Randomised versus fixed trial order in repeated exposure VR studies of pedestrian interaction with automated vehicles

MD SHADAB ALAM*, Eindhoven University of Technology, The Netherlands
PAVLO BAZILINSKYY, Eindhoven University of Technology, The Netherlands

Repeated exposure experiments often use the trial position to study learning, fatigue, habituation, or carryover. Such analyses are difficult to interpret if the order of conditions is fixed, because later trials may contain a different mixture of conditions from earlier trials. We examine this issue in a virtual reality study of pedestrian interaction with automated vehicles. The same experiment was available in two runs: 50 participants completed the 40 experimental trials in participant specific randomised orders, and 50 participants completed the same trials in one shared fixed sequence. We asked where this ordering difference changed the conclusions. The average trigger based responses were highly similar between the groups. Mean continuous unsafety was 0.34 in the randomised order group and 0.34 in the fixed sequence group, and the fraction of time marked unsafe was 0.34 and 0.34, respectively. No behavioural or head yaw outcome differed after correction for multiple comparisons. The ordering difference mattered more for trial position based conclusions. In the fixed sequence group, the position of the test was related to yielding, eHMI status, visibility, and distance to the pedestrian. The mixed effects models showed a corrected ordering scheme by trial position interaction for Q3, the rating of understanding of vehicle intention ($\beta = 0.33$, $q = 0.01$), while the corresponding interaction for fraction of time marked unsafe was weaker after correction ($\beta = 0.04$, $q = 5.30 \times 10^{-2}$). These effects weakened when trial position was replaced by prior exposure to the relevant condition. The results show that a reused fixed sequence may preserve broad average responses while making apparent learning effects ambiguous. For repeated exposure VR studies, participant specific randomisation or stronger counterbalancing is especially important when the research question concerns learning, expectation, carryover, or other trial by trial changes.

Additional Key Words and Phrases: Statistical analysis, Virtual reality, Repeated measures, Trial order, Randomisation, Human factors, Automated vehicles, Pedestrian behaviour

ACM Reference Format:
Md Shadab Alam and Pavlo Bazilinskyy. 2026. Randomised versus fixed trial order in repeated exposure VR studies of pedestrian interaction with automated vehicles. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Repeated measures experiments are common in human factors research because each participant can experience several experimental conditions. This design is efficient: responses can be compared within the same participant rather than only between different participants [8, 11]. The same feature also creates a methodological challenge. As participants move through a session, their responses may change due to practice, fatigue, habituation, contrast with previous trials, or carryover from previous trials [19, 20, 27]. Randomisation is one of the main safeguards against this problem, because it reduces the chance that a condition is consistently paired with a particular position in the session [7, 18].

Authors' Contact Information: Md Shadab Alam, m.s.alam@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands; Pavlo Bazilinskyy, p.bazilinskyy@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands.

Manuscript submitted to ACM

The problem becomes important when the trial position itself is analysed. In many repeated exposure studies, researchers use changes in trials as evidence of learning, adaptation, fatigue, or changing trust. Such an interpretation is only straightforward when the trial position is reasonably independent of the conditions shown at that point in the experiment. If one type of scenario appears mostly early and another mostly late, a trial position effect can combine two sources: change due to repeated exposure and change due to the changing balance of conditions [24, 26, 30]. In that case, an apparent learning curve may partly reflect the order in which the conditions were presented.

Several design strategies are available to reduce this problem. Complete randomisation and restricted randomisation aim to separate condition assignment from trial position [18, 26]. Blocking and counterbalancing can improve balance when an experiment contains many conditions or when full randomisation is undesirable [7, 24]. Latin square and Williams designs are used when order and first order carryover need to be balanced more explicitly [22, 34]. The common principle is simple: the schedule should not make a condition systematically more likely to occur early or late in the session. Statistical adjustment can help after data collection, but cannot fully restore interpretability when trial position and condition assignment are closely linked [15, 30].

VR studies of interactions between automated vehicles (AVs) and pedestrians provide a useful applied case. Pedestrians judge whether an approaching vehicle is likely to yield using cues such as speed, distance, and deceleration [14, 28]. Their judgements may also be influenced by external human machine interfaces that communicate vehicle intent [9, 21]. VR makes it possible to repeat such encounters under controlled conditions, including situations that would be difficult or unsafe to test in real traffic [4, 10, 17]. Repetition is useful, but it also means that participants can experience many similar encounters in one session. Previous AV pedestrian work has shown that repeated exposure can affect crossing behaviour, trust, perceived safety, and responses to external communication [12, 13, 23].

For this type of study, trial ordering can matter in two different ways. First, it may change average responses, such as trigger based reports of unsafety, post trial questionnaire ratings, or head orientation summaries. Second, it may change conclusions about how responses develop throughout the session. A fixed sequence can produce average responses that are similar to those from a randomised order, while still making trial position effects difficult to interpret if some conditions occur more often early or late. The present paper focuses on this distinction. Rather than asking whether randomisation is preferable in general, we quantify which conclusions changed when one reused fixed sequence was compared with participant specific randomisation in the same repeated exposure VR paradigm.

1.1 Aim of the study

This study asks how much the trial ordering scheme changed the conclusions drawn from a repeated measures VR experiment. We compared two runs of the same AV pedestrian study. In one run, each participant received a participant specific randomised order of the experimental trials. In the other run, all participants received the same fixed sequence. We examine three questions. First, did the ordering scheme change the average trigger responses, questionnaire ratings, or head yaw summaries? Second, did it change conclusions about responses about the trial position? Third, did the fixed sequence make some experimental conditions more common earlier or later in the session?

2 Method

2.1 Participants

The study received ethical approval from the Ethics Review Board of the Eindhoven University of Technology, and all participants provided their informed consent. The present analysis combines two runs of the same VR street crossing

experiment on interactions between an AV and pedestrians. The fixed sequence run was conducted first. In that run, all participants received the same sequence of the 40 experimental trials. The randomised order run was conducted later and has previously been reported by Alam et al. [1]. In that run, the same 40 experimental trials were used, but their order was randomised separately for each participant. The previous report describes the shared experimental protocol; the present paper compares the shared fixed sequence with participant specific randomisation.

In addition to the ordering scheme, the two runs used the same laboratory room, apparatus, software, virtual environment, procedure, trial structure, experimental conditions, response instructions, and outcome measures. The analysis included 100 participants with complete trial records: 50 in the randomised order group and 50 in the fixed sequence group. In the randomised order group, the mean age was 28.30 years (SD = 7.30). The gender counts were 28 male, 21 female, and 1 other. Recent VR use was reported as not in the past month by 32 participants, less than once per week by 10 and regular use by 7; one participant chose not to answer this question.

In the fixed sequence group, the mean age was 25.28 years (SD = 4.61). The gender counts were 29 male, 20 female, and 1 other. The nationalities were India (13), the Netherlands (13), China (5), Bulgaria (4), Austria (2), Italy (2), Portugal (2), and a participant each from Brazil, Canada, Indonesia, Iran, Japan, Malta, Morocco, Romania, and Somalia. Recent VR use was reported as not in the past month by 41 participants, less than once per week by 7 and regular use by 1; one participant chose not to answer this question.

2.2 Apparatus and virtual environment

Both runs used the same immersive VR setup, laboratory room, hardware, software, and virtual street crossing environment. The experiment was conducted using a Meta Quest 3[1] head mounted display and handheld controller. The virtual environment was implemented in Unity 2022.3.5f1 and depicted a straight residential street with an approaching AV and a life sized virtual co-pedestrian on the pavement. The participants remained at the kerb throughout the experiment. The behaviour data were recorded at 50 Hz, including the pose of the headset and the analogue trigger value of the controller.

2.3 Experiment design

Both runs followed the same full design within the participant. Each participant completed 42 trials in total: two practice trials followed by 40 experimental trials. These 40 experimental trials formed the analysed repeated-measures sequence and covered the full factorial design, with one trial for each condition combination. The factors within the participant were vehicle yielding with two levels (yielding, non yielding), eHMI status with two levels (off, on), pedestrian position with two levels (participant first, participant second on the vehicle path) and inter pedestrian distance with five levels (2, 4, 6, 8 and 10 m). After each experimental trial, participants provided three ratings on 0 to 100 scales: Q1, the influence of the other pedestrian on the decision to cross; Q2, the influence of the distance between pedestrians; and Q3, understanding of the intention of the vehicle.

2.4 Trial order

The two study runs differed only in the ordering of the 40 experimental trials. In the randomised order group, the 40 experimental scenarios were randomly assigned for each participant using `System.Random` C # together with a LINQ

[1] https://www.meta.com/ch/en/quest/quest-3/

shuffle. In the fixed sequence group, the same scenario order was generated once and then reused for all participants, so that each participant experienced the same trial sequence.

2.5 Procedure and measures

Both runs followed the same study procedure. Participants first read an information sheet and completed a pre-experiment questionnaire. The experimenter then explained the task and demonstrated the trigger mechanism. The participants were instructed to imagine that they intended to cross the road in front of an approaching AV and to press and hold the controller trigger whenever initiating a crossing felt unsafe. Releasing the trigger indicated that crossing would feel safe, and participants could press and release the trigger multiple times within a trial. Throughout the manuscript, we use *unsafety* for this trigger based response: a time resolved report that initiating a crossing felt unsafe. The participants remained on the pavement throughout the experiment. After each experimental trial, they completed the three trial wise questions Q1 to Q3 described above. The original randomised study included two practice trials and optional short breaks after trials 14 and 26, and the same procedure was followed here.

2.6 Data processing and outcomes

The analysis used the timestamp, analogue trigger value, and HMD pose streams recorded at 50 Hz. All trial level features were recomputed from the recorded time series for the present comparison rather than copied from the earlier study. For each trial, the timestamps and trigger values were coerced to numeric values, sorted in temporal order, and deduplicated if repeated timestamps were present. The time was then expressed in seconds relative to the first valid sample in the trial.

2.6.1 Trigger windowing and state definition. The primary trigger analyses used the part of each trial from trial onset until the approaching vehicle had passed the participant's crossing position. This endpoint was defined as the moment when the front of the vehicle, that is, the bonnet, passed the relevant pedestrian position: position 1 when the participant was the first pedestrian on the vehicle path and position 2 when the participant was second.

Within this window, the unsafe state was defined from the analogue controller trigger. Because the trigger was pressure sensitive, a sample was treated as unsafe only when the trigger value exceeded 0.05. This 5% threshold was used to avoid classifying very small trigger deflections or numerical noise as intentional trigger presses. The binary unsafe indicator in sample k was therefore $U_k = 1$ if the trigger value exceeded 0.05 and $U_k = 0$ otherwise. We use the term *unsafety* as shorthand for the participant's momentary report that initiating a crossing felt unsafe under this response rule.

2.6.2 Trigger derived trial features. The two primary behavioural outcomes were derived from the trigger signal within the crossing window. The mean continuous unsafety was defined as the average magnitude of the trigger in the analysed part of the trial. Fraction of time marked unsafe was defined as the proportion of the analysed trial window in which the trigger exceeded the unsafe threshold described above.

Several secondary trigger summaries were also extracted to check whether the ordering scheme affected other aspects of the response. These summaries described response variability, response peaks, changes in trigger pressure over time, transitions between safe and unsafe states, and the timing and duration of unsafe responses. They were used as supporting analyses rather than as primary outcomes.

2.6.3 Yaw processing. Head orientation was processed from the HMD rotation data. For each sample, the recorded headset quaternion was converted to Euler angles and the yaw component was retained. The yaw values were expressed in degrees and wrapped in the interval $[-180°, 180°]$ [4].

The extracted yaw summaries were mean yaw, mean absolute yaw, yaw standard deviation, yaw range, and the fraction of samples in which the headset was facing approximately forward. The forward-facing summary was defined as an HMD yaw within ±15° of the scene-aligned forward direction [31].

2.6.4 Primary outcomes and participant level summaries. To make the inferential hierarchy explicit, we define three primary outcomes for the main analyses: mean continuous unsafe, fraction of time marked unsafe, and Q3. The two trigger based outcomes capture the continuous behavioural response and the proportion of the trial spent in an unsafe state. Q3 was selected as the primary outcome of the questionnaire because it asks about the participant's judgement of the vehicle's intention, which is the central post trial judgement in this study. Q1 and Q2 were analysed as secondary questionnaire outcomes because they refer to specific explanatory cues, namely the presence of another pedestrian and the distance between pedestrians, rather than to the participant's overall judgement of vehicle intention.

For participant level mean comparisons, each outcome was first averaged across the experimental trials with available data, and then the two ordering groups were compared using two sample Welch tests t [33]. The control of the false discovery rate within each outcome family used the Benjamini–Hochberg procedure [5].

2.6.5 Trial position and factor drift. To assess whether the ordering scheme linked the trial position with the experimental factors, we computed Pearson correlations between the trial position and each factor separately within each ordering group. Yielding, eHMI, and visibility were coded as binary indicators, whereas inter pedestrian distance was coded using its ordered distance level. Pearson correlations were used as descriptive measures of factor drift throughout the trial sequence.

2.6.6 Primary mixed effects models. The temporal analyses tested whether the responses changed throughout the session in the same way for the randomised order and fixed sequence groups. Because each participant completed many trials, the trial observations of the same participant could not be treated as independent. We therefore used linear mixed effects models, which account for repeated observations within participants [3]. The models also allowed each participant to have their own starting level and their own trend in the trial position, because some participants may respond more strongly than others and others may change more during the session [2].

For mean continuous unsafety and Q3, the model was:

$$Y_{ij} = \beta_0 + \beta_1 \,\mathrm{Order}_i + \beta_2 \,\mathrm{TrialPosition}_{ij} + \beta_3 \,(\mathrm{Order}_i \times \mathrm{TrialPosition}_{ij}) + u_{0i} + u_{1i} \,\mathrm{TrialPosition}_{ij} + \varepsilon_{ij}. \tag{1}$$

Here, Y_{ij} is the response from participant i on trial j. Order_i identifies whether the participant belonged to the randomised order or fixed sequence group. The randomised order group was the reference group. $\mathrm{TrialPosition}_{ij}$ represents where the trial occurred in the session. The terms u_{0i} and u_{1i} represent participant specific deviations from the overall starting level and trial position trend.

The coefficient of interest was β_3. This term asks whether the fixed sequence group had a different change in the trial position compared to the randomised order group. In other words, it tests whether the two ordering schemes produced different time on task patterns.

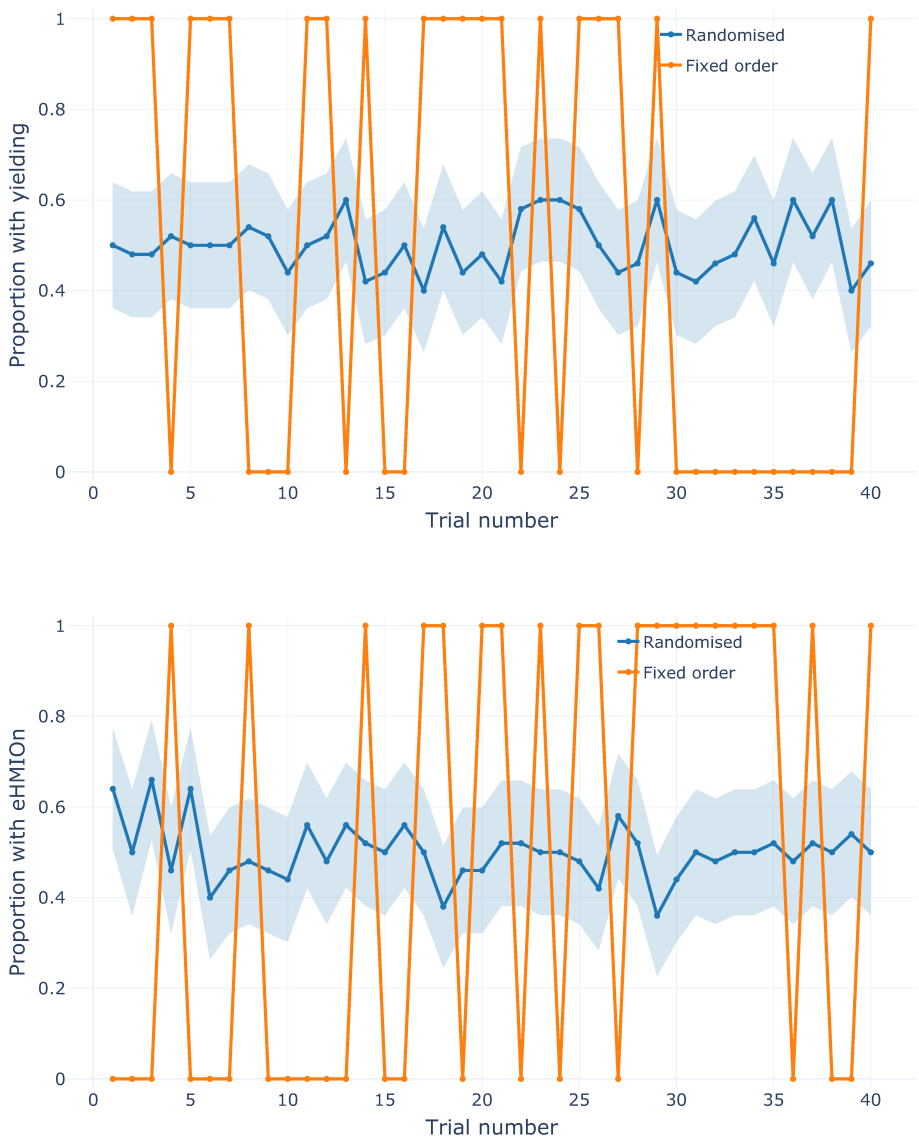

Fig. 1. Average trial composition by trial position across participants. The top panel shows the proportion of participants who encountered a yielding trial at each trial position, and the bottom panel shows the proportion who encountered an eHMI on trial at each trial position. Shaded bands indicate 95% binomial confidence intervals for the randomised order group only; the fixed sequence has no confidence band because all participants received the same trial order.

For the fraction of time marked unsafe, the same model was used after applying a logit transformation, because this outcome is a proportion and is bounded between 0 and 1 [6, 32]. Values equal to 0 or 1 were moved slightly inside

the interval before transformation so that the logit could be calculated. The transition-count analysis was treated as secondary and used a $\log(1 + x)$ transformation because the outcome was a count. For each result, we report the estimate, the 95% confidence interval, the p value, and the Benjamini–Hochberg adjusted value q for β_3.

Table 1. Participant level comparison of behavioural and yaw outcomes by ordering group.

Outcome	n_r	M_r	n_f	M_f	t	q
Maximum increase rate of unsafety	50	151.32	50	32.76	1.73	0.85
Unsafety volatility (SD of rate of change)	50	5.02	50	1.79	1.61	0.85
Fraction of time marked unsafe	50	0.34	50	0.34	-0.05	0.99
Mean continuous unsafety rating	50	0.34	50	0.34	8.00×10^{-3}	0.99
95th percentile unsafety rating	50	0.78	50	0.80	-0.38	0.99
Unsafety rating variability (SD)	50	0.34	50	0.35	-0.09	0.99
Latency to first unsafe response (s)	50	5.52	50	5.77	-0.47	0.99
Latency to first return to safe (s)	50	9.50	50	9.77	-1	0.99
Mean unsafe bout duration (s)	50	3.98	50	3.97	0.02	0.99
First press to first release interval (s)	50	4.25	50	4.30	-0.11	0.99
Yaw forward fraction (within 15°)	50	0.91	50	0.90	0.56	0.99
Mean yaw (deg)	50	1.40	50	0.34	1.25	0.99
Yaw SD (deg)	50	7.17	50	7.68	-0.67	0.99
Mean absolute yaw (deg)	50	6.16	50	6.32	-0.26	0.99
Yaw range (deg)	50	41.14	50	40.96	0.04	0.99

n_r and n_f denote the number of participants in the randomised order and fixed sequence groups. M_r and M_f are participant means across the 40 experimental trials with available data. Welch t tests compare the two ordering groups at the participant level; q values are Benjamini–Hochberg adjusted across outcomes.

2.6.7 Exposure-based and sequential models. The trial-position models tested whether the two ordering groups changed differently from early to late trials. However, trial position can mix two things. It can reflect real changes within the session, such as learning, habituation, or changing expectations. It can also reflect the order of the conditions. For example, if more yielding trials occur late in the fixed sequence, a late-session change may partly reflect increased exposure to yielding trials rather than time in the session itself.

To check this, we fitted a second set of models based on prior exposure. For each participant and each trial, we counted how many previous trials had belonged to a relevant condition category, such as eHMI in trials. We then refitted the temporal models using this prior-exposure count instead of trial position. This analysis tested whether the two ordering groups still differed after the participants were compared at similar levels of prior exposure to the relevant condition.

These exposure-based models were used as a sensitivity check. If a difference between the two ordering groups became smaller after using prior exposure, this suggested that the original trial-position effect was partly related to the sequence of conditions. If the difference remained, this suggested that it was less dependent on the specific condition order. We also fitted secondary sequential models that included the previous trial outcome, whether the yielding state changed from the previous trial, the current trial factors, and the previous trial factors. These models were used to check whether short-term carryover from the immediately preceding trial contributed to the observed pattern patterns at the trial-level.

Table 2. Participant level questionnaire ratings by ordering group.

Outcome	Randomised order	Fixed sequence	t	q	d
Q1	21.30 (17.22)	15.25 (16.91)	1.77	0.12	0.35
Q2	27.70 (18.79)	18.81 (16.13)	2.54	0.04	0.51
Q3	68.10 (17.30)	65.26 (16.33)	0.84	0.40	0.17

Values are participant means across the experimental trials, with standard deviations in parentheses. Both ordering groups had $n = 50$ participants for each outcome. Welch t tests compare the two ordering groups at the participant level; q values are Benjamini–Hochberg adjusted across Q1–Q3. Cohen's d is reported as the standardised mean difference.

Table 3. Association between trial position and experimental factors (Pearson r), by ordering group.

	Randomised order	Fixed sequence
Correlation with yielding condition	4.00×10^{-3}	-0.36
Correlation with eHMI on condition	-0.02	0.41
Correlation with visibility condition	-8.00×10^{-3}	0.25
Correlation with inter pedestrian distance level	-0.01	0.10

2.6.8 Equivalence and event-defined latency analyses. We used a targeted equivalence analysis to check whether the yielding effect was similar under the two ordering schemes. The test focused on the ordering scheme by yielding interaction from the mixed model. This term asks whether the difference between yielding and non-yielding trials changed when the experiment used one fixed sequence instead of participant specific randomisation.

Equivalence was assessed using two one sided tests. The equivalence bound was set to $\Delta = 0.20 \times \mathrm{SD}(Y)$, where Y is the dependent variable analysed on its fitted scale. This bound represents a small standardised effect and allowed the same rule to be used across outcomes with different scales. The coefficient was treated as equivalent only if it was within $[-\Delta, +\Delta]$.

The latency outcomes were analysed separately because they are event-defined measures. The latency to the first unsafe response was calculated only for trials in which the trigger exceeded the unsafe threshold. The latency to the first return to safe was calculated only for trials in which an unsafe response was followed by a return to safe state. We therefore modelled whether these latency events were observed using generalised logistic estimating equations with the participant as the clustering variable. The predictors were ordering scheme, yielding, eHMI, trial position, and the ordering scheme by trial position interaction.

3 Results

All 100 participants had complete trial logs. The randomised order and fixed sequence groups each contributed 50 participants and 2,000 experimental trial observations for trigger and questionnaire analyses.

3.1 Overall comparisons

The average response levels were very similar in the two ordering groups. The primary trigger outcomes in Table 1 showed almost no difference between the randomised order and fixed sequence groups. Fraction of time marked unsafe was 34.00×10^{-2} in the randomised order group and 34.20×10^{-2} in the fixed sequence group. The mean continuous unsafety was 33.80×10^{-2} and 33.70×10^{-2}, respectively. None of the behavioural or yaw outcomes showed a statistically reliable difference after correction for multiple comparisons.

The questionnaire ratings are reported in Table 2. Q1 and Q2 were lower in the fixed sequence group and Q3 was slightly lower. After correction in Q1–Q3, only Q2 met the adjusted threshold of $q < 0.05$. Q2 was 27.70 in the randomised order group and 18.81 in the fixed sequence group, indicating a lower rating of the influence of inter pedestrian distance in the fixed sequence group. The two primary trigger outcomes and Q3 did not show corresponding mean level differences.

Equivalence results and latency event absence are reported in Table 5 and Table 6. Equivalence for the yielding versus non yielding contrast was supported only for the number of transitions. It was not supported for mean continuous unsafe, Q3, yaw SD, or fraction of time marked unsafe. The first unsafe response was not observed in 17.30% of the trials in the randomised order group and 14.60% in the fixed sequence group. A first return to safe was not observed in 48.70% and 47.50% of the trials, respectively.

3.2 Trial order and condition structure

The randomised order group showed little association between trial position and the experimental factors (Table 3). The correlations ranged from -0.02 for eHMI on condition to 4.00×10^{-3} for yielding condition.

The fixed sequence group showed clearer changes in the balance of conditions throughout the session. Yielding trials were more common earlier in the sequence ($r = -0.36$), whereas eHMI on trials became more common later ($r = 0.41$). Visibility and inter pedestrian distance also increased across trial position ($r = 0.25$ and $r = 0.10$, respectively). Figure 1 shows these changes for yielding and eHMI status.

Condition level and factor level summaries are reported in the Appendix. They did not change the interpretation based on the primary outcomes.

3.3 Primary temporal analyses

The mixed effects results are reported in Table 4, and the corresponding raw trajectories are shown in Figure 2. The ordering scheme by interaction of the position of the test survived correction for Q3 ($b = 0.33$, $q = 9.00 \times 10^{-3}$), with the positive coefficient indicating a greater increase in Q3 in the position of the test in the fixed sequence group than in the randomised order group. The fraction of time marked unsafe and the mean continuous unsafety interactions were in the same direction but did not survive correction ($q = 5.30 \times 10^{-2}$ and $q = 0.07$, respectively). The number of transitions and the yaw SD also did not show a corrected ordering scheme by the effects of the position of the test.

The exposure based models did not retain any corrected ordering scheme by exposure interactions for mean continuous unsafety, Q3, or unsafety volatility. Therefore, the trial position difference observed for Q3 was not retained when the comparison was expressed in terms of prior exposure. The Appendix reports the carryover, reliability, break matched, ROC AUC, cross metric correlation, and latency event absence analyses.

Table 4. Ordering scheme×trial position interaction terms from the mixed effects learning model.

Outcome	b	SE	95% CI	p	q
Q3	0.33	0.10	[0.12, 0.53]	2.00×10^{-3}	9.00×10^{-3}
Fraction time unsafe	0.04	0.02	$[6.00 \times 10^{-3}, 0.08]$	0.02	5.30×10^{-2}
Mean unsafety (continuous)	2.00×10^{-3}	1.00×10^{-3}	$[< 1.00 \times 10^{-3}, 4.00 \times 10^{-3}]$	0.04	0.07
Number of transitions	-2.00×10^{-3}	2.00×10^{-3}	$[-5.00 \times 10^{-3}, 2.00 \times 10^{-3}]$	0.29	0.36
Yaw SD	0.02	0.03	[-0.03, 0.08]	0.39	0.39

Positive coefficients indicate a steeper increase over trial position in the fixed sequence group. q values are Benjamini–Hochberg adjusted across outcomes.

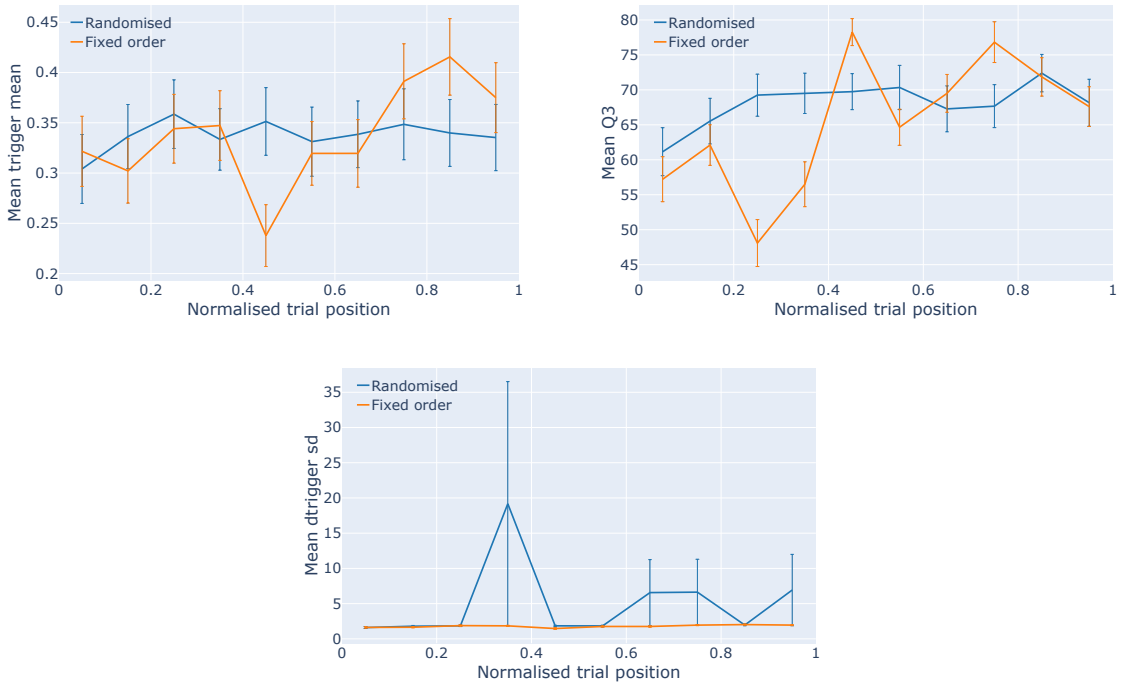

Fig. 2. Raw time on task trajectories by ordering group. Trial position is normalised from 0 to 1, where 0 indicates the start of the session and 1 indicates the end. The top left panel shows mean unsafety, the top right panel shows Q3, the rating of understanding of vehicle intention, and the bottom panel shows unsafety volatility.

4 Discussion

This study examined whether the conclusions from a repeated exposure VR pedestrian experiment changed when the same experimental trials were presented in a reused fixed sequence rather than in participant specific randomised orders. The main finding was that the influence of the ordering scheme depended on the type of conclusion to be drawn. The average trigger based outcomes were largely stable across the two ordering groups, whereas analyses based on trial position were more sensitive to the fixed sequence.

For average comparisons at the participant level, the two primary trigger outcomes were almost unchanged between the ordering groups. The mean continuous unsafe and the fraction of time marked unsafe did not show a corrected

group difference, and the same was true for the broader set of behavioural and yaw outcomes (subsection 3.1; Table 1). This indicates that, in this experiment, the reused fixed sequence did not substantially alter broad average estimates of the trigger based response. In other words, if the analysis had been limited to average response levels, the fixed sequence and the randomised order would have led to similar conclusions.

The results were different for analyses that depended on trial position. In the fixed sequence group, the trial position was associated with the experimental factors: yielding, eHMI, visibility, and inter pedestrian distance were not evenly distributed across the session (subsection 3.2; Table 3). This was not observed in the randomised order group, where the corresponding correlations were close to zero. The mixed effects models also showed a corrected ordering scheme by trial position interaction for Q3, the rating of understanding of vehicle intention (subsection 3.3; Table 4). Thus, the clearest consequence of the fixed sequence was not a shift in average unsafety, but a change in how responses appeared to vary across the session.

The exposure based analyses help to interpret this finding. When the trial position was replaced by prior exposure to the relevant experimental conditions, the ordering scheme by exposure interactions no longer survived correction. This suggests that the trial position effects in the fixed sequence should not be interpreted as pure learning, habituation, or adaptation effects. A later trial in the fixed sequence was not only later in time; it was also more likely to have a different combination of experimental conditions. The main methodological contribution of this paper is therefore to show, in a concrete repeated exposure VR study, that a fixed sequence can preserve average response estimates while weakening the interpretability of temporal effects.

The results of the questionnaire provide a related indication that the type of outcome matters. Q2 differed between the ordering groups, while Q3 and the two primary trigger based outcomes did not show the corresponding mean level differences (subsection 3.1; Table 2). Because Q2 asked about the influence of inter pedestrian distance, this result may reflect a difference in how participants evaluated a specific scene factor rather than a difference in their average trigger based response. This distinction is important because trial order may affect retrospective ratings and continuous trigger measures in different ways.

The latency results require a more restricted interpretation. The first press latency is defined only when a participant gives an unsafe response, and the first release latency is defined only when the participant later returns to the safe state. A substantial proportion of the trials did not contain these events, particularly the first return to safe events (Table 6). These latency measures are therefore useful for describing the timing of observed trigger changes, but they are less general than outcomes that are defined for every trial. They should not be treated as the main basis for conclusions about ordering effects.

These findings have implications for the design and analysis of repeated exposure VR and human factors studies. When the research question concerns broad average response levels, a fixed sequence may sometimes produce estimates similar to participant specific randomisation, although this should still be checked empirically. When the research question concerns learning, habituation, expectation, carryover, or other changes in the session, the trial schedule becomes part of the inference. In such cases, participant specific randomisation, restricted randomisation, or stronger counterbalancing is needed to reduce the dependence between trial position and condition exposure.

The substantive behavioural findings are secondary to this methodological point. The broad trigger based responses were stable, and the head orientation summaries did not drive the main conclusions. The central result is that the same fixed sequence had little effect on average unsafety estimates but made trial position effects harder to interpret. This distinction is important for repeated measures VR studies in which many similar trials are presented within a single session, and apparent changes over time may partly reflect the structure of the trial sequence.

5 Limitations and future studies

The study provides comparative methodological evidence, not a direct causal test of randomisation. The two runs were closely matched in apparatus, procedure, virtual environment, and outcome measures (subsection 2.1; subsection 2.2; subsection 2.5), but they were collected from different participant samples. Residual differences between the two groups therefore cannot be ruled out. In addition, the fixed sequence condition was represented by one shared order. The observed factor drift and carryover patterns should therefore be understood as properties of that particular schedule, not as estimates of all possible fixed sequence designs. A stronger next step would be a prospective comparison in which several order control strategies are tested within the same data collection, such as participant specific randomisation, restricted randomisation, blocked randomisation, Latin square designs, or Williams type designs [26, 30].

The scope of the conclusions should also be kept narrow. Similar participant level means do not imply that fixed and randomised ordering are generally interchangeable. The targeted equivalence analysis addressed only the stability of the yielding contrast, which is a central contrast in AV pedestrian interaction research (method in subsubsection 2.6.8; results in Table 5). Even for that narrow question, equivalence was supported only for the number of transitions and not for mean continuous unsafety, Q3, yaw SD, or fraction time unsafe. Future work would benefit from prospectively specified smallest effects of interest, formal equivalence tests for the principal outcomes, and sample size planning that includes temporal effects as well as mean differences [25, 26].

A measurement related constraint concerns the scope of the derived outcomes. The latency measures are conditional on the relevant trigger event being observed (subsubsection 2.6.2). They describe when an unsafe response began or when it returned to safe among trials in which those events occurred; they should not be interpreted as general summaries of every trial. The event absence rates in Table 6 therefore indicate the limits of the latency summaries, not missing trigger recordings. The trigger summaries themselves are also one operationalisation of perceived unsafety; they do not capture all aspects of pedestrian appraisal or action preparation. The head yaw provides only a coarse description of the head orientation (subsubsection 2.6.3). Without eye tracking, it cannot identify fixed targets, visual search sequences, or covert attention. Future studies would benefit from richer multimodal measurement, including eye tracking, gaze event coding, locomotor or postural measures, and explicit crossing decisions [17].

Finally, the conclusions are strongest for repeated measures VR street crossing studies with a similar exposure structure and a relatively sparse social scene. Order effects may differ in studies with denser pedestrian contexts, multiple vehicles, more dynamic pedestrian behaviour, different eHMI reliabilities, longer exposure schedules, or more complex traffic environments. Future work should test whether the same pattern holds across a wider range of AV pedestrian paradigms [9, 28, 29]. It would also be useful to evaluate candidate trial schedules before data collection by checking factor drift, carryover balance, and cumulative exposure balance, the same diagnostics used in this comparison (subsubsection 2.6.5; subsubsection 2.6.7), so that problematic schedules can be rejected before participants are tested [15, 24].

6 Supplementary material

In line with current open science practices and recommendations for transparency in automotive user research [16], the authors openly provide these research artefacts to support reproducibility, collaboration and further advancements in the field. The materials used in the study, the analysis code and anonymised responses of the participants are available at https://www.dropbox.com. The maintained versions of the analysis code and the VR environment are available at https://github.com/Shaadalam9/multiped-learning.

References

[1] Md Shadab Alam, Debargha Dey, Marieke H Martens, and Pavlo Bazilinskyy. 2025. You'll Never Walk Alone: Inter-Pedestrian Distance, eHMIs, and Crossing Decisions in Virtual Reality. *Available at SSRN 5987383* (2025). doi:10.2139/ssrn.5987383

[2] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68, 3 (2013), 255–278. doi:10.1016/j.jml.2012.11.001

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67 (2015), 1–48. doi:10.18637/jss.v067.i01

[4] Pavlo Bazilinskyy, Md Shadab Alam, and Roberto Merino-Martınez. 2025. Pedestrian crossing behaviour in front of electric vehicles emitting synthetic sounds: A virtual reality experiment. In *Proceedings of 54th International Congress Exposition on Noise Control Engineering (INTER-NOISE)*. São Paulo, Brazil. doi:10.3397/IN_2025_1076086

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

[6] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24, 3 (2009), 127–135. doi:10.1016/j.tree.2008.10.008

[7] George E. P. Box, J. Stuart Hunter, and William G. Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery* (2 ed.). Wiley Series in Probability and Statistics, Vol. 559. Wiley-Interscience, Hoboken, NJ.

[8] Markus Brauer and John J Curtin. 2018. Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological methods* 23, 3 (2018), 389. doi:10.1037/met0000159

[9] Sarah Brill, William Payre, Ashim Debnath, Ben Horan, and Stewart Birrell. 2023. External human–machine interfaces for automated vehicles in shared spaces: A review of the human–computer interaction literature. *Sensors* 23, 9 (2023), 4454. doi:10.3390/s23094454

[10] Fanta Camara, Patrick Dickinson, and Charles Fox. 2021. Evaluating pedestrian interaction preferences with a game theoretic autonomous vehicle in virtual reality. *Transportation research part F: traffic psychology and behaviour* 78 (2021), 410–423. doi:10.1016/j.trf.2021.02.017

[11] Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization* 81, 1 (2012), 1–8. doi:10.1016/j.jebo.2011.08.009

[12] Mark Colley, Elvedin Bajrovic, and Enrico Rukzio. 2022. Effects of pedestrian behavior, time pressure, and repeated exposure on crossing decisions in front of automated vehicles equipped with external communication. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–11. doi:10.1145/3491102.3517571

[13] Mark Colley, Daniel Kornmüller, Debargha Dey, Wendy Ju, and Enrico Rukzio. 2024. Longitudinal effects of external communication of automated vehicles in the usa and Germany: A comparative study in virtual reality and via a browser. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–33. doi:10.1145/3699778

[14] Koen De Clercq, Andre Dietrich, Juan Pablo Núñez Velasco, Joost De Winter, and Riender Happee. 2019. External human-machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. *Human factors* 61, 8 (2019), 1353–1370. doi:10.1177/0018720819836343

[15] JCF De Winter and D Dodou. 2021. Pitfalls of statistical methods in traffic psychology. *International encyclopedia of transportation* (2021), 87–95. doi:10.1016/B978-0-08-102671-7.10665-7

[16] Patrick Ebel, Pavlo Bazilinskyy, Mark Colley, Courtney Michael Goodridge, Philipp Hock, Christian P Janssen, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, and Philipp Wintersberger. 2024. Changing lanes toward open science: Openness and transparency in automotive user research. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 94–105. doi:10.1145/3640792.3675730

[17] Yan Feng, Dorine Duives, Winnie Daamen, and Serge Hoogendoorn. 2021. Data collection methods for studying pedestrian behaviour: A systematic review. *Building and Environment* 187 (2021), 107329. doi:10.1016/j.buildenv.2020.107329

[18] Ronald A. Fisher. 1966. *The Design of Experiments* (8 ed.). Oliver and Boyd, Edinburgh and London. xv+248 pages.

[19] GH Freeman. 1991. Designing Experiments and Analysing Data: A Model Comparison Perspective. doi:10.4324/9781315642956

[20] Anthony G. Greenwald. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83, 2 (1976), 314. doi:10.1037/0033-2909.83.2.314

[21] Azra Habibovic, Victor Malmsten Lundgren, Jonas Andersson, Maria Klingegård, Tobias Lagström, Anna Sirkka, Johan Fagerlönn, Claes Edgren, Rikard Fredriksson, Stas Krupenia, et al. 2018. Communicating intent of automated vehicles to pedestrians. *Frontiers in psychology* 9 (2018), 284756. doi:10.3389/fpsyg.2018.01336

[22] Byron Jones and Michael G. Kenward. 2003. *Design and Analysis of Cross-Over Trials* (2 ed.). Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Vol. 98. Chapman & Hall/CRC, New York. 408 pages. doi:10.1201/9781420036091

[23] Anees Ahamed Kaleefathullah, Natasha Merat, Yee Mun Lee, Yke Bauke Eisma, Ruth Madigan, Jorge Garcia, and Joost de Winter. 2022. External human–machine interfaces can be misleading: An examination of trust development and misuse in a CAVE-based pedestrian simulation environment. *Human factors* 64, 6 (2022), 1070–1085. doi:10.1177/0018720820970751

[24] Geoffrey Keppel and Thomas D. Wickens. 2004. *Design and Analysis: A Researcher's Handbook* (4 ed.). Pearson Prentice Hall, Upper Saddle River, NJ.

[25] Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science* 8, 4 (2017), 355–362. doi:10.1177/1948550617697177

[26] Douglas C. Montgomery. 2017. *Design and Analysis of Experiments* (9 ed.). John Wiley & Sons, Inc., Hoboken, NJ.

[27] EC Poulton. 1973. Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin* 80, 2 (1973), 113. doi:10.1037/h0034731

[28] Amir Rasouli and John K Tsotsos. 2019. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE transactions on intelligent transportation systems* 21, 3 (2019), 900–918. doi:10.1109/TITS.2019.2901817

[29] Alexandros Rouchitsas and Håkan Alm. 2019. External human–machine interfaces for autonomous vehicle-to-pedestrian communication: A review of empirical work. *Frontiers in psychology* 10 (2019), 2757. doi:10.3389/fpsyg.2019.02757

[30] Stephen S. Senn. 2002. *Cross-over Trials in Clinical Research* (2 ed.). John Wiley & Sons, Chichester, UK. 364 pages. doi:10.1002/0470854596

[31] Walter J Talamonti Jr, Wenyan Huang, Louis Tijerina, and Dev Kochhar. 2013. Eye glance and head turn correspondence during secondary task performance in simulator driving. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 57. SAGE Publications Sage CA: Los Angeles, CA, 1968–1972. doi:10.1177/1541931213571439

[32] David I Warton and Francis KC Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92, 1 (2011), 3–10. doi:10.1890/10-0340.1

[33] Bernard L Welch. 1947. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika* 34, 1-2 (1947), 28–35. doi:10.2307/2332510

[34] Evan James Williams. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research Series A: Physical Sciences* 2, 2 (1949), 149–168. doi:10.1071/CH9490149

Appendix

Additional analyses

This appendix contains analyses that complement the main results. Each block below states the purpose of the analysis before the corresponding table or figure. The analyses are not part of the primary inferential hierarchy, but they help to document the robustness of the conclusions and the behaviour of secondary outcomes.

Equivalence of the yielding contrast. Table 5 reports the targeted equivalence test for the ordering scheme×yielding term. This analysis asks whether the estimated yielding effect was practically similar in the randomised order and fixed sequence groups. It should not be read as a test that the two ordering schemes were equivalent for all outcomes.

Table 5. Targeted equivalence testing (TOST) for the ordering scheme×yielding term. This analysis addresses whether the estimated yielding contrast was practically similar across ordering groups; it is not a test of overall interchangeability between ordering schemes.

Outcome	b	95% CI	$\pm\Delta$	p_{TOST}	q	Equivalent
Number of transitions	-0.01	[-0.21, 0.19]	0.25	9.00×10^{-3}	4.70×10^{-2}	Yes
Mean unsafety (continuous)	0.10	[0.02, 0.17]	0.05	0.86	0.95	No
Q3	15.84	[3.83, 27.85]	6.01	0.95	0.95	No
Yaw SD	-1.25	[-4.44, 1.95]	1.39	0.47	0.95	No
Fraction time unsafe	1.30	[-0.55, 3.16]	0.05	0.91	0.95	No

Event-defined latency outcomes. Table 6 summarises trials in which a latency value was not defined because the corresponding trigger event was not observed. These cases are not missing trigger recordings. Rather, they indicate that the participant did not give a first unsafe response, or gave an unsafe response without a subsequent return to the safe state within the analysed window.

Table 6. Latency event absence descriptives and participant-clustered logistic models (GEE). No first unsafe response indicates that the trigger never exceeded the unsafe threshold during the analysed trial window. No first return to safe after unsafe response indicates that an unsafe response occurred but no subsequent return to the safe state was observed. The all-trial descriptive value for no first return to safe includes both trials without an unsafe response and trials with an unsafe response but no subsequent return to safe. Odds ratios (OR) are reported with 95% confidence intervals.

Outcome	Term	OR	p	95% CI
No first unsafe response	Ordering scheme (fixed sequence vs randomised order)	0.50	0.23	[0.16, 1.56]
No first unsafe response	Yielding (1 vs 0)	2.26	0.02	[1.14, 4.46]
No first unsafe response	eHMI on (1 vs 0)	1.21	0.23	[0.89, 1.65]
No first unsafe response	Trial position	0.99	0.35	[0.98, 1.01]
No first unsafe response	Ordering scheme × trial position	0.99	0.38	[0.97, 1.01]
No first return to safe after unsafe response	Ordering scheme (fixed sequence vs randomised order)	0.94	0.89	[0.41, 2.20]
No first return to safe after unsafe response	Yielding (1 vs 0)	0.05	6.01×10^{-11}	[0.02, 0.12]
No first return to safe after unsafe response	eHMI on (1 vs 0)	1.27	0.04	[1.01, 1.59]
No first return to safe after unsafe response	Trial position	0.99	0.22	[0.98, 1.01]
No first return to safe after unsafe response	Ordering scheme × trial position	1.00	0.95	[0.98, 1.02]
Randomised order	No first unsafe response / no first return to safe	0.17 / 0.49		
Randomised order	No first return to safe after unsafe response	0.38		
Fixed sequence	No first unsafe response / no first return to safe	0.15 / 0.47		
Fixed sequence	No first return to safe after unsafe response	0.39		

Figure 3 gives the same event absence information by trial position. It is included to show whether the absence of latency defining events followed a clear time on task pattern in either ordering group.

Fig. 3. Trial-position profiles of latency event absence by ordering group. The left panel shows the mean probability that no first unsafe response was observed at each trial position, and the right panel shows the mean probability that no first return to safe was observed. Shaded bands indicate 95% binomial confidence intervals across participants.

Learning, drift, and carryover summaries. Table 7 reports the participant level learning, drift, and carryover summaries that differed between ordering groups after correction. These outcomes provide additional context for the main temporal analysis by showing which trial by trial summaries were most sensitive to the ordering scheme.

Table 7. Participant-level learning and sequential metrics differing between randomised order and fixed sequence groups ($q < .05$).

Outcome	n_r	M_r (SD)	n_f	M_f (SD)	t	q	d
Carryover of previous trial camera on mean unsafety	50	-0.01 (0.04)	50	0.04 (0.04)	-5.91	2.03×10^{-6}	-1.18
Early to late drift in Q3 (late minus early)	50	3.80 (15.74)	50	16.51 (15.31)	-4.09	2.00×10^{-3}	-0.82
Carryover of previous trial yielding on Q3	50	1.10 (7.75)	50	-4.64 (7.40)	3.79	3.00×10^{-3}	0.76
Carryover of previous trial pedestrian distance on Q3	50	0.27 (2.49)	50	-1.40 (2.36)	3.43	9.00×10^{-3}	0.69
Linear slope of Q3 across trial position	49	0.15 (0.58)	50	0.47 (0.47)	-3.04	0.03	-0.61
Carryover of previous trial yielding on yaw SD	50	0.47 (1.58)	50	-0.61 (2.27)	2.77	0.04	0.55
Carryover of previous trial yielding on mean unsafety	50	-3.00×10^{-3} (0.05)	50	-0.03 (0.04)	2.75	0.04	0.55
Early to late drift in number of transitions	50	0.16 (0.65)	50	-0.15 (0.49)	2.68	0.04	0.54

Figure 4 visualises selected learning and carryover summaries from Table 7. The plots show the distribution of participant estimates in each ordering group, rather than only the group means.

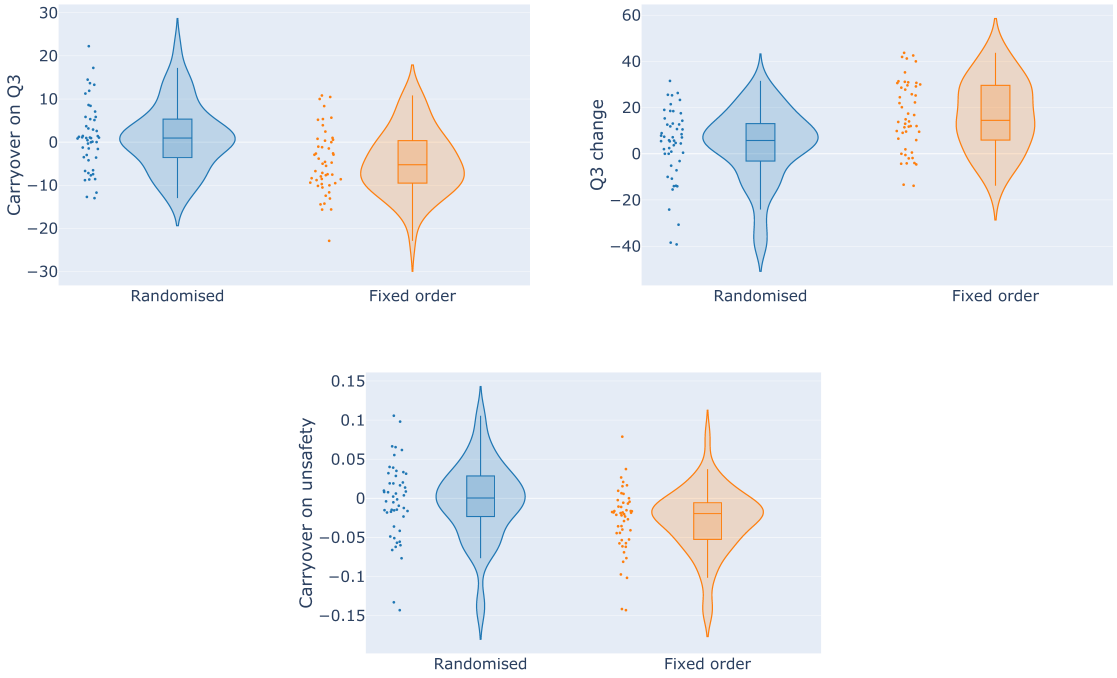

Fig. 4. Participant level learning and sequential metrics by ordering group. The top left panel shows the carryover of previous trial yielding on Q3, and the top right panel shows early to late drift in Q3. The bottom panel shows the carryover of previous trial yielding on mean unsafety.

Figure 5 separates carryover estimates by previous trial factor. These plots show whether Q3 carryover patterns were associated with the previous trial's eHMI status, camera condition, or inter pedestrian distance.

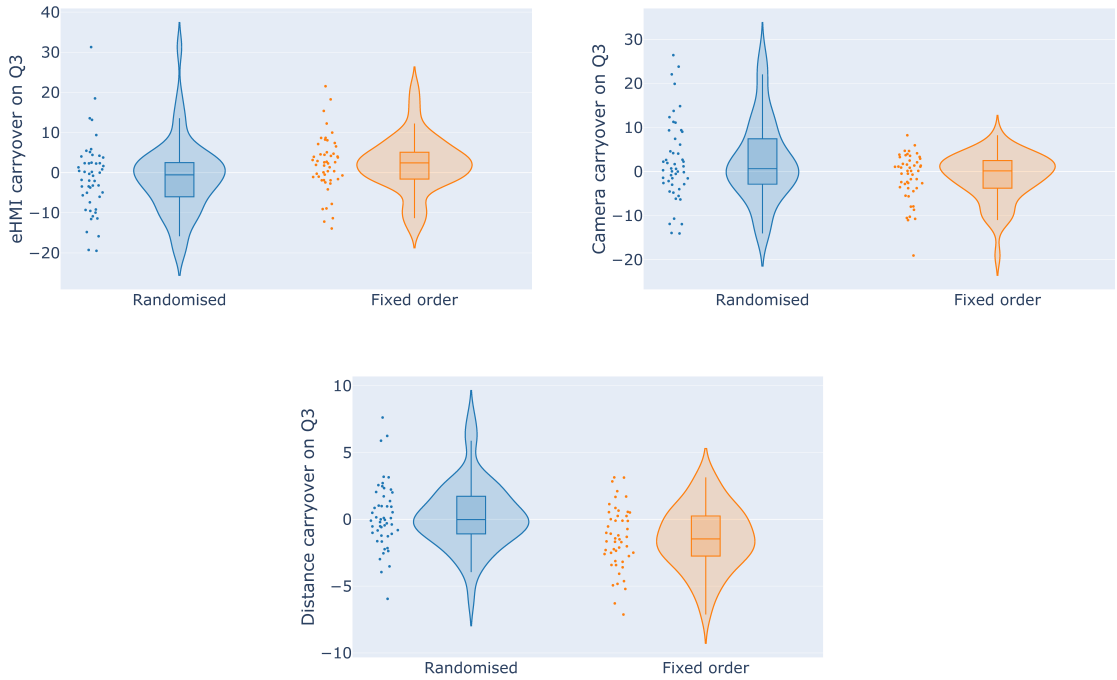

Fig. 5. Additional carryover estimates by previous trial factor. The top panel shows the carryover of previous trial eHMI status on Q3, the middle panel shows the carryover of previous trial camera condition on Q3, and the bottom panel shows the carryover of previous trial inter pedestrian distance on Q3.

Trigger response shape and event-aligned checks. Figure 6 describes the interval from the first unsafe response to the first return to safe. This is an event-defined timing measure and therefore applies only to trials in which both events occurred.

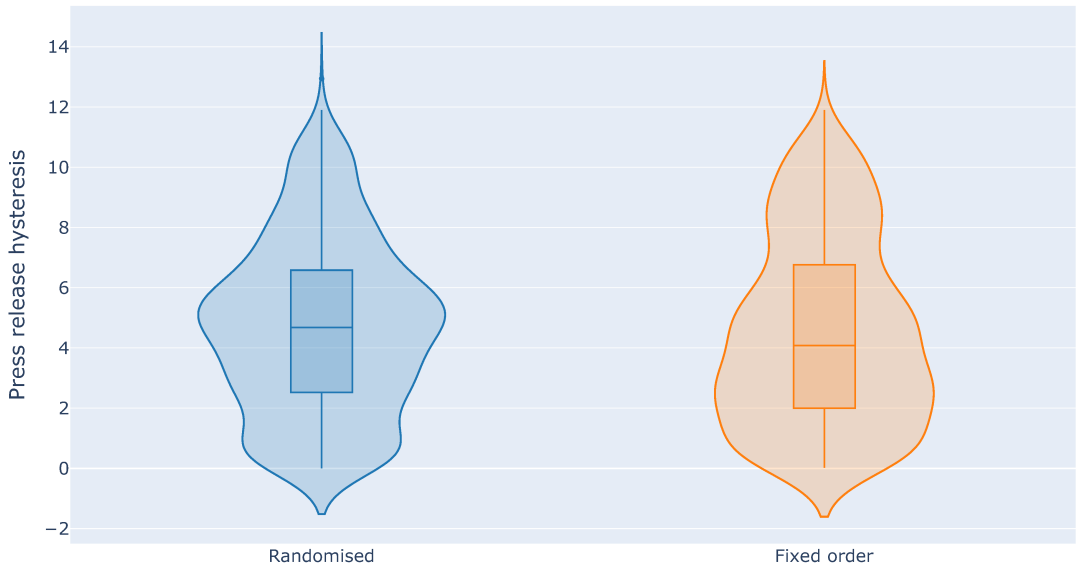

Fig. 6. First press to first release interval by ordering group. Larger values indicate that an initial unsafe commitment was maintained for longer once initiated.

Table 8 provides descriptive trigger shape summaries. These measures complement the primary trigger outcomes by describing integrated unsafety, occupancy above alternative thresholds, and the time from vehicle stop onset to return to safe.

Table 8. Trigger response shape metrics (randomised order vs fixed sequence; descriptive summaries).

Outcome	M_r	M_f	Δ
Integrated unsafety (integral)	4.05	4.05	-8.00×10^{-3}
Time above 10% unsafety threshold	0.34	0.34	1.00×10^{-3}
Time above 30% unsafety threshold	0.34	0.34	-1.00×10^{-3}
Time above 50% unsafety threshold	0.34	0.34	-1.00×10^{-3}
Time from stop onset to first return to safe (s)	1.90	1.73	-0.17

M_r and M_f are grand means across all experimental trials in the randomised order and fixed sequence groups. Participant level inference for these outcomes yielded no reliable differences after correction.

Table 9 reports trigger values at specific moments in the scenario, including yield onset, yield resume, and the participant's crossing point. These checks assess whether ordering related differences appeared at specific task events rather than only in trial averaged summaries.

Table 9. Event-aligned unsafety markers (randomised order vs fixed sequence).

Outcome	n_r	M_r (SD)	n_f	M_f (SD)	Δ	p	q
Unsafety rating at yield onset (yielding trials)	1050	0.47 (0.50)	1050	0.38 (0.48)	-0.09	0.27	0.30
Unsafe state at yield onset (yielding trials)	1050	0.47 (0.50)	1050	0.39 (0.49)	-0.08	0.30	0.30
Unsafety rating at yield resume (yielding trials)	1050	0.15 (0.35)	1050	0.23 (0.42)	0.08	0.14	0.30
Unsafe state at yield resume (yielding trials)	1050	0.15 (0.36)	1050	0.23 (0.42)	0.08	0.14	0.30
Unsafety rating at crossing point P2	2100	0.46 (0.50)	2100	0.52 (0.50)	0.05	0.30	0.30
Unsafe state at crossing point P2	2100	0.47 (0.50)	2100	0.53 (0.50)	0.06	0.24	0.30

Δ: fixed sequence minus randomised order mean difference. p values use OLS with participant clustered robust standard errors. q values are Benjamini–Hochberg adjusted across the six event aligned outcomes.

Factor composition and exposure checks. Table 10 gives a direct early versus late view of the factor drift described in the main results. It compares the average condition mix in the early and late parts of the session for each ordering group.

Table 10. Early versus late factor composition, by ordering group. Values are means, with proportions used for binary factors.

	Randomised		Fixed-sequence	
Factor	Early	Late	Early	Late
Yielding (proportion)	0.50	0.50	0.64	0.10
eHMI on (proportion)	0.52	0.50	0.18	0.70
Visibility condition (proportion at level 1)	0.51	0.50	0.46	0.70
Pedestrian distance factor (mean level)	2.98	2.94	2.91	3.20

Figure 7 re-expresses time on task as cumulative prior exposure to a focal condition. This figure complements the trial position models by showing whether between group differences were reduced when participants were compared by prior exposure rather than by raw trial position.

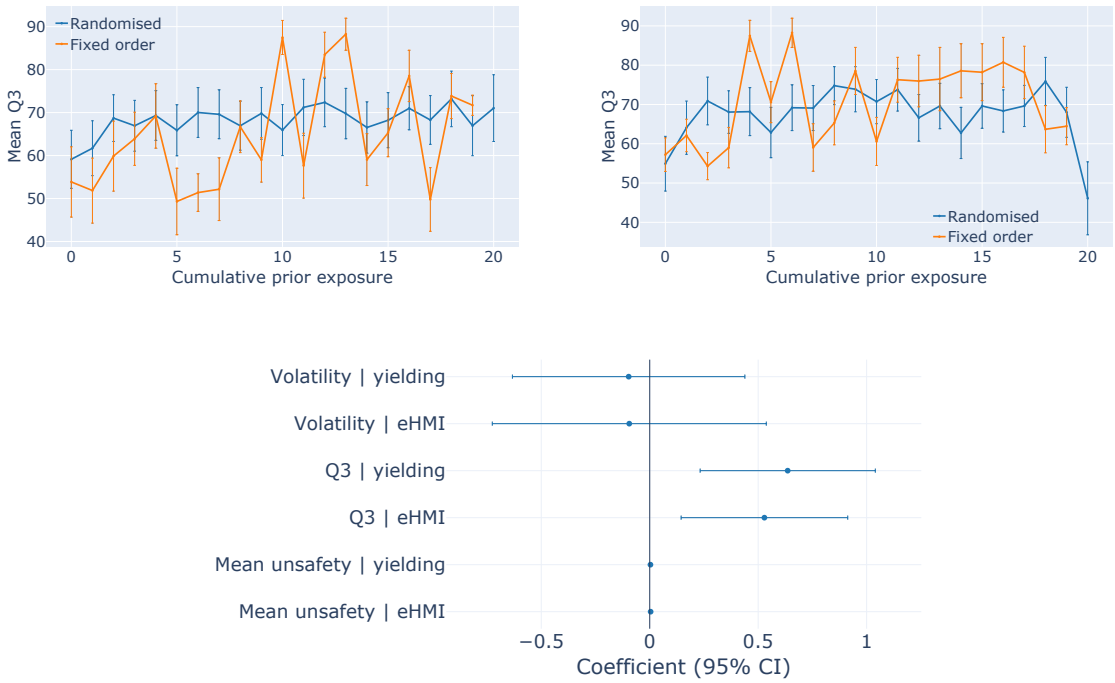

Fig. 7. Exposure based learning analysis. The top left panel shows Q3 as a function of cumulative prior exposure to yielding trials, and the top right panel shows Q3 as a function of cumulative prior exposure to eHMI on trials. The bottom panel summarises the ordering group by exposure interaction terms. In these analyses, time on task is expressed as prior exposure to the focal condition rather than raw trial position.

Reliability, break matched checks, and discriminability. Figure 8 and Figure 9 show within participant stability for the two main repeated outcomes. Odd even splits assess consistency across interleaved trials, whereas early late splits assess whether participant level responses remained stable across the session.

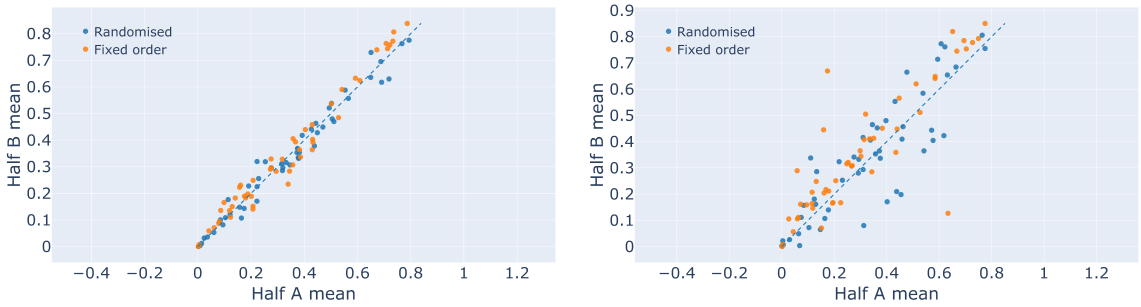

Fig. 8. Within participant stability of mean unsafety responding. The left panel shows the odd versus even split half correlation of participant means, and the right panel shows the early versus late split half correlation.

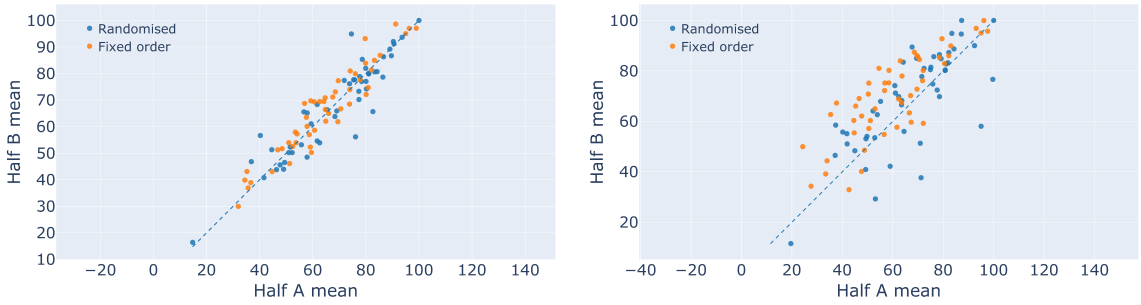

Fig. 9. Within participant stability of Q3 ratings. The left panel shows the odd versus even split half correlation of participant means, and the right panel shows the early versus late split half correlation.

Table 11 compares the changes before and after the break after the matching in yield and eHMI composition. This check examines whether apparent reset effects around the break remained after accounting for two key experimental factors.

Table 11. Break window reset effects controlling for factor composition (matched yielding×eHMI). Participant level outcomes are post−pre differences.

Outcome	n_r	M_r	n_f	M_f	t	q	d
Mean continuous unsafety rating	50	-8.00×10^{-3}	50	0.03	-2.82	0.02	-0.56
Q3	50	-9.00×10^{-3}	50	-1.26	0.87	0.58	0.17
Unsafety volatility (SD of rate of change)	50	2.66	50	0.10	0.39	0.70	0.08

Table 12 and Figure 10 report how well individual signals distinguished yielding from non yielding trials. Values near 0.50 indicate chance level discrimination, whereas larger values indicate stronger separation.

Table 12. Discriminability of yielding versus non yielding trials (ROC AUC) by signal and ordering group.

Signal	Randomised order	Fixed sequence	Δ (Randomised − Fixed)
Number of transitions	0.68	0.62	0.05
Q3	0.52	0.55	-0.04
Yaw forward fraction (within 15°)	0.51	0.54	-0.03
Unsafety volatility	0.48	0.41	0.07
Yaw SD	0.49	0.46	0.02
Mean absolute yaw	0.50	0.46	0.04
Mean unsafety rating	0.37	0.38	-8.00×10^{-3}
Fraction time unsafe	0.37	0.38	-7.00×10^{-3}
Trigger SD	0.47	0.48	-0.01

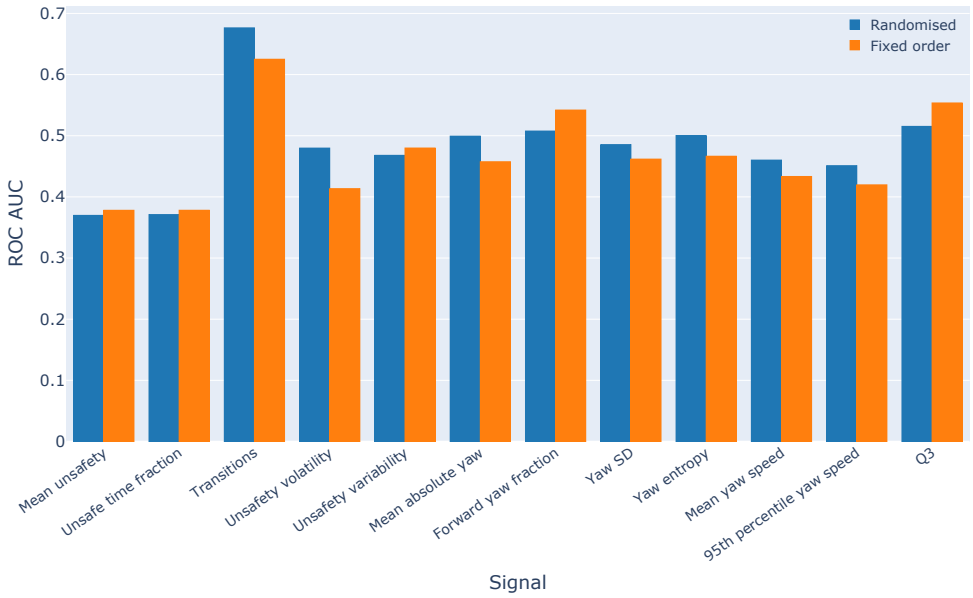

Fig. 10. ROC AUC by signal and ordering group for discriminating yielding versus non yielding. Dashed line indicates chance (AUC = 0.50).

Received 20 February 2026; revised 12 March 2026; accepted 5 June 2026