders of contributions are obtained and negligible increases in VAF are observed, such that it is concluded that first-order dynamics are sufficient to explain the

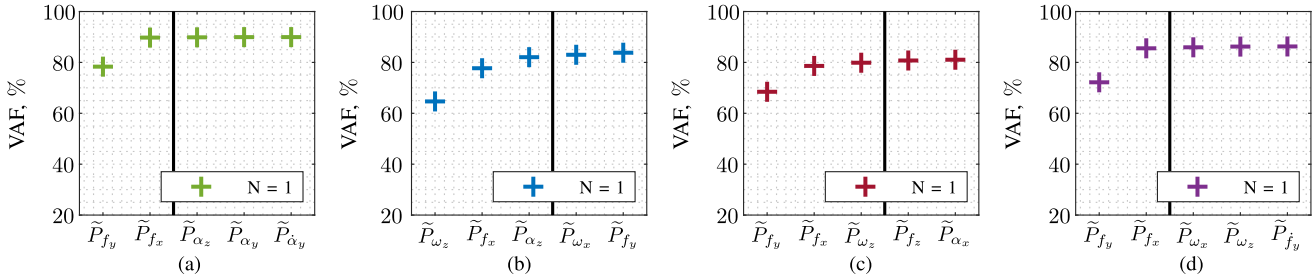

Fig. 9. VAF values of the ARX FR method, showing the consecutive contribution of the mismatch signals from left to right for the delay providing the highest VAF. The vertical bars indicate a 1% cutoff rule, such that signals that provide a lower contribution (to the right of the bar) are not considered in further analyses. (a) NTC. (b) CWA. (c) ORC. (d) ALL.

TABLE I
VAF VALUES OF THE ARX FR METHOD, SHOWING THE CONSECUTIVE CONTRIBUTION OF THE MISMATCH SIGNAL FOR THE FIRST-, SECOND-, AND THIRD-ORDER SYSTEMS

	NTC		CWA		ORC		ALL	
	$\widetilde{P}_{(\cdot)}$	VAF	$\widetilde{P}_{(\cdot)}$	VAF	$\widetilde{P}_{(\cdot)}$	VAF	$\widetilde{P}_{(\cdot)}$	VAF
$N=1$	$\boldsymbol{f_y}$	78.38%	$\boldsymbol{\omega_z}$	65.10%	$\boldsymbol{f_y}$	68.99%	$\boldsymbol{f_y}$	72.21%
	$\boldsymbol{f_x}$	89.72%	$\boldsymbol{f_x}$	78.70%	$\boldsymbol{f_x}$	78.49%	$\boldsymbol{f_x}$	87.27%
	α_z	89.92%	$\boldsymbol{\alpha_z}$	82.50%	$\boldsymbol{w_z}$	79.96%	ω_x	87.42%
	α_y	89.93%	w_x	83.42%	f_z	81.22%	w_z	87.42%
	$\dot{\alpha}_y$	89.93%	$\dot{\alpha}_x$	83.66%	α_x	81.53%	$\dot{f}_y$	87.45%
$N=2$	f_y	78.90%	w_z	65.75%	f_y	69.11%	f_y	72.46%
	f_x	89.90%	f_x	78.53%	f_x	78.52%	f_x	87.24%
	α_z	89.93%	f_y	82.56%	w_z	79.74%	f_z	87.25%
	$\dot{f}_x$	89.95%	w_x	83.46%	f_z	81.30%	w_z	87.34%
	α_y	89.96%	$\dot{\alpha}_x$	83.71%	w_x	82.03%	$\dot{f}_x$	87.35%
$N=3$	f_y	78.99%	ω_z	65.83%	f_y	69.26%	f_y	72.59%
	f_x	89.92%	f_x	78.60%	f_x	78.79%	f_x	87.24%
	$\dot{f}_x$	89.97%	f_y	82.67%	w_z	80.25%	f_z	87.28%
	α_z	90.00%	w_x	83.69%	f_z	81.75%	w_z	87.39%
	α_y	90.02%	$\dot{\alpha}_x$	83.88%	w_x	82.27%	$\dot{f}_x$	87.41%

Bold face indicates the selected model components.

TABLE II
INFLUENCE FACTORS OF THE IDENTIFIED CHANNELS FOR THE ARX FR AND PARAMETRIC MODELS, AS WELL AS REPORTED VALUES BY [9]

		$I_{\widetilde{P}_{f_x}}$ [%]	$I_{\widetilde{P}_{f_y}}$ [%]	$I_{\widetilde{P}_{f_z}}$ [%]	$I_{\widetilde{P}_{\omega_x}}$ [%]	$I_{\widetilde{P}_{\omega_y}}$ [%]	$I_{\widetilde{P}_{\omega_z}}$ [%]	$I_{\widetilde{P}_{\alpha_z}}$ [%]
ARX FR	a-NTC	25.4	74.6	-	-	-	-	-
	a-CWA	28.0	-	-	-	-	72.0	0.0
	a-ORC	13.0	76.2	-	-	-	10.9	-
	a-ALL	24.6	75.4	-	-	-	-	-
Parametric	p-NTC	31.0	69.1	-	-	-	-	-
	p-CWA	35.5	-	-	-	-	64.5	0.0
	p-ORC	24.9	63.0	-	-	-	12.2	-
	p-ALL	34.0	66.0	-	-	-	-	-
	[9]	*0*	*37*	*18*	*26*	*2*	*17*	-

Dashes indicate channels were not present in the model.

rating data and are thus used for further analysis. The bode plots in Fig. 10(a)–(h) show the estimated first-order dynamics. The responses resemble those of low-pass filters, such that participants apply smoothing to form their ratings. Furthermore, the phase responses in each model are generally equal, due to the equal $A(z)$ terms in all mismatch channels [as shown in (3)]. These phase shifts are within $0°$ and $-90°$, indicating that the low-pass filters have positive gains: An increase of the mismatches also leads to an increase in the rating. The phase responses further reveal that possibly an additional response exists at high frequencies, however, with negligible impact on the magnitude ($<10^{-3}$).

TABLE III
ESTIMATED PARAMETERS FOR THE FOUR PARAMETRIC MODELS

Model	τ [s]	ω_c [rad/s]	K_{f_x} [-]	K_{f_y} [-]	K_{ω_z} [-]	K_{α_z} [-]
p-NTC	0.00	0.33	0.89	1.66	–	–
p-CWA	0.04	0.37	0.78	–	6.71	3.71
p-ORC	0.07	0.52	0.62	1.11	1.08	–
p-ALL	0.00	0.36	0.91	1.50	–	–

2) Parametric Model: The parametric models (denoted "p-") are based on the estimated dynamics of the ARX FR method. The additional dynamics at high frequencies, as estimated by the ARX FR method, are not included, as it provided only negligible contributions to the magnitude of the estimated dynamics. In addition, as the lack of a time delay cannot be readily explained, a delay is still included in the parametric model; the model is fit in the form

$$H(s) = \sum_m K_{\widetilde{P}_m} \frac{\omega_c}{s+\omega_c} e^{-\tau s}. \tag{9}$$

Each mismatch channel has a gain $K_{\widetilde{P}_m}$, whereas ω_c is the cutoff frequency and τ the time delay constant, assumed equal in all mismatch channels. The parameter sets that describe the dynamics are $\Theta = [\tau\ \omega_c\ K_{\widetilde{P}_{f_y}}\ K_{\widetilde{P}_{f_x}}]^T$ for p-NTC and p-ALL, $\Theta = [\tau\ \omega_c\ K_{\widetilde{P}_{f_y}}\ K_{\widetilde{P}_{f_x}}\ K_{\widetilde{P}_{\omega_z}}]^T$ for model p-ORC, and $\Theta = [\tau\ \omega_c\ K_{\widetilde{P}_{\omega_z}}\ K_{\widetilde{P}_{f_x}}\ K_{\widetilde{P}_{\alpha_z}}]^T$ for model p-CWA. The resulting parameters are shown in Table III. Generally, similar values are obtained between the models, indicative of the same rating dynamics and similar weightings being applied by participants between the various conditions.

Notable is that the time delay, as in the ARX FR, is estimated as 0 s, although it is expected that humans would require a processing delay [31]. The work of [9] also found a nonzero delay of 1.45 s. Their applied model for the RS dynamics was a moving average of the form $(1+z^{-1}+z^{-2}\ldots+z^{-N_{ma}+1})/N_{ma}$ with a window N_{ma} of 300 samples (= 3 s).

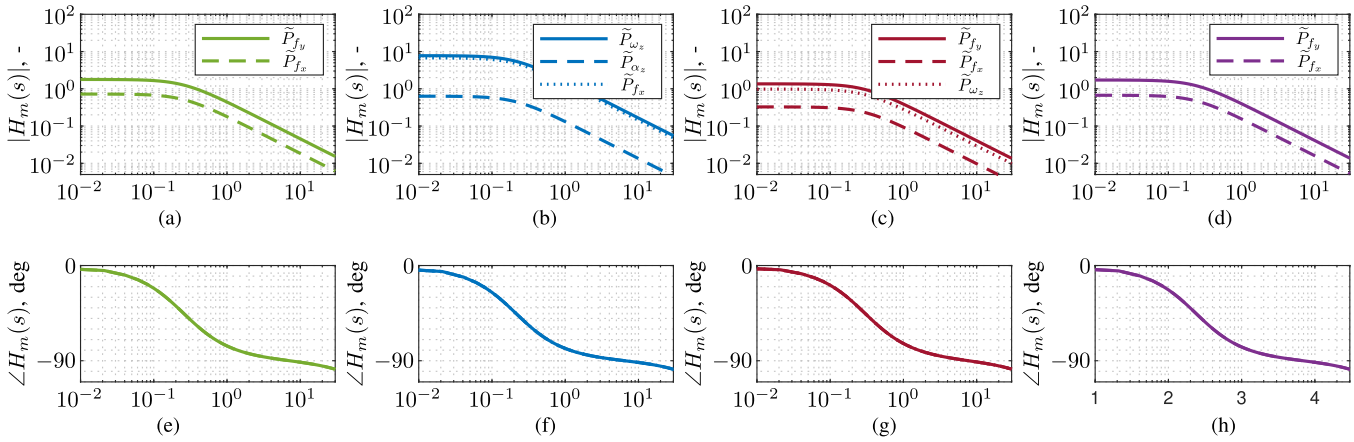

Fig. 10. Bode diagrams (magnitude (top row) and phase (bottom) of $H_m(s)$ as a function of frequency in radians/sec) of the first-order ARX FR estimations, showing low-pass filter dynamics in all mismatch channels. (a) NTC, magnitude. (b) CWA, magnitude. (c) ORC, magnitude. (d) ALL, magnitude. (e) NTC, phase. (f) CWA, phase. (g) ORC, phase. (h) ALL, phase.

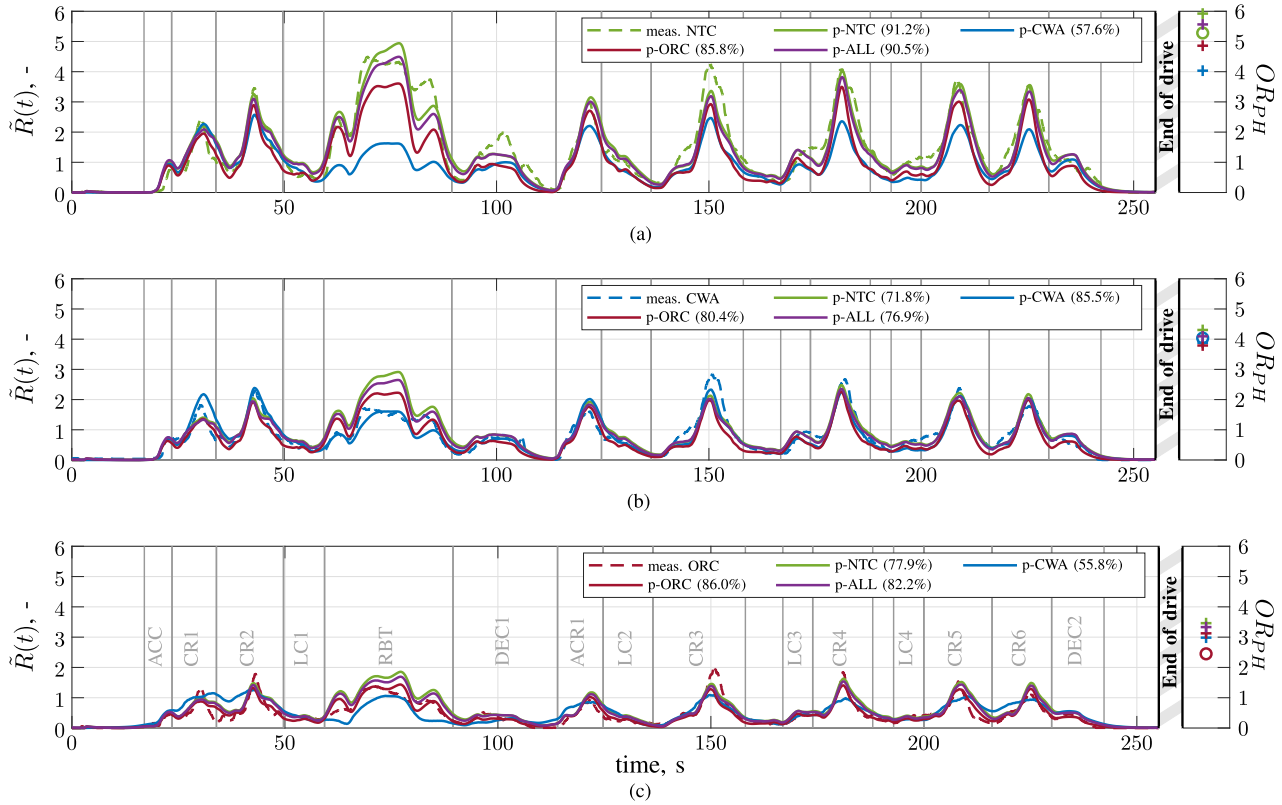

Fig. 11. Measured continuous (left) and overall (right) ratings of three conditions, each with the four applied models. Percentages in the legend indicate the VAF values for the continuous rating models. (a) NTC. (b) CWA. (c) ORC.

If the same model structure is used on our data and the delay $\tau_{ma} = N_{ma}/100$ is estimated for maximization of the cross-correlation between the ratings and moving averaged mismatch signals, a similar value of $\tau_{ma} = 1.88$ s is obtained for all conditions grouped together. This shows that although a phase shift is present between the mismatches and ratings, the phase of the estimated low-pass filter response currently captures all of the phase present in the system.

3) Model Fits and Generalizability: The model fits are shown in Fig. 11. Each figure shows the measured ratings of that condition, as well as how well the four parametric models predict the ratings in terms of VAF. Note that each condition has two models that were fit on the ratings—the respective condition and the p-ALL model. The two other conditions were fit on the other two conditions and thus provide an insight in the generalizability between the conditions. From these results, it is clear that the

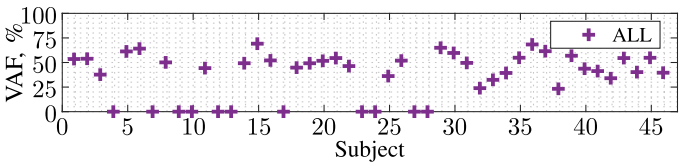

Fig. 12. Prediction quality of the "ALL" model for each subject.

model p-CWA generalizes the worst. However, for the CWA data, the p-NTC, p-ORC, and p-ALL models provide reasonable VAF values at 71.8%, 80.4%, and 76.9%, respectively. When considering all three conditions, using only two model terms, the model p-ALL explains most of the measured rating data well. Thus, a surprisingly simple model description can be used to predict the continuous rating data of all three conditions.

A notable exception is maneuver CR3, where all models underestimate the actual ratings as given by the participants. One explanation that followed from participant feedback is that this corner is specifically tight and was taken at a relatively high velocity, which might have resulted in measured ratings that are higher than the models predict.

Note that the right of the figures also includes the measured overall ratings ("$\circ$"-symbols), as well as the predicted ("+"-symbols) overall ratings, using the relation $\mathrm{OR}_{\mathrm{PH}} = 2.0 + 0.8 \cdot \max[\tilde{R}(t)]$. This again shows the generalizability of the p-ALL model, which can predict the overall ratings of all three conditions with reasonable accuracy.

4) Individual Predictions: The developed models deliver a prediction for the "average" participant. However, to form an indication on prediction power of individual ratings, Fig. 12 shows the VAF values calculated between the "ALL" model and the three datasets together. On the individual level, individual scaling differences in the rating strategy become prominent, which lead to low VAF values. In three cases, the VAF is lower than 0. Therefore, the values are manually set to a value of 0. With an average VAF of 34.5%, these values are lower than the model fits of the average rating data.

V. Discussion

The presented experiment applied the continuous rating task of Cleij et al. [9], who tested short drives each with a single maneuver, in a realistic setting: a long scenario combining a large number of maneuvers characteristic for urban driving. Overall, the 51 participants were well able to distinguish the differences in incongruences between the motion cueing conditions and rate these accordingly. Whether the rating task provides *useable* results is discussed below, in terms of how the continuous and overall ratings correspond, their reliability, and the ability to model and predict the acquired ratings.

A. Continuous and Overall Rating Correlations

Analyzing the correlation between the maximum of the continuous ratings per maneuver and the overall ratings revealed that the most *in* congruent motion dominates a participant's overall impression of the provided simulator physical motion. Recency effects, in which maneuvers occurring later in the scenario have a stronger influence on the overall rating, were not observed. This confirms the findings of [9], but also extends this finding for longer-duration and realistic urban driving scenarios containing a large number of maneuvers.

B. Reliability

Reliability estimates, mainly based on the estimation of McDonald's omega, show that the urban driving scenario is generally rated in a consistent manner with reliability levels of α $(0.6 - 0.8)$ similar as reported by [9], [14], [15], [17]. The most striking result regarding reliability estimates of the continuous MIR data is that they were found to be *inversely related* to the rating power: the lower the ratings, i.e., the better the motion is rated, generally the less reliable the ratings are. A possible explanation is that it is easier for participants to point out that something is wrong, incongruent, rather than that something is right, congruent. This also explains why the worse-rated maneuvers correlate more to the overall ratings.

This leads to a paradoxical situation, as the more one improves the simulator motion cueing, the less reliable the subjective assessment methods to confirm so become. This conclusion is independent of the choice between Cronbach's alpha and McDonald's omega. However, in continuous rating studies where reliability estimates are used as a cutoff requirement (such as in [15]), i.e., by removing data that do not meet a certain value of reliability, omega *can* be beneficial, as it is shown that for more congruent motion, the difference between omega and alpha becomes significant. Thus, it is at these points that alpha often underestimates the reliability, which can lead to the wrongful conclusion that certain rating data are unreliable. Generally, if incongruences are to be further reduced, reliability can become an issue, such that increasing the number of repetitions or deliberately inducing incongruences in the motion are required to boost reliability.

C. Model Predictions

In the model predictions, no effect of the reliability is directly observed. The four models (NTC, CWA, ORC, and ALL) provide reasonable fits and a decent level of cross-validation when predicting ratings of the other datasets. Overall, the system identification results show that the ratings can be modeled by a low-pass filter response and are dominated by the lateral specific force mismatch. In [9], a similar finding was reported, with 37% of the measured ratings attributed to this channel. In our case, this contribution ranges from 63.0% to 69.1%. The lagged response to the mismatches can likely be attributed to the rating dynamics. For example, operating the rating knob to change the rating from a 1 to a 7 requires rotating the rating knob through all in-between rating values.

Similar as in [9], a contribution of the yaw rate mismatch was found, but only in the CWA and ORC conditions. One explanation, strengthened by participants' comments, is that whereas the lateral specific force mismatches were more prominent and easier to identify, yaw motion mismatches were not. That is, the yaw rate mismatches *can* be sensed, but are secondary to the lateral specific force mismatches. NTC and CWA had

identical yaw rate mismatches, but NTC also had large lateral specific force mismatches, which therefore became dominant in the rating. For ORC, the yaw rate mismatch was smaller and might have been less noticeable, thus also resulting in a smaller contribution.

A notable difference is that an f_x term is identified between 24.9% - 35.5%, which is 0% in [9]. The obtained values are indeed realistic for an urban scenario in which strong accelerations and decelerations are present. Significant is also that other channels, such as ω_x (which was tuned above the perceptual threshold of 3 deg/s) and ω_y, did not provide a meaningful contribution, such that these were not noticeable or too short to have a meaningful impact on the rating.

When only fitting on one condition, and validating on the other, some generalizability issues are revealed. Due to its different terms and associated weightings, the CWA condition performs less in cross-validating the other two conditions. However, the ALL model (with only contributions of $\widetilde{P}_{f_y}$ and $\widetilde{P}_{f_x}$), which is fit to *all* data together, provides a reasonably good quality of the fit on all conditions and could thus be used as a general model for predicting incongruences, independent of the motion cueing architecture.

D. Future Work

1) Experiment Differences: The present experiment investigated the applicability of measuring and modeling continuous and overall ratings in a realistic urban scenario. The main motivation for this investigation is to use the gained knowledge to make predictions on the motion cueing of future driving simulation experiments. This can, for example, be used to support decision-making when selecting an appropriate simulator and motion cueing settings, and offline tuning of MCA parameters. Other, future experiments for which these evaluations are used might be performed on the exact same urban scenario, on a different urban scenario, or on a completely different scenario type (e.g., highway, rural). Therefore, it is suggested to investigate how the ratings are affected under each of these three steps.

First, if the scenario would be exactly the same, future work should investigate how ratings are affected by a different participant group and/or a different simulator or motion cueing settings. Cleij [24] showed that when two experiments expose a different range of motions (for example, by using a larger and a smaller simulator), the obtained ratings of these experiments need to be corrected for through a linear scaling factor. The next step, using a different urban scenario could explicitly investigate whether possibly the length or a different order in which maneuvers are presented affects the provided ratings. Finally, extending the results to completely different scenario types would be an important step. For example, a highway scenario might have more interaction with surrounding traffic, which could induce different types of motion (e.g., more lane changes), which might affect the balance between the mismatch channels. Furthermore, maneuvers might be harder to rate, as their occurrence might be harder to anticipate than the visually clear corner maneuvers in an urban scenario. As a result, such scenarios might inherently have a lower reliability.

2) Open-Loop Driving Experiments: A main motivation for the presented work is to leverage continuous rating prediction models, which can only be extracted from data collected in open-loop driving experiments, for predicting the quality of closed-loop driving simulator experiments. However, a central assumption so far made in the existing continuous rating literature is that open-loop ratings are also representative for closed-loop driving. However, it is possible that differences between open-loop and closed-loop driving occur due to perceptual differences [38], [39]. Thus, future work should explicitly investigate whether motion cueing in closed-loop and open-loop driving is in fact rated equivalently. Explicitly proving this would further increase the validity of the continuous rating method for closed-loop testing. An example for which the continuous rating method may be applied effectively is studying the perceived effects of masking of cues [40].

3) Error Types: In the present work, the mismatches in the motion were analyzed through objective difference functions between the vehicle reference and simulator motion. As a result, the predictive models linearly depend on the overall magnitude of the mismatch, without making any distinction between what *type* of cueing error is present. However, humans may have different sensitivities to different error types. For example, Grant and Reid [41] defined three different types of errors for flight simulation motion cueing: false cues, missing/scaling error cues, and phase-error cues. In their definition, false cue motion results in errors in the opposite direction of the true vehicle motion, or a motion cue whereas no motion is expected from the vehicle. A scaled cue is correct in its direction, but mismatched in magnitude compared to the vehicle reference motion, of which the missing cue is a special case (i.e., no simulator motion). Phase errors were also defined by [41], in which the simulator's motion is shifted in time (i.e., leading or lagging) with respect to the vehicle reference motion. Variations of these definitions exist, such as defined in [18] and [24]. Grant and Reid [41] noted that false cue motion was generally perceived as worse than scaled or missing motion, although without providing experimental proof. Following on preliminary investigations by [24], future research should investigate explicitly how these error types compare and if predictive rating models may be improved when different error types are weighted independently in the rating model.

VI. Conclusion

The difficult tradeoff and selection of motion cueing settings would greatly benefit from accurate prediction methods of subjective ratings. This article describes the application of continuous and overall motion incongruence ratings in a realistic urban driving experiment through reliability and predictability. From analyzing the correlation between the continuous and overall ratings, it is concluded that incongruent motion strongly determines the overall impression of drivers. This is explained by the reliability of the acquired continuous ratings, which is generally high, but *inversely related* to the incongruence ratings: the more congruent the presented motion is, the less the acquired ratings can be trusted. Reducing incongruent motion thus requires more effort in the subjective confirmation. This is

done either through the deliberate presentation of incongruent motion cues or by increasing the number of repetitions. For the rating data presented in this article, the reliability of the data is sufficient, as the estimates are similar to values in literature and no effects on the predictability of the rating data are observed. Nondelayed, first-order linear low-pass filtered responses to the lateral (66.0%) and longitudinal (34.0%) specific force mismatches are sufficient to predict the measured motion incongruence ratings in an urban scenario. Through this model and the gained knowledge on its associated reliability, incongruences can be more accurately targeted and reduced in the development, selection, and tuning of future motion cueing.

REFERENCES

[1] W. Zhan, C. Liu, C.-Y. Chan, and M. Tomizuka, "A non-conservatively defensive strategy for urban autonomous driving," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst.*, Rio de Janeiro, Brazil, 2016, pp. 459–464.

[2] M. Turner and M. Griffin, "Motion sickness in public road transport: The effect of driver, route and vehicle," *Ergonomics*, vol. 42, no. 12, pp. 1646–1664, 2000.

[3] S. Salter, C. Diels, P. Herriotts, S. Kanarachos, and D. Thake, "Motion sickness in automated vehicles with forward and rearward facing seating orientations," *Appl. Ergonom.*, vol. 78, pp. 54–61, 2019.

[4] T. Irmak, D. M. Pool, and R. Happee, "Objective and subjective responses to motion sickness: The group and the individual," *Exp. Brain Res.*, vol. 239, no. 2, pp. 515–531, 2021.

[5] M. Bruschetta, C. Cenedese, A. Beghi, and F. Maran, "A motion cueing algorithm with look-ahead and driver characterization: Application to vertical car dynamics," *IEEE Trans. Hum.- Mach. Syst.*, vol. 48, no. 1, pp. 6–16, Feb. 2018.

[6] M. R. C. Qazani, H. Asadi, and S. Nahavandi, "A motion cueing algorithm based on model predictive control using terminal conditions in urban driving scenario," *IEEE Syst. J.*, vol. 15, no. 1, pp. 445–453, Mar. 2021.

[7] F. Ellensohn, "Urban motion cueing algorithms - trajectory optimization for driving simulators," Ph.D. dissertation, Dept. Inst. Appl. Mechanics, Techn. Univ. Munich, München, Germany, 2020.

[8] A. Berthoz et al., "Motion scaling for high-performance driving simulators," *IEEE Trans. Hum.- Mach. Syst.*, vol. 43, no. 3, pp. 265–276, May 2013.

[9] D. Cleij, J. Venrooij, P. Pretto, D. M. Pool, M. Mulder, and H. H. Bülthoff, "Continuous subjective rating of perceived motion incongruence during driving simulation," *IEEE Trans. Hum.- Mach. Syst.*, vol. 48, no. 1, pp. 17–29, Feb. 2018.

[10] C. Himmels, J. Venrooij, M. Gmünder, and A. Riener, "The influence of simulator and driving scenario on simulator sickness," in *Proc. Driving Simul. Conf. Europe*, Strasbourg, France, 2022, pp. 29–36.

[11] P. Biemelt, S. Böhm, S. Gausemeier, and A. Trächtler, "Subjective evaluation of filter- and optimization-based motion cueing algorithms for a hybrid kinematics driving simulator," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2021, pp. 1619–1626.

[12] C. Rengifo, J.-R. Chardonnet, H. Mohellebi, and A. Kemeny, "Impact of human-centered vestibular system model for motion control in a driving simulator," *IEEE Trans. Hum.- Mach. Syst.*, vol. 51, no. 5, pp. 411–420, Oct. 2021.

[13] J. R. Van der Ploeg, D. Cleij, D. M. Pool, M. Mulder, and H. H. Bülthoff, "Sensitivity analysis of an MPC-based motion cueing algorithm for a curve driving scenario," in *Proc. Driving Simul. Conf. Europe*, Antibes, France, 2020, pp. 37–44.

[14] F. Ellensohn, M. Spannagl, S. Agabekov, J. Venrooij, M. Schwienbacher, and D. Rixen, "A hybrid motion cueing algorithm," *Control Eng. Pract.*, vol. 97, 2020, Art. no. 104342.

[15] F. Ellensohn, D. Hristakiev, M. Schwienbacher, J. Venrooij, and D. Rixen, "Evaluation of an optimization based motion cueing algorithm suitable for online application," in *Proc. Driving Simul. Conf Europe*, Strasbourg, France, 2019, pp. 93–100.

[16] F. Ellensohn, J. Venrooij, M. Schwienbacher, and D. Rixen, "Experimental evaluation of an optimization-based motion cueing algorithm," *Transp. Res. Part F. Traffic Psychol. Behav.*, vol. 62, pp. 115–125, Apr. 2019.

[17] D. Cleij et al., "Comparison between filter- and optimization-based motion cueing algorithms for driving simulation," *Transp. Res. Part F: Traffic Psychol. Behav.*, vol. 61, pp. 53–68, 2019.

[18] M. Kolff, J. Venrooij, M. Schwienbacher, D. M. Pool, and M. Mulder, "Motion cueing quality comparison of driving simulators using oracle motion cueing," in *Proc. Driving Simul. Conf. Europe*, Strasbourg, France, 2022, pp. 111–118.

[19] B. B. Murdock Jr., "Serial position effect of free recall," *J. Exp. Psychol.*, vol. 64, no. 5, pp. 482–488, 1962.

[20] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes," *J. Pers. Social Psychol.*, vol. 65, no. 1, pp. 45–55, 1993.

[21] F. Ellensohn, M. Schwienbacher, J. Venrooij, and D. Rixen, "Motion cueing algorithm for a 9 DoF driving simulator: MPC with linearized actuator constraints," *SAE Int. J. Connected Autom. Veh.*, vol. 2, no. 3, pp. 145–155, 2019.

[22] K. Sijtsma, "On the use, the misuse, and the very limited usefulness of Cronbach's Alpha," *Psychometrika*, vol. 74, no. 1, pp. 107–120, 2009.

[23] M. Kolff, J. Venrooij, M. Schwienbacher, D. M. Pool, and M. Mulder, "Quality comparison of motion cueing algorithms for urban driving simulations," in *Proc. Driving Simul. Conf. Europe*, Munich, Germany, 2021, pp. 141–148.

[24] D. Cleij, "Measuring, modelling and minimizing perceived motion incongruence," Ph.D. dissertation, Dept. Fac. Aeros. Eng., Delft Univ. Technol., Delft, Netherlands, Feb. 2020.

[25] H. G. Stassen, G. Johannsen, and N. Moray, "Internal representation, internal model, human performance and mental workload," *Automatica*, vol. 26, no. 4, pp. 811–820, 1990.

[26] D. T. McRuer and H. R. Jex, "A review of quasi-linear pilot models," *IEEE Trans. Hum. Factors Electron.*, vol. HFE-8, no. 3, pp. 231–249, Sep. 1967.

[27] M. Mulder et al., "Manual control cybernetics: State-of-the-art and current trends," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 5, pp. 468–485, Oct. 2018.

[28] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, pp. 297–334, 1951.

[29] I. Trizano-Hermosilla and J. M. Alvarado, "Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements," *Front. Psychol.*, vol. 7, 2016, Art. no. 769.

[30] R. P. McDonald, "Test theory: A unified treatment," *J. Amer. Stat. Assoc.*, vol. 95, no. 451, pp. 1012–1013, 2000.

[31] K. van der El, S. Padmos, D. M. Pool, M. M. van Paassen, and M. Mulder, "Effects of preview time in manual tracking tasks," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 5, pp. 486–495, Oct. 2018.

[32] M. Demir, N. McNeese, J. Gorman, N. Cooke, C. Myers, and D. Grimm, "Exploration of teammate trust and interaction dynamics in human-autonomy teaming," *IEEE Trans. Hum.- Mach. Syst.*, vol. 51, pp. 696–705, Dec. 2021.

[33] G. Reymond and A. Kemeny, "Motion cueing in the Renault driving simulator," *Veh. Syst. Dyn.*, vol. 34, no. 4, pp. 249–259, 2000.

[34] B. Conrad, J. G. Douvillier, and S. F. Schmidt, "Washout circuit design for multi-degrees-of-freedom moving base simulators," in *Proc. AIAA Vis. Motion Simul. Conf.*, 1973, Paper 73-929.

[35] L. D. Reid and M. A. Nahon, "Flight simulation motion-base drive algorithms: Part 1 developing and testing the equations," Inst. Aerosp. Stud., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. UTIAS 296, 1985.

[36] A. Stratulat, V. Roussarie, J. Vercher, and C. Bourdin, "Does tilt/translation ratio affect perception of deceleration in driving simulators?," *J. Vestibular Res.: Equilibrium Orientation*, vol. 21, no. 3, pp. 127–139, 2011.

[37] V. Savalei and S. P. Reise, "Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018)," *Collabra: Psychol.*, vol. 5, no. 1, 2019, Art. no. 36.

[38] A. Nesti, S. Nooij, M. Losert, H. H. Bülthoff, and P. Pretto, "Roll rate perceptual thresholds in active and passive curve driving simulation," *Simulation*, vol. 92, no. 5, pp. 417–426, 2016.

[39] A. R. Valente Pais, D. M. Pool, A. M. de Vroome, M. M. van Paassen, and M. Mulder, "Pitch motion perception thresholds during passive and active tasks," *J. Guid., Control, Dyn.*, vol. 35, no. 3, pp. 904–918, 2012.

[40] G. L. Greig, "Masking of motion cues by random motion: Comparison of human performance with a signal detection model," Ph.D. dissertation, Dept. Inst. Aeros. Stud., Univ. Toronto, Toronto, ON, Canada, 1987.

[41] P. R. Grant and L. D. Reid, "Motion washout filter tuning: Rules and requirements," *J. Aircr.*, vol. 34, no. 2, pp. 145–151, 1997.