Fixed or randomised trial order? A case study of pedestrian interaction with an automated vehicle

Md Shadab Alam[a,*], Marieke H. Martens[a,b] and Pavlo Bazilinskyy[a]

[a] *Eindhoven University of Technology, Eindhoven, The Netherlands*
[b] *TNO, Helmond, The Netherlands*

ARTICLE INFO

Keywords:
Trial randomisation
Trial order
Virtual reality
Automated vehicles
Pedestrian behaviour
Experimental design
Repeated measures

ABSTRACT

In a virtual reality study of pedestrian–automated vehicle interaction, the same trial order was used by mistake for every participant. A second study presented the same 40 conditions to a new group, randomising the order for each participant. We compared the two groups of 50 participants to establish which conclusions a fixed sequence alters. Participants stood at the kerb near a virtual pedestrian, held a controller trigger whenever they felt it was unsafe to cross and rated the influence of the other pedestrian and of the distance to them. The effects of vehicle yielding, the external human–machine interface (eHMI) and pedestrian position on perceived crossing safety were the same in both groups; none of 12 group differences was significant after adjustment. Overall, perceived crossing safety did not differ significantly (difference 2.34 percentage points, 95% CI $[-5.16, 9.83]$). Of 600 condition-by-condition comparisons, one differed after adjustment: in the condition that every fixed order participant met first, the fixed order group rated the influence of distance 23.34 points lower (adjusted $p = .011$). When the conditions were ignored, judged unsafety rose by 21.76 percentage points from the first to the last ten trials in the fixed order group, against 2.88 in the randomised order group, because the sequence ended with non-yielding vehicles. As the groups were tested at different times and not randomly assigned, differences cannot be attributed to trial order alone. A fixed sequence can still support conclusions about average factor effects, but not about single early conditions or change over the session.

1. Introduction

Imagine a virtual reality (VR) study of how pedestrians interact with approaching automated vehicles. The researchers want to know how pedestrians perceive the light signals that such vehicles use to show what they are going to do. When a participant meets the vehicle for the first time, they may not know what its light signal means. After several trials, the same participant may recognise the signal and anticipate the vehicle's movement. Ratings can also change for a simpler reason. In the first trial there is no earlier situation to compare with, whereas later ratings are given with the previous trials in mind, so the same situation may be rated higher or lower depending on what preceded it (Poulton, 1973). If every participant receives the same trial order, these changes over time always accompany the same experimental conditions.

This was the problem in our original VR study: by mistake, the same trial order was used for all participants. We subsequently ran a second study with the same conditions and a new group of participants, this time with the trial order randomised for each participant. In the rest of this paper, we call the first data collection the fixed order study and the

*Corresponding author.
m.s.alam@tue.nl (M.S. Alam); m.h.martens@tue.nl (M.H. Martens); p.bazilinskyy@tue.nl (P. Bazilinskyy)
ORCID(s): 0000-0001-9184-9963 (M.S. Alam); 0000-0002-1661-7019 (M.H. Martens); 0000-0001-9565-8240 (P. Bazilinskyy)

second the randomised order study, and refer to their participants as the fixed order group and the randomised order group. Having both datasets allowed us to examine what randomisation changes in the results. Such a comparison is rarely possible, because the same conditions are seldom tested with two different ordering procedures. This paper reports that comparison.

Below, we first explain why trial order matters when the same participants experience several conditions. We then describe the crossing situation in our experiment, the factors that can make it uncertain for a pedestrian, and why responses to it may change over repeated encounters, before stating the aim of the study.

In a repeated measures design, the same participants experience several conditions. This reduces the influence of individual differences on comparisons between conditions (Greenwald, 1976; Charness et al., 2012). Such designs are common in human factors research, including studies of pedestrians and automated vehicles. Many of these studies use VR, so that vehicle behaviour, communication and the traffic scene can be varied without exposing participants to real traffic (Camara et al., 2021; Feng et al., 2021; Tran et al., 2021).

However, the response to a condition may also depend on when it is presented in the session. With a fixed trial sequence, a response to a particular condition may be affected by the experience that participants gained in the preceding trials. Randomising the trial order for each participant is a common way to deal with this. It does not stop participants from becoming familiar with the task or tired during the session, but it prevents these changes from always coinciding with the same condition (Fisher, 1966; Montgomery, 2017). Each condition then appears early for some participants and late for others, so the influence of experience is spread over all conditions rather than being mixed up with the effect of any one of them. Counterbalancing does the same in a planned way: the trial orders are constructed so that each condition appears equally often at each position and, in some designs, equally often after each of the other conditions (Williams, 1949; Jones and Kenward, 2003; Brooks, 2012).

How much trial order matters in a particular study depends on what participants have to interpret, and therefore on the situation they face. Most studies of external human–machine interfaces (eHMIs) have placed a single pedestrian at the kerb (Dey et al., 2021b; Tran et al., 2024). In real traffic, however, pedestrians often wait to cross next to other people, and an approaching vehicle then has to be interpreted with those people in mind. Our experiment therefore included a second pedestrian. The participant stood at the kerb, and a virtual pedestrian stood on the same pavement, either before or after the participant along the vehicle's path. The participant judged whether it was safe to cross in front of an approaching automated vehicle. Three features of this situation were varied: how the vehicle moved, whether it displayed an eHMI, and where a second pedestrian stood. Each of these can leave the pedestrian uncertain about what the vehicle will do, and each is introduced below.

The first is the movement of the vehicle. When deciding whether it is safe to cross, pedestrians rely mainly on how the vehicle moves: its speed, its distance and whether it slows down (Rasouli and Tsotsos, 2019; Dietrich et al.,

2020). Dey et al. (2019) found that driving behaviour influenced willingness to cross more than whether the vehicle was described as automated or manually driven. A vehicle that slows down, however, does not always make clear whether it will stop.

The second is the eHMI. In today's traffic, explicit signals between drivers and pedestrians, such as hand gestures or eye contact, are used only occasionally (Lee et al., 2021). Even so, a pedestrian can usually look at the driver to check whether they have been seen. In an automated vehicle there may be no driver, or the person in the driver's seat may not be watching the road, so this check is no longer possible. eHMIs have been proposed to fill this gap. They are displays on the outside of the vehicle that communicate, for example, that it is driving automatically, that it has detected the pedestrian or that it is going to stop. Proposed designs include light bands, symbols, text and projections on the road surface (Habibovic et al., 2018; Dey et al., 2020a; Faas et al., 2020).

Whether eHMIs are needed at all is still debated, because the vehicle's movement already conveys much of this information (de Winter and Dodou, 2022). Their benefit seems greatest when the movement is ambiguous. In a VR study, pedestrians felt safe to cross for a larger part of the approach when a yielding vehicle displayed an eHMI (De Clercq et al., 2019). Dey et al. (2021a) found that an eHMI helped mainly when the way the vehicle braked left it unclear whether it would stop. Studies of eHMIs whose message changed with the distance to the pedestrian, and of an explicit signal that the vehicle would not yield, addressed similar situations (Dey et al., 2020b, 2022).

The third is the presence of another pedestrian. A pedestrian waiting alone can assume that a vehicle that stops is stopping for them. When another pedestrian is nearby, a vehicle may signal that it will stop without making clear whom it is stopping for (Dey et al., 2021b). In a simulator with two interacting participants, Tran et al. (2024) found that participants sometimes sought information from one another when the vehicle's communication was unclear. The crossing behaviour of a group of pedestrians can also influence an individual's decision (Hübner et al., 2025). The virtual pedestrian in our experiment remained stationary and never crossed, so it could not show the participant when to cross. What changed between conditions was its position: we varied the distance between the two pedestrians and which of them the vehicle reached first. A yielding vehicle therefore stopped closer to one of the two pedestrians, which could leave the participant unsure whom it was stopping for.

Participants therefore have to interpret the vehicle's movement, a signal that is usually new to them, and the position of the other pedestrian, and their interpretation can change during the experiment. Several studies have shown that pedestrians' responses to automated vehicles change when similar situations are encountered repeatedly. First, participants learn what an unfamiliar eHMI means. Lee et al. (2022) followed how pedestrians came to interpret a light band they had not seen before, and Colley et al. (2022), Colley et al. (2024) and Shahi et al. (2026) examined crossing decisions over repeated exposures, the latter two across several sessions. Second, responses can change in ways other than learning the signal. In a situation where it was unclear whether the vehicle was stopping for them,

participants became more cautious across trials (Lee et al., 2025). Third, a single experience can carry over to later trials. After eHMI malfunctions, participants' trust in and acceptance of the signals changed (Hensch et al., 2022), and participants who had experienced a reliable yielding signal started to cross when a vehicle showed the signal but did not stop (Kaleefathullah et al., 2022). Together, these studies show that the response in a given trial partly depends on the trials that came before it. Such changes are not always learning of the interface; participants may also become more cautious, lose trust or simply get used to the task. In a fixed sequence, however, each of these changes accompanies the same conditions for every participant and cannot be separated from the effect of those conditions.

Studies of pedestrians and automated vehicles have ordered their trials in different ways. Some randomise the order of all conditions, some group conditions into blocks, and some use predefined sequences that are balanced across participants. Dey et al. (2022) grouped the trials into blocks by eHMI type and randomised both the order of the blocks and the order of the trials within each block. Tran et al. (2024) gave participants different predefined sequences arranged as a balanced Latin square, so that each condition appeared equally often at each position. In Dey et al. (2020b), the block without an eHMI always came first and the remaining blocks were randomised, so participants always experienced the vehicle without an eHMI before seeing any of the signals. These choices affect how results can be interpreted. In the last example, every comparison between the vehicle with and without an eHMI was also a comparison between earlier and later trials. Because each of these studies used a single ordering procedure, they could not show how the results would have differed under another one. Our two datasets allow this for one specific case: reusing a single sequence for every participant compared with randomising the order for each participant. The two datasets were collected at different times, however, and participants were not randomly assigned to them.

1.1. Aim of the study

This comparison was not planned when the fixed order study was designed. Using the two datasets, we examined whether the fixed order group and the randomised order group differed in their overall responses, averaged across the 40 conditions, in their responses to each of the 40 conditions separately, and in how their responses changed over the session. The aim was to establish which conclusions were similar across the datasets and where a fixed sequence made the findings difficult to interpret.

2. Method

2.1. Study design and participants

The experiment was originally designed to study how the presence of another pedestrian affects a pedestrian's willingness to cross in front of an automated vehicle. It examined whether the distance to the other pedestrian, and which of the two pedestrians the vehicle reached first, changed how safe participants felt to cross, and whether this

depended on the vehicle's yielding behaviour and on an eHMI. The results of the randomised order study are reported by Alam et al. (2026). The present paper uses the two datasets for a different question: whether the way trials were ordered affected the results.

The fixed order study and the randomised order study each included 50 participants, and different people took part in each. Both studies used the same apparatus, virtual environment, task and 40 experimental conditions; only the trial order differed. Participants in the fixed order study all received the same sequence, whereas participants in the randomised order study each received their own randomised sequence. Participants were not randomly assigned between the two groups.

The randomised order group had a mean age of 28.30 years ($SD = 7.30$); 28 participants identified as male, 21 as female, and one selected another or undisclosed category. The fixed order group had a mean age of 25.28 years ($SD = 4.61$); 29 identified as male, 20 as female, and one selected another or undisclosed category. The study was approved by the Ethics Review Board of the Eindhoven University of Technology. All participants provided informed consent.

2.2. Experimental setup and design

Participants wore a Meta Quest 3 head-mounted display and held a Meta Quest controller in their right hand. The controller had no influence on the vehicle or on anything else in the virtual scene. Participants used its trigger to indicate when they considered it unsafe to cross (Section 2.3) and used the controller to answer the questions shown after each trial. The virtual environment was developed in Unity 2022.3.5f1. Its starting point was the VR environment of Dey et al. (2021b), which had been built for a remote study with smartphone-based headsets; we adapted it for the Meta Quest 3 and developed it further for a related study (Alam et al., 2026). The analogue controller trigger and the position and orientation of the head-mounted display were recorded at 50 Hz.

The scene and the vehicle were based on Dey et al. (2021b), a VR study with two pedestrians that examined whether eHMIs could make clear which of them an automated vehicle was yielding to. We kept the main elements of that study, so that the second pedestrian would create the same kind of ambiguity, and varied the factors that our study set out to test (Section 2.1). The virtual environment showed a straight residential street without a zebra crossing or an intersection. As in Dey et al. (2021b), participants therefore could not rely on having right of way and had to base their judgement on the approaching vehicle. Participants stood at a fixed spot on the kerb and were asked to imagine that they planned to cross the road. They could turn their head and look around freely, but they could not walk, so their position in the scene did not change during a trial. A life-sized virtual pedestrian, the other pedestrian whose influence the experiment was designed to study, stood on the same pavement and faced the approaching vehicle. It was an adult man in casual clothing. We kept this avatar unchanged from the environment of Dey et al. (2021b), so that the

scene stayed comparable with that study, and every participant saw the same avatar. Its appearance was not matched to the participant's gender or age. Its behaviour was the same in every trial: it stood still at the kerb, looked towards the approaching vehicle and did not walk or cross. Only its position varied between conditions. Dey et al. (2021b) placed the two pedestrians 10 m apart, a distance chosen to be close enough for a yielding message to apply to either pedestrian but far enough to avoid group behaviour. Because our study asked how this distance affects the participant, we varied it from 2 to 10 m in steps of 2 m.

The vehicle approached from the participant's left, and its movement followed Dey et al. (2021b). It accelerated to 50 km/h within the first 1.25 s of the trial, more than 100 m before the pedestrians. In non-yielding trials, it kept this speed and drove past both pedestrians. In yielding trials, it braked at a constant 2.4 m/s^2, the deceleration that Deligianni et al. (2017) observed in normal driving and that Dey et al. (2021b) also used. Braking from 50 km/h at this rate takes 40.19 m, so the vehicle started braking 43.19 m before the first pedestrian and stopped 3 m before that pedestrian. It remained stationary for 3 s and then drove off at 2 m/s^2. The vehicle always yielded to the pedestrian it reached first. When the participant was second, the vehicle therefore stopped for the other pedestrian, which is the situation in which a participant could mistakenly believe that the vehicle was stopping for them. Non-yielding trials were included, as in Dey et al. (2021b), so that participants could not expect every vehicle to stop.

The eHMI was the light bar used by Dey et al. (2021b), which developed earlier light-band concepts (Habibovic et al., 2018; Dey et al., 2020b). It was mounted on the front of the vehicle. When the eHMI was on, the bar glowed steadily in cyan while the vehicle approached, and two light segments repeatedly wiped from the outer ends of the bar towards its centre while the vehicle yielded. When the eHMI was off, the bar stayed dark. We chose this design because it communicates only that the vehicle is yielding, not whom it is yielding to. Dey et al. (2021b) found that such a general yielding message does not make clear which pedestrian the vehicle intends to stop for when two pedestrians are present. This made it suitable for studying how the position of the other pedestrian affects the participant's judgement.

The experiment used a repeated measures design of $2 \times 2 \times 2 \times 5$. The four factors were:

- vehicle behaviour: yielding or non-yielding;
- eHMI: off or on;
- pedestrian position: participant first or participant second along the vehicle's path; and
- distance between the participant and the other pedestrian: 2, 4, 6, 8, or 10 m.

The four factors produced 40 conditions. A condition is one combination of the four factors; for example, one condition was a yielding vehicle with the eHMI off, the participant first and the pedestrians 6 m apart. Each participant experienced each condition once.

The pedestrian position factor describes which of the two pedestrians stood nearer to the approaching vehicle. When the participant was first, they were the nearer pedestrian, and a yielding vehicle stopped 3 m before them. When the participant was second, the virtual pedestrian was nearer, and a yielding vehicle stopped 3 m before the virtual pedestrian. The vehicle then stood 5, 7, 9, 11 or 13 m before the participant, depending on the distance between the two pedestrians (Figure 1). The virtual pedestrian was present in every experimental trial, but whether it was in view depended on its position. When the participant was second, the virtual pedestrian stood between the participant and the approaching vehicle, so it was in view whenever the participant looked towards the vehicle. When the participant was first, it stood on the side away from the vehicle, and the participant had to turn their head away from the vehicle to see it. The two practice trials did not include a virtual pedestrian.

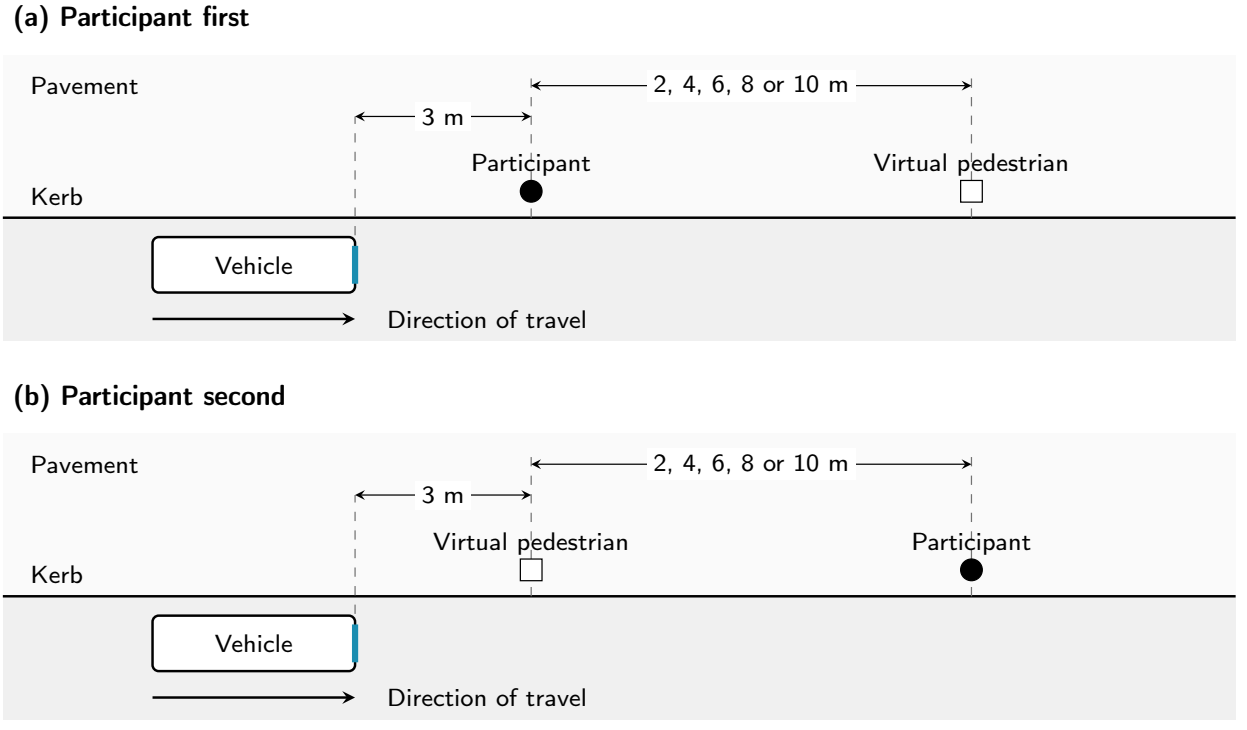


Figure 1: Bird's-eye schematic of the two pedestrian positions (not to scale). The filled circle represents the participant and the open square the stationary virtual pedestrian. The vehicle approaches from the left and is shown at its stopping position in yielding trials, 3 m before the first pedestrian. In non-yielding trials it continues past both pedestrians. The cyan mark identifies the location of the eHMI; its activation was varied separately. Participants remained at the kerb and used a controller trigger to indicate perceived crossing safety.

Both studies used the same randomisation procedure in the Unity C# code. It assigned a random number to each of the 40 conditions and sorted the conditions by these numbers, which produced a random order; the two practice trials always came first. The studies differed only in how often this procedure was run. In the fixed order study, it was run once, before data collection, and the resulting sequence was saved and then used for every participant. In the

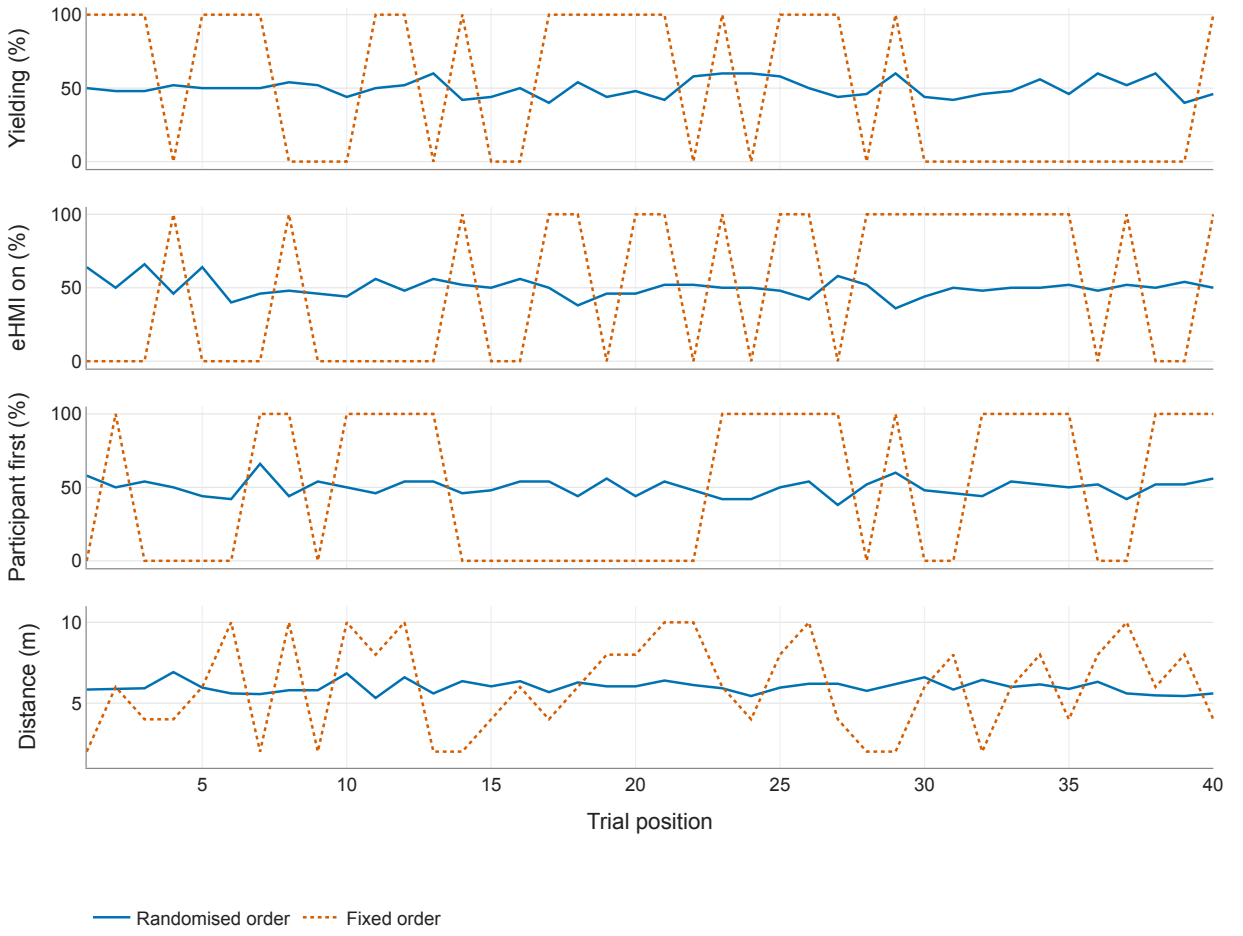


Figure 2: Experimental factors at each trial position. The upper three panels show the percentage of trials at each position that had a yielding vehicle, an active eHMI and the participant first; the lowest panel shows the mean distance between the pedestrians. The randomised order line is the mean over the 50 participants in that group. The fixed order line is the single sequence that all 50 participants in the fixed order group received, so in the upper three panels its values are either 0% or 100%.

randomised order study, it was run again for each participant, so that every participant received a different random sequence. The fixed sequence was therefore random in itself, but it was the same for everyone.

Figure 2 shows the distribution of experimental factors over trial positions. Each fixed order condition occupied the same position for all participants; the randomised conditions appeared at different positions between participants.

2.3. Procedure and dependent measures

Participants first read an information sheet, provided informed consent and completed a demographic questionnaire. The experimenter then explained and demonstrated the use of the controller trigger. Participants were instructed to hold the trigger whenever they considered it unsafe to start crossing the road and to release it when they considered it safe.

This continuous measure follows Dey et al. (2021b), in which participants indicated their willingness to cross in real time while the vehicle approached. It differs in two respects: Dey et al. (2021b) asked participants to hold a button while they felt safe rather than unsafe, and their button was binary, whereas the trigger used here recorded values between 0 and 1. We asked participants to hold the trigger while they felt unsafe because the vehicle was not yet visible at the start of a trial. Holding the trigger from the start of the trial until the situation became unsafe would have been less intuitive and easier to forget (Alam et al., 2026). They could press and release the trigger more than once during a trial. Participants remained at the kerb and did not physically cross the virtual road.

Participants completed two practice trials before the 40 experimental trials. Optional breaks were offered after trials 14 and 26 to reduce fatigue; participants could take or skip each break (Alam et al., 2026). The instruction to hold the trigger was shown again after each break. Participants were told that another pedestrian would be beside them at the kerb, sometimes in front of them and sometimes behind them. The virtual pedestrian remained stationary throughout the experimental trials and did not enter the road.

Perceived crossing safety, the primary dependent measure, was derived from the controller trigger. We analysed the final 5 s before the vehicle reached the participant. In yielding conditions, the vehicle reached the participant only after it had stopped, waited for 3 s and started moving again.

The 5 s period was divided into 50 non overlapping intervals of 100 ms. The trigger value ranged from 0 when the trigger was released to 1 when it was fully pressed. An interval was classified as one in which crossing was judged unsafe when at least one valid trigger value in that interval exceeded 0.10, that is, when the trigger was pressed beyond one tenth of its travel. This threshold therefore refers to how far the trigger was pressed, not to time. We chose a low threshold so that an interval counted as unsafe as soon as the participant began to press, while ignoring very small values that did not reflect a deliberate press. In practice, participants either released the trigger or pressed it almost fully: of all trigger values recorded during the trials, 71.58% were 0, 25.36% were above 0.95, and only 3.07% lay in between. For each trial, we calculated the percentage of intervals in which crossing was judged unsafe. A value of 0% means that the participant judged crossing to be safe throughout the 5 s period, and a value of 100% that they judged it unsafe throughout. A value of 40%, for example, means that crossing was judged unsafe in 20 of the 50 intervals.

After each trial, three questions were shown on a panel in the virtual environment, so participants kept the headset on. They answered each question by moving a slider from 0 to 100 with the controller:

Q1 "On a scale of 0 to 100, how much did the behaviour of the other pedestrian influence your decision to cross the road?"

Q2 "On a scale of 0 to 100, how much did the distance between you and the other pedestrian affect your decision to cross the road?"

Q3 "On a scale of 0 to 100, rate how well you understood the intention of the vehicle."

In Q1, the behaviour of the other pedestrian refers to what the virtual pedestrian did in the scene, namely standing at the kerb and looking towards the approaching vehicle. Q1 and Q2 ask how much the other pedestrian and the distance to that pedestrian influenced the participant's decision, and Q3 asks how well the participant understood the vehicle. None of the three asks whether the participant was willing to cross; that judgement was measured separately, and continuously, with the trigger.

In addition to the percentage of time that crossing was judged unsafe, we examined several aspects of controller use. These included the average and maximum trigger value, whether the trigger was activated at least once, the time of the first activation, whether the trigger was subsequently released, and the time of the first release.

Horizontal head orientation was derived from the head-mounted display. The participant's average orientation from 0.02 to 0.30 s after the start of each trial was used as the reference orientation. Subsequent head angles were expressed relative to this starting position, with negative values indicating that the participant turned towards the approaching vehicle on the left.

The head orientation when the vehicle reached the participant was calculated as the average orientation during the 200 ms interval centred on that moment. We also examined the largest head turn towards the vehicle before it reached the participant, the change in orientation immediately before and after this moment, variation in head orientation during the 5 s period, and yaw rate. To check whether participants turned towards the virtual pedestrian when it was out of view, we also determined the largest head turn away from the vehicle, from 0.5 s after the start of the trial until 0.2 s before the vehicle reached the participant. A turn of more than 45° was counted as a turn towards the side on which the virtual pedestrian stood when the participant was first. This check was descriptive and was not compared statistically between the groups. These measures describe head movement and should not be interpreted as measures of eye gaze or visual attention.

2.4. Statistical analysis

We compared the two groups in three ways, which correspond to the three questions in the aim (Section 1.1). First, we compared their overall responses, averaged across all 40 conditions, for example whether crossing was judged unsafe for a larger part of the time in one group than in the other. Second, we compared the groups within each of the 40 conditions separately, to find out whether differences were confined to particular conditions. Third, we compared how responses changed from the beginning to the end of the session, to find out whether the two groups changed differently as they gained experience. The analysis distinguished one primary outcome from several secondary outcomes. The primary outcome was perceived crossing safety, as defined in Section 2.3: for each trial, the number of 100 ms intervals, out of 50, in which the participant judged crossing to be unsafe. We made this the primary outcome because it is the most

direct measure of the participant's judgement of whether it was safe to cross, and it was recorded continuously while the vehicle approached. The main conclusion about overall differences between the groups therefore rests on this one comparison. The other measures describe further aspects of the participants' responses and were treated as secondary. Because there were many of them, the chance that at least one would differ between the groups by coincidence was high, so their p values were corrected for multiple comparisons, as described below.

There were 14 secondary outcomes. Twelve were continuous measures, which can take any value within a range: the three ratings (Q1 to Q3, each from 0 to 100); the mean and the maximum trigger value in the 5 s period; the time until the trigger was first pressed and until it was first released; and five measures of head orientation, namely the heading when the vehicle reached the participant, the largest turn towards the vehicle, the change in heading as the vehicle reached the participant, the variability of the heading in the 5 s period, and the yaw rate. The other two were binary measures, which were either yes or no in each trial: whether the trigger was pressed at all, and whether it was released again after being pressed.

Each participant completed 40 trials, and trials from the same participant tend to be more alike than trials from different participants. All models took this into account. For the primary outcome, we used binomial regression, which models the probability that an interval was judged unsafe, with standard errors that allowed for the grouping of trials within participants. For the secondary outcomes, we used generalised estimating equations, which make the same allowance (Liang and Zeger, 1986): linear models for the continuous measures and logistic models for the binary measures.

All models described the responses in terms of the four experimental factors. They included the effects of vehicle behaviour, eHMI, pedestrian position and distance; all interactions between vehicle behaviour, eHMI and pedestrian position; and the interaction of distance with each of these three factors. Every one of these terms was estimated separately for the two groups, so that the model could show whether a factor had a different effect in the fixed order group than in the randomised order group.

The models were then used for the first two comparisons. For the overall comparison, we used the model to predict the response in each of the 40 conditions for each group, and averaged the 40 predictions, giving each condition the same weight. The overall difference is the difference between these two averages. For the comparison within conditions, we first tested whether the difference between the groups was the same in all 40 conditions. This was a single test of all the terms that allowed the effects of the factors to differ between the groups. We then compared the two groups in each of the 40 conditions separately. Finally, we compared the conclusions that each dataset would support about the experimental factors themselves, which were the questions the experiment was designed to answer. For perceived crossing safety, we estimated in each group the effect of each of three factors (vehicle behaviour, eHMI and pedestrian position) separately for each combination of the other two, averaged over the five distances. This gave 12 effects per

group. We compared each effect between the groups and adjusted the p values with the Holm method within each set of four effects of the same factor. Throughout the paper, differences are reported as the value in the fixed order group minus the value in the randomised order group, so a positive difference means a higher value in the fixed order group.

Because many comparisons were made, some would be expected to appear significant by chance alone. We therefore adjusted the p values with the Holm method, which takes into account how many comparisons belong to the same set (Holm, 1979). The sets differed for the two kinds of comparison.

In the comparisons within conditions, each outcome gave 40 comparisons, one for each condition, and the 15 outcomes (the primary outcome and the 14 secondary outcomes) gave 600 comparisons in total. For each comparison we calculated two adjusted p values: one that accounts for the 40 comparisons of the same outcome, and a stricter one that accounts for all 600. The differences in single conditions reported in the Results are judged against the stricter adjustment. For these comparisons we also report simultaneous 95% confidence intervals. An ordinary 95% confidence interval is intended for a single comparison. A simultaneous interval is wider, so that the intervals for all 40 conditions of an outcome together reach 95% confidence; we used the Bonferroni method for this.

In the overall comparisons, the primary outcome was a single test and was not adjusted. The 14 secondary outcomes were divided into four families of related measures: the three ratings; the mean trigger value, the maximum trigger value and whether the trigger was pressed; the time of the first press, whether the trigger was released again and the time of the first release; and the five head-orientation measures. For each secondary outcome, we report p values adjusted within its family and adjusted across all 14 secondary outcomes. All other confidence intervals in the paper are ordinary 95% intervals.

For the third comparison, we divided each session into three parts at the two optional breaks, after trials 14 and 26, which gave parts of 14, 12 and 14 trials. We used the breaks as boundaries for two reasons. They avoided the need to choose arbitrary cut points, and they were the only moments at which participants were offered a rest. Participants could take or skip each break. We estimated the response in each part for each group, using models that included the experimental factors and their interactions with the trial order group, as in the overall comparison. The main question was whether the change from the first to the final part differed between the groups. For the secondary outcomes, the p values of these comparisons were adjusted with the Holm method across all outcomes. To show what a simpler analysis of change would have concluded, we also calculated for each participant the difference between the mean response in the last ten trials and in the first ten trials, without taking the conditions into account, and compared these differences between the groups with Welch t tests. The p values of these tests were adjusted with the Benjamini–Hochberg method (Benjamini and Hochberg, 1995).

We also examined whether the effect of the eHMI changed as participants gained experience. For perceived crossing safety and for understanding of the vehicle's intention (Q3), we estimated the difference between yielding trials with

the eHMI on and off, and how this difference changed over the session. This change was modelled over time, using all 40 trials. Time was expressed in two ways. First, as the trial number: the models allowed the eHMI difference to increase or decrease steadily from the first to the last trial, and we summarise this trend by the change it implies between trial 1 and trial 40. We used a steady trend rather than a freely shaped curve because, in the fixed order group, every trial position contained one particular condition. A curve flexible enough to follow changes from trial to trial would therefore also follow the conditions that happened to occur at those positions. Second, as the number of earlier yielding trials with the eHMI on, which ranged from zero to nine. This measure counts only the encounters in which the meaning of the eHMI could be learned, and we summarise its trend by the change from zero to nine such encounters. In both cases, the change was compared directly between the groups. The tests of change within each group were adjusted with the Holm method for the 12 tests in this analysis. In the fixed order group, each condition appeared at the same trial position for every participant, so none of these analyses can separate the effect of a condition from the effect of its position in the sequence.

Finally, we checked whether the overall difference in perceived crossing safety depended on particular choices in the analysis. These checks concerned only the primary outcome. First, we repeated the analysis with trigger thresholds of 0.05 and 0.50 instead of 0.10, to see whether the result depended on how far the trigger had to be pressed for an interval to count as unsafe. Second, we estimated the uncertainty of the difference without relying on the assumptions of the regression model. We calculated each participant's mean percentage over the 40 trials and then formed 5,000 new samples, each consisting of 50 participants drawn at random, with replacement, from each group. The difference between the groups was recalculated in every sample, and the middle 95% of the 5,000 differences gives a confidence interval. This procedure is known as a bootstrap; we also used a Bayesian version with 20,000 draws, which gives a 95% credible interval. Third, we recalculated the difference 100 times, each time leaving out a different participant, to see whether the result depended on any single person. Fourth, because participants were not randomly assigned to the two groups and the groups differed in mean age, we repeated the primary comparison with age, gender and recent VR experience added to the model.

We also examined whether ratings given in a first trial differ in general. For this, we used the randomised order group, in which every condition could appear first. We compared the ratings (Q1 to Q3) given in trial 1, and in trials 2 and 3, with those given in trials 4 to 40, with the experimental condition included in the model. In addition, we report the mean Q2 rating in the condition that came first in the fixed sequence, separately for randomised order participants who encountered it in trials 1 to 5, 6 to 20 and 21 to 40. This check was exploratory, and its p values were not adjusted.

3. Results

3.1. Overall responses

All 100 participants completed 40 trials, giving 4,000 analysed trials. A complete 5 s analysis period was available for every trial.

Head turns away from the vehicle, towards the side of the virtual pedestrian, were more frequent when the participant was first. When the participant was first, and the virtual pedestrian therefore stood on the side away from the vehicle, participants turned their head more than 45° towards that side in 59.20% of trials (58.20% in the fixed order group and 60.20% in the randomised order group), and 97 of the 100 participants did so at least once. When the participant was second, and the virtual pedestrian stood between the participant and the vehicle, such turns occurred in 17.55% of trials.

Table 1 shows the overall comparisons for all measures. Perceived crossing safety did not differ significantly between the groups. Crossing was judged unsafe in 40.48% of the intervals in the randomised order group and in 42.82% in the fixed order group (difference 2.34 percentage points, 95% CI $[-5.16, 9.83]$, $p = .54$). The difference stayed between 1.52 and 3.20 percentage points in all checks of the analysis, and none of the intervals excluded zero (Table 2).

Table 1
Overall responses in the randomised order and fixed order groups, averaged over all 40 conditions.

Measure	Randomised	Fixed	Difference	95% CI	Adjusted p
Time crossing judged unsafe (%)	40.48	42.82	2.34	$[-5.16, 9.83]$	.54
Mean trigger value (0–1)	0.40	0.42	0.02	$[-0.05, 0.10]$	.61
Trials with trigger activation (%)	72.38	78.29	5.91	$[-3.21, 15.03]$	.61
Influence of the other pedestrian's behaviour	21.30	15.25	−6.05	$[-12.67, 0.57]$	.15
Influence of the distance between pedestrians	27.70	18.81	−8.89	$[-15.68, -2.09]$	.031
Understanding of vehicle intention	68.10	65.26	−2.84	$[-9.37, 3.69]$	.39
Head orientation when vehicle reached participant (degrees)	−5.77	−0.92	4.85	$[-2.47, 12.16]$	.78

Differences are fixed minus randomised. The primary p value is unadjusted. Secondary p values use Holm correction within each family of measures; confidence intervals are pointwise. Questionnaire ratings range from 0 to 100.

Table 2
Checks of the overall difference in perceived crossing safety (fixed minus randomised, percentage points).

Analysis	Difference	95% interval	p
Main analysis (trigger threshold 0.10)	2.34	$[-5.16, 9.83]$	.54
Trigger threshold 0.05	2.44	$[-5.05, 9.93]$	.52
Trigger threshold 0.50	2.14	$[-5.37, 9.65]$	.58
Bootstrap of participants	2.34	$[-5.13, 9.74]$	–
Bayesian bootstrap of participants	2.35	$[-4.86, 9.58]$	–
Leaving out one participant at a time	1.52 to 3.20	–	–
Adjusted for age, gender and VR experience	1.80	$[-6.12, 9.73]$	.66

The Bayesian bootstrap gives a credible interval and its value is the mean of the 20,000 draws. For the leave-one-out analysis, the range of the 100 differences is shown; none changed sign.

Averaged over all 40 conditions, the fixed order group rated the influence of the distance to the other pedestrian (Q2) lower than the randomised order group (18.81 versus 27.70; difference −8.89 points, 95% CI [−15.68, −2.09]). This difference was significant when its p value was adjusted for the three ratings (Q1 to Q3; $p = .031$), but not when it was adjusted for all 14 secondary outcomes ($p = .14$). None of the other differences in Table 1 was significant.

3.2. Responses across experimental conditions

The two groups showed the same effects of the experimental factors on perceived crossing safety (Table 3). In both groups, crossing was judged unsafe much less often when the vehicle yielded, and in yielding trials it was judged unsafe less often with the eHMI on and more often when the participant was first. The effect had the same direction in both groups in 10 of the 12 comparisons; the two exceptions were small effects whose confidence intervals included zero in both groups. In 10 of the 12 comparisons, the 95% confidence interval excluded zero either in both groups or in neither. None of the 12 differences between the groups was significant after adjustment (smallest adjusted $p = .09$).

Table 3
Effects of the experimental factors on perceived crossing safety in each group (percentage points).

Effect and situation	Randomised	Fixed	Difference [95% CI]	Adjusted p
Yielding minus non-yielding				
eHMI off, participant second	−48.04	−40.53	7.51 [−4.47, 19.49]	.88
eHMI off, participant first	−37.70	−32.79	4.91 [−8.69, 18.51]	.96
eHMI on, participant second	−51.62	−49.10	2.51 [−9.60, 14.62]	.96
eHMI on, participant first	−47.13	−39.94	7.19 [−5.98, 20.36]	.88
eHMI on minus off				
Non-yielding, participant second	−0.43	2.42	2.86 [−2.44, 8.15]	1.00
Non-yielding, participant first	0.17	0.28	0.11 [−5.41, 5.64]	1.00
Yielding, participant second	−4.01	−6.15	−2.14 [−9.25, 4.96]	1.00
Yielding, participant first	−9.26	−6.86	2.39 [−5.64, 10.42]	1.00
Participant first minus participant second				
Non-yielding, eHMI off	−2.38	1.73	4.10 [0.31, 7.90]	.10
Non-yielding, eHMI on	−1.78	−0.42	1.36 [−2.61, 5.33]	1.00
Yielding, eHMI off	7.96	9.46	1.50 [−4.25, 7.26]	1.00
Yielding, eHMI on	2.71	8.75	6.04 [0.89, 11.19]	.09

Effects are averaged over the five distances. Negative values mean that crossing was judged unsafe less often. Differences are fixed minus randomised; p values are adjusted with the Holm method within each set of four effects.

For perceived crossing safety, the difference between the groups was not the same in all 40 conditions ($p = .032$). When the groups were compared in each condition separately, none of the 40 differences in perceived crossing safety was significant. The largest differences occurred in two yielding conditions with the eHMI on and the participant first, with the pedestrians 2 m apart (11.36 percentage points, adjusted $p = .39$) and 10 m apart (10.77 percentage points, adjusted $p = .22$); both p values are adjusted for the 40 comparisons of perceived crossing safety, and both were 1.00 after adjustment for all 600 comparisons.

Of the 600 comparisons within conditions, one was significant after adjustment for all 600. It concerned the influence of pedestrian distance (Q2). In this condition, the vehicle yielded with its eHMI off, and the virtual pedestrian stood 2 m from the participant on the side of the approaching vehicle. The vehicle therefore stopped 3 m before the

virtual pedestrian and 5 m before the participant. The estimated rating was 53.41 in the randomised order group and 30.08 in the fixed order group. The difference was −23.34 points (simultaneous 95% CI [−40.93, −5.74], adjusted $p = .011$). In the same condition, crossing was judged unsafe in 22.84% of the intervals in the fixed order group and in 19.64% in the randomised order group (difference 3.20 percentage points, adjusted $p = 1.00$).

This was always the first trial for the fixed order group. In the randomised order group, it appeared at 29 different positions, spanning trials 1–40 (Figure 3). The corresponding 4 m condition was always trial 3 in the fixed sequence. Its ratings were also lower in the fixed order group, but the difference was not significant (−18.39 points, adjusted $p = .52$).

In the randomised order group, Q2 ratings given in trial 1 did not differ from those given in trials 4 to 40 (difference 1.84 points, 95% CI [−6.60, 10.29], $p = .67$), whereas Q3 ratings were 14.39 points lower in trial 1 (95% CI [−23.58, −5.20], $p = .002$). Four randomised order participants encountered the condition that came first in the fixed sequence within their first five trials. Their mean Q2 rating in this condition was 32.50, compared with 51.72 for the 25 participants who encountered it in trials 6 to 20, 50.33 for the 21 participants who encountered it in trials 21 to 40, and 24.24 in the fixed order group.

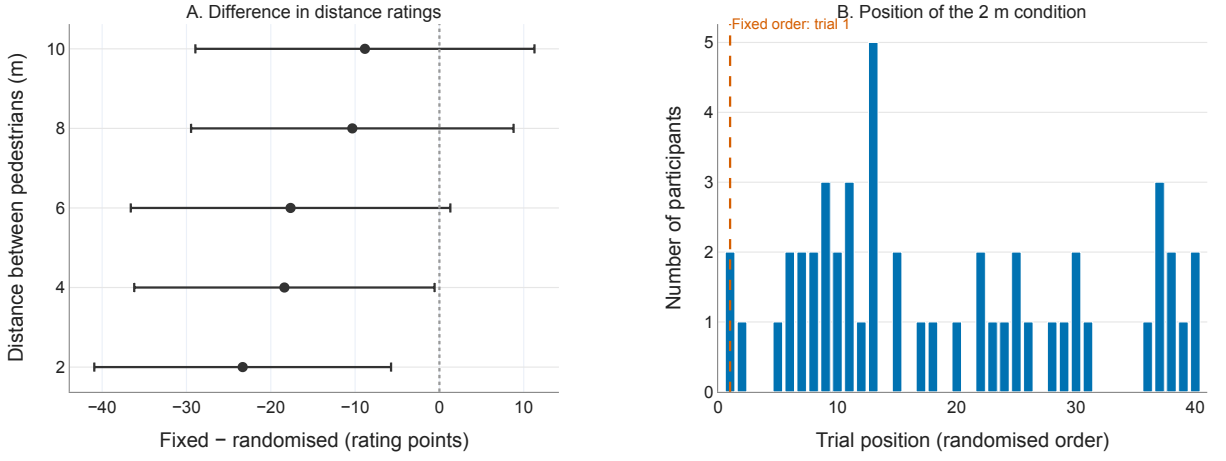


Figure 3: Reported influence of the distance to the other pedestrian (Q2) in yielding trials with the eHMI off in which the vehicle stopped for the virtual pedestrian (participant second). A: fixed minus randomised differences for the five distances; bars are simultaneous 95% confidence intervals for all 40 Q2 conditions. Negative values mean lower ratings in the fixed order group. B: trial positions at which participants in the randomised order group encountered the 2 m condition; in the fixed order group, this condition was always trial 1. These five conditions were selected for the figure after the condition comparisons had been examined.

3.3. Changes over the experiment

From the first to the final part of the session, the percentage of intervals in which crossing was judged unsafe changed by 1.94 percentage points in the randomised order group (95% CI [−1.55, 5.43]) and by 0.35 percentage

points in the fixed order group (95% CI [−3.40, 4.09]; Figure 4). The difference in change between the groups was −1.60 percentage points (95% CI [−6.72, 3.53], $p = .54$). Understanding of vehicle intention increased by 4.35 points in the randomised order group (adjusted $p = .37$) and by 5.68 points in the fixed order group (adjusted $p = .18$). The difference in change between the groups was 1.34 points (95% CI [−4.68, 7.36], adjusted $p = 1.00$). Head yaw rate decreased by 3.79 degrees per second in the randomised order group (adjusted $p < .001$) and by 2.44 degrees per second in the fixed order group (adjusted $p = .54$). The difference in change between the groups was 1.35 degrees per second (95% CI [−1.64, 4.35], adjusted $p = 1.00$).

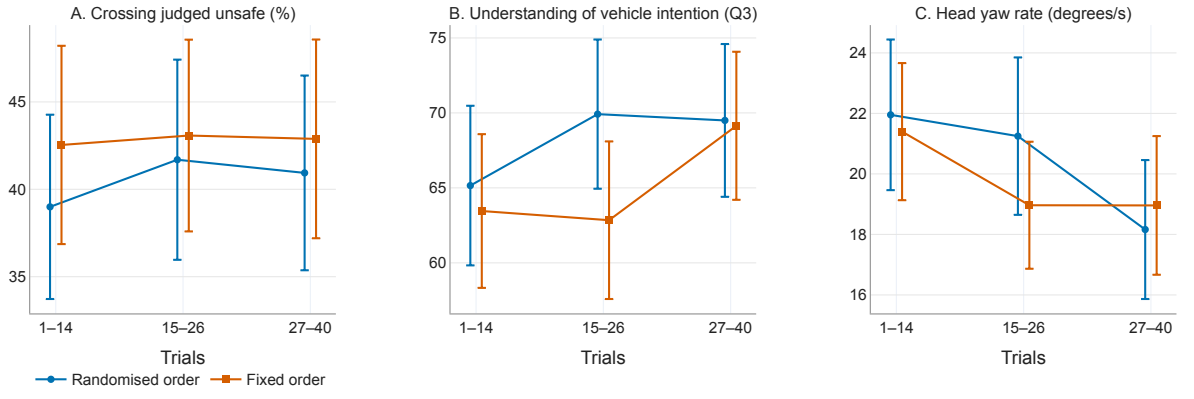


Figure 4: Responses in the three parts of the session, separated by the optional breaks after trials 14 and 26: A, percentage of intervals in which crossing was judged unsafe; B, understanding of the vehicle's intention (Q3); C, head yaw rate. The estimates come from the models that included the experimental conditions (Section 2.4) and give equal weight to all 40 conditions. Error bars are pointwise 95% confidence intervals.

Without taking the conditions into account, the two groups showed very different changes over the session (Figure 5). In the fixed order group, crossing was judged unsafe in 21.76 percentage points more of the intervals in the last ten trials than in the first ten; in the randomised order group, the increase was 2.88 percentage points (difference 18.88 percentage points, 95% CI [11.04, 26.72], adjusted $p < .001$). Unadjusted differences between the groups were also found for head heading when the vehicle reached the participant (29.10 and −0.99 degrees; difference 30.09 degrees, 95% CI [21.20, 38.98], adjusted $p < .001$) and for understanding of the vehicle's intention (14.62 and 5.53 points; difference 9.08 points, 95% CI [2.41, 15.76], adjusted $p = .015$). In the fixed sequence, trials 30 to 39 all had a non-yielding vehicle, and only one of the last ten trials had a yielding vehicle (Figure 2).

In yielding trials, the eHMI on/off difference in reported understanding became larger over the session in both groups (Figure 6B). The estimated increase was 17.81 points in the randomised order group (95% CI [8.59, 27.04], adjusted $p = .002$) and 11.85 points in the fixed order group (95% CI [−13.92, 37.62], adjusted $p = 1.00$). The direct comparison did not show a difference in change (−5.96 points, 95% CI [−33.23, 21.30], $p = .67$). The conclusion was unchanged when experience was expressed as the number of previous yielding encounters with an active eHMI.

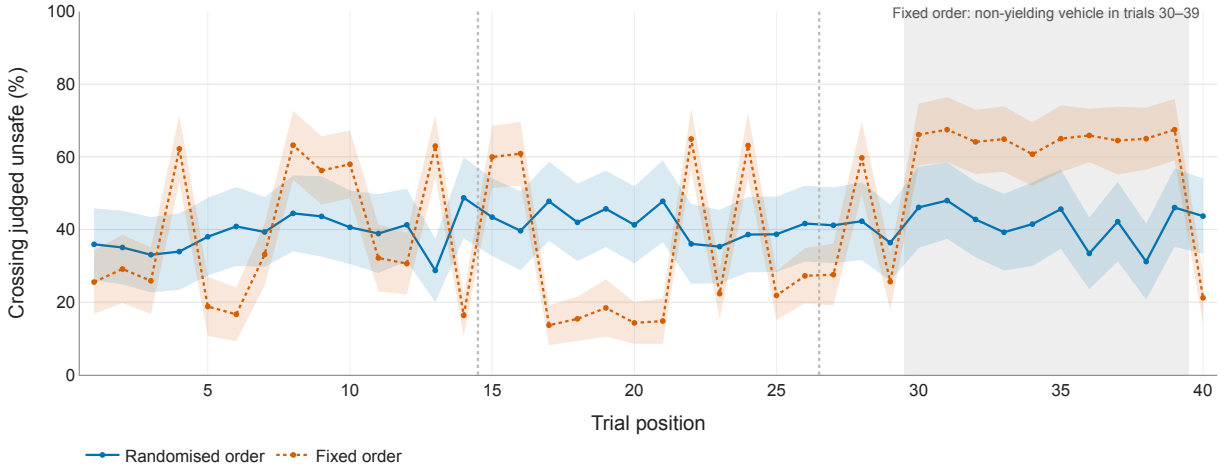


Figure 5: Percentage of intervals in which crossing was judged unsafe at each trial position, without adjustment for condition. Lines show the mean of the 50 participants in each group and shaded areas the 95% confidence intervals of the means. In the fixed order group, each trial position always contained the same condition. The grey area marks trials 30 to 39, which all had a non-yielding vehicle in the fixed sequence. Dotted vertical lines mark the optional breaks after trials 14 and 26.

For perceived crossing safety, the eHMI on/off difference changed by −0.61 percentage points from trial 1 to trial 40 in the randomised order group (95% CI [−8.33, 7.11]) and by 5.22 percentage points in the fixed order group (95% CI [−9.36, 19.80]); the difference in change between the groups was 5.83 percentage points (95% CI [−11.00, 22.66], $p = .50$).

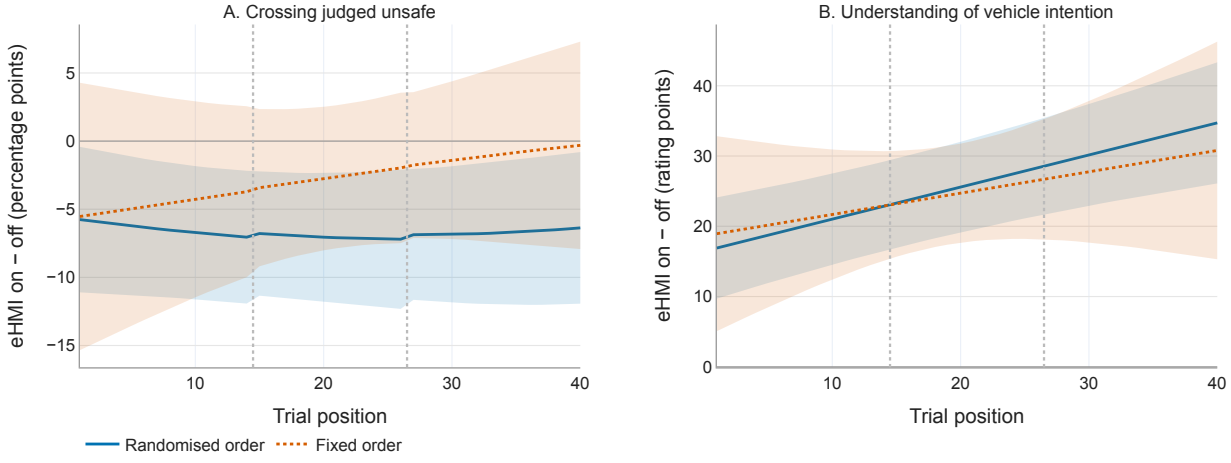


Figure 6: Difference between yielding trials with the eHMI on and off over the 40 trials. A: percentage of intervals in which crossing was judged unsafe; negative values mean that crossing was judged unsafe less often with the eHMI on. B: understanding of the vehicle's intention (Q3); positive values mean higher ratings with the eHMI on. Lines are model estimates and shaded areas pointwise 95% confidence intervals. Dotted vertical lines mark the optional breaks after trials 14 and 26.

4. Discussion

In a fixed sequence, every condition is tied to one trial position. In principle, therefore, none of the 40 conditions in the fixed order study can be separated from the stage of the session at which it was presented. What the comparison with the randomised order study adds is evidence of where this mattered in practice. Only one condition differed significantly between the groups, and it was the condition that every participant in the fixed order group encountered first. In this condition, the fixed order group reported that the distance to the other pedestrian had influenced their decision less. The corresponding condition with the pedestrians 4 m apart, which was always the third trial, showed a difference in the same direction, although it was not significant. In our data, the clearest consequence of the fixed sequence therefore appeared at the very start of the session, when participants had the least experience of the task.

In the condition that came first, the vehicle yielded with its eHMI off and stopped for the virtual pedestrian, who stood 2 m from the participant on the side of the approaching vehicle. When participants in the fixed order group encountered it, they had completed only the two practice trials, whereas randomised order participants could encounter the condition anywhere in the session. Our data do not show why the ratings differed. One possibility is that a first encounter is judged differently from later ones because participants have not yet seen other situations to compare it with (Poulton, 1973). Two observations from the randomised order group bear on this. First, Q2 ratings given in trial 1 were in general not lower than later ratings, so a general tendency to rate the first situation lower does not explain the result. Second, the four randomised order participants who encountered this particular condition within their first five trials rated it similarly to the fixed order group. This points to something specific to meeting this situation early, but four participants are too few for a firm conclusion. There is also a more basic alternative: the two groups consisted of different people, who were tested at different times and were not randomly assigned to a group. The lower ratings in the fixed order group could therefore also reflect who these participants were, rather than when they encountered the condition; the groups differed, for example, in mean age (Section 2.1). In particular, it cannot show how the fixed order group would have responded had this condition appeared later in the session.

For the question addressed here, an average over all 40 conditions says little. A fixed sequence affects each condition through its own position, so the consequences can differ from one condition to the next, and a difference confined to one condition largely disappears in the average. This is what happened in our data: the clear difference in distance ratings (Q2) in the condition that came first became much smaller, and no longer significant after adjustment for all secondary outcomes, when the ratings were averaged over all 40 conditions. A researcher who compared only the overall results of a fixed and a randomised study might therefore wrongly conclude that the fixed sequence did no harm. The difference in this condition was also limited to one measure. It was found in Q2, the rating that participants gave after each trial of how much the distance to the other pedestrian had influenced their decision. In the same condition, the two groups did not differ in how long they judged crossing to be unsafe while the vehicle approached. Participants who met this

condition first therefore did not judge the situation as more or less safe; they reported afterwards that the distance had mattered less to their decision. The difference concerned what participants reported about the situation, not their judgement of whether it was safe to cross.

In the models that took the conditions into account, the groups also did not differ significantly in how their responses changed over the session. One of these changes concerns the eHMI. In yielding trials, participants rated their understanding of the vehicle's intention higher when the eHMI was on than when it was off, and this gap became wider as the session went on (Figure 6B). In other words, the light bar helped participants more to understand that the vehicle was going to stop at the end of the session than at the start. This is what would be expected if participants learned the meaning of the light bar by encountering it, and it is consistent with studies of repeated exposure to eHMIs (Lee et al., 2022; Colley et al., 2022, 2024). The gap widened in both groups, and the groups did not differ significantly in how much it widened. This is what should be expected, because randomisation is not meant to change how participants learn. Its purpose is to spread the effect of learning over all conditions, so that the effect is not attributed to whichever conditions happen to come late in the session. The two groups differ in how the widening can be interpreted. In the randomised order group, every condition occurred both early and late across participants, so the widening can be attributed to experience. In the fixed order group, the late trials were always the same conditions, so the same widening could also reflect those particular conditions. Learning the meaning of the light bar did not lead to a corresponding change in perceived crossing safety: the eHMI on/off difference in how often crossing was judged unsafe did not change significantly over the session in either group (Figure 6A). Understanding what a signal means and feeling safe to cross are different judgements, and earlier studies suggest three reasons why the first need not change the second. First, pedestrians base their crossing decision mainly on how the vehicle moves, and Dey et al. (2021a) found that an eHMI mattered mostly when the movement itself was ambiguous. In our yielding trials, the braking of the vehicle already showed that it would stop. Second, when two pedestrians are present, a yielding signal does not say for whom the vehicle is stopping (Dey et al., 2021b; Tran et al., 2024). A participant who understood that the light bar meant that the vehicle was stopping could still be unsure whether it was stopping for them. Third, a similar gap between how people rate an interface and how they act has been found before: Carlowitz et al. (2023) reported positive ratings of an eHMI without a change in the time pedestrians took to decide to cross.

Would the original study have reached different conclusions if it had not been followed by the randomised order study? For the questions it was designed to answer, it would not. Both datasets show the same effects of vehicle behaviour, the eHMI and pedestrian position on perceived crossing safety, in the same direction and of similar size, and none of these effects differed significantly between the groups. The consequences of the fixed sequence were confined to single conditions, most clearly the condition presented first, and to a rating given after the trial rather than to the judgement of safety made while the vehicle approached. The mistake therefore mattered less than might have

been feared. This is partly because each level of a factor, such as yielding or an active eHMI, was spread over many positions in the fixed sequence, so its average effect did not depend on a single position.

The answer is different for any question about change over the session. Looked at without taking the conditions into account, the fixed order data suggest that participants judged crossing unsafe far more often towards the end of the session and turned their heads differently. A researcher with only these data could easily have concluded that participants became more cautious, or more tired, as the session went on. The randomised order data show no such change, and the models that accounted for the conditions found no difference between the groups. The apparent change arose because the fixed sequence happened to end with a long run of non-yielding vehicles, which are judged unsafe much more often. This was the largest distortion that the fixed sequence produced in our data, and it concerns precisely the questions about learning and fatigue that a fixed sequence cannot answer.

This suggests how results from a study that used one sequence for everyone can still be interpreted. Conclusions about the average effect of a factor are the most robust, provided that each level of the factor occurs throughout the sequence; this can be checked by plotting the factors against trial position, as in Figure 2. Results for single conditions are much less certain, especially for conditions in the first few trials, when participants are least familiar with the task. Such results should be reported together with the position of the condition, and they can be checked against similar conditions that appeared later in the sequence. Changes over the session should not be interpreted as learning, because in a fixed sequence they always coincide with the conditions that happen to come late.

The practical implication applies to any study in which each participant experiences several conditions one after another. Across participants, each condition should appear at different positions in the session and after different preceding conditions. Randomising the order separately for each participant achieves this. Randomising the order once and giving the resulting sequence to everyone, as in our fixed order study, does not: a random sequence that all participants share still ties each condition to one position. Counterbalanced sequences, such as balanced Latin squares, are an alternative when the order needs to be controlled more precisely (Williams, 1949; Jones and Kenward, 2003; Brooks, 2012; Tran et al., 2024). Neither approach prevents participants from becoming familiar with the task or tired; both prevent these changes from always coinciding with the same condition.

4.1. Limitations and future studies

The two groups were tested at different times, and participants were not randomly assigned to one group or the other. A difference between the groups may therefore be due to the trial order, but it may also be due to differences between the people in the two groups or to anything else that changed between the two data collections. Our comparison shows where the results of the two studies differed, but it cannot prove that the trial order caused these differences. In addition, which conditions are affected by a fixed sequence depends on the sequence itself. With a different sequence, a

different condition would have been presented first, and differences might have appeared in other conditions. A stronger test would be a single study in which participants are randomly assigned, at the same time, either to a fixed sequence or to a randomised order, preferably with several different fixed sequences. Finally, the fact that the groups did not differ significantly in perceived crossing safety does not mean that they responded in exactly the same way. With 50 participants per group, the study could detect only fairly large overall differences, so a smaller difference could have gone unnoticed. Here, no significant difference means that there is no clear evidence of a difference, not that there is evidence of no difference.

How much a fixed sequence matters depends on how much participants learn during the session, and the features of our experiment limited this. The street scene was the same in every trial, the other pedestrian never moved, and the eHMI was a light bar that was either steady or showed one animation. There was therefore relatively little for participants to learn beyond the meaning of the light bar. In experiments with more varied scenes, pedestrians who move or cross, or eHMIs that use text, symbols or projections, there is more to learn, and early trials may differ more from later ones. The consequences of a fixed sequence could then be larger than in our data. The same may apply when participants actually cross the road, because they then also experience the outcome of each decision. Head orientation did not measure eye gaze. The frequent head turns towards the virtual pedestrian when it stood out of view suggest that participants were aware of it in those trials, but they do not show exactly where participants looked. The virtual pedestrian was always an adult man, whereas both groups included a substantial number of participants who identified as female (Section 2.1). Participants may be influenced more by another pedestrian who resembles them, which could affect how strongly they reported being influenced by the other pedestrian (Q1) and by the distance to them (Q2). Because the avatar was the same and the gender composition was almost identical in the two groups, this is unlikely to explain the differences between groups, but it may limit how far the ratings themselves generalise.

That trial order should be randomised or counterbalanced is not new; it is a standard recommendation for repeated measures designs (Fisher, 1966; Montgomery, 2017; Brooks, 2012). What our accidental comparison adds is a view of what this recommendation protects against in an actual pedestrian experiment. Usually, the consequences of a fixed order can only be assumed, because a study is rarely run a second time with the same conditions in a different order. Here they could be observed. The average effects of the experimental factors were unaffected, a single early condition was affected, and an analysis of change over the session would have been misleading. In this sense, an error was needed to confirm what was already known. The comparison also goes further than the standard advice, because it shows which conclusions from a study with a fixed order can still be trusted. That is useful because fixed elements of trial order are still used, sometimes deliberately. In several eHMI studies, the condition without an eHMI was always presented first (Dey et al., 2020b, 2021b), and in a recent study all participants rated a crossing scenario without an eHMI before the same scenario with an eHMI (Marcus et al., 2025). Our results suggest that such designs can support conclusions

about the average effect of a factor, but that results for the first conditions, and any change over the session, should be interpreted with particular care.

Data availability

The materials used in the study, analysis code and anonymised responses of the participants are available at https://www.dropbox.com/scl/fo/4pphss4roc9stcue0904o/AN76AaVvkSZHDkR4rW9_xvs?rlkey=sfnc6k4neajka793zxlw845wx.

Funding

This research received no external funding.

Declaration of competing interest

The authors declare no competing interests.

References

Alam, M.S., Dey, D., Martens, M.H., Bazilinskyy, P., 2026. Perceived crossing risk in multipedestrian encounters with automated vehicles: Effects of yielding, eHMIs, and pedestrian position. URL: https://shaadalam9.github.io/publications/alam2026multiped. manuscript in preparation.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Brooks, J.L., 2012. Counterbalancing for serial order carryover effects in experimental condition orders. Psychological Methods 17, 600–614. doi:10.1037/a0029310.

Camara, F., Dickinson, P., Fox, C., 2021. Evaluating pedestrian interaction preferences with a game theoretic autonomous vehicle in virtual reality. Transportation Research Part F: Traffic Psychology and Behaviour 78, 410–423. doi:10.1016/j.trf.2021.02.017.

Carlowitz, S., Madigan, R., Lee, Y.M., Tango, F., Merat, N., 2023. Pedestrians' perceptions of automated vehicle movements and light-based ehmis in real world conditions: A test track study. Transportation Research Part F: Traffic Psychology and Behaviour 95, 83–97. doi:10.1016/j.trf.2023.02.010.

Charness, G., Gneezy, U., Kuhn, M.A., 2012. Experimental methods: Between subject and within subject design. Journal of Economic Behavior & Organization 81, 1–8. doi:10.1016/j.jebo.2011.08.009.

Colley, M., Bajrovic, E., Rukzio, E., 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, pp. 1–11. doi:10.1145/3491102.3517571.

Colley, M., Kornmüller, D., Dey, D., Ju, W., Rukzio, E., 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, 1–33. doi:10.1145/3699778.

De Clercq, K., Dietrich, A., Núñez Velasco, J.P., De Winter, J., Happee, R., 2019. External human machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. Human Factors 61, 1353–1370. doi:10.1177/0018720819836343.

Deligianni, S.P., Quddus, M., Morris, A., Anvuur, A., Reed, S., 2017. Analyzing and modeling drivers' deceleration behavior from normal driving. Transportation Research Record: Journal of the Transportation Research Board 2663, 134–141. doi:10.3141/2663-17.

Dey, D., Habibovic, A., Berger, M., Bansal, D., Cuijpers, R.H., Martens, M.H., 2022. Investigating the need for explicit communication of non-yielding intent through a slow-pulsing light band (SPLB) eHMI in AV-pedestrian interaction, in: Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, ACM. pp. 307–318. doi:10.1145/3543174.3546086.

Dey, D., Habibovic, A., Löcken, A., Wintersberger, P., Pfleging, B., Riener, A., Martens, M., Terken, J., 2020a. Taming the ehmi jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces. Transportation Research Interdisciplinary Perspectives 7, 100174. doi:10.1016/j.trip.2020.100174.

Dey, D., Holländer, K., Berger, M., Eggen, B., Martens, M., Pfleging, B., Terken, J., 2020b. Distance-dependent eHMIs for the interaction between automated vehicles and pedestrians, in: Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, ACM. pp. 192–204. doi:10.1145/3409120.3410642.

Dey, D., Martens, M., Eggen, B., Terken, J., 2019. Pedestrian road-crossing willingness as a function of vehicle automation, external appearance, and driving behaviour. Transportation Research Part F: Traffic Psychology and Behaviour 65, 191–205. doi:10.1016/j.trf.2019.07.027.

Dey, D., Matviienko, A., Berger, M., Pfleging, B., Martens, M., Terken, J., 2021a. Communicating the intention of an automated vehicle to pedestrians: The contributions of eHMI and vehicle behavior. it - Information Technology 63, 123–141. doi:10.1515/itit-2020-0025.

Dey, D., van Vastenhoven, A., Cuijpers, R.H., Martens, M., Pfleging, B., 2021b. Towards scalable eHMIs: Designing for AV–VRU communication beyond one pedestrian, in: Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 274–286. doi:10.1145/3409118.3475129.

Dietrich, A., Tondera, M., Bengler, K., 2020. Automated vehicles in urban traffic: The effect of kinematics and ehmis on pedestrian crossing behavior. Advances in Transportation Studies 2, 73–84.

Faas, S.M., Mathis, L.A., Baumann, M., 2020. External HMI for self-driving vehicles: Which information shall be displayed? Transportation Research Part F: Traffic Psychology and Behaviour 68, 171–186. doi:10.1016/j.trf.2019.12.009.

Feng, Y., Duives, D., Daamen, W., Hoogendoorn, S., 2021. Data collection methods for studying pedestrian behaviour: A systematic review. Building and Environment 187, 107329. doi:10.1016/j.buildenv.2020.107329.

Fisher, R.A., 1966. The Design of Experiments. 8 ed., Oliver and Boyd, Edinburgh.

Greenwald, A.G., 1976. Within-subjects designs: To use or not to use? Psychological Bulletin 83, 314–320. doi:10.1037/0033-2909.83.2.314.

Habibovic, A., Lundgren, V.M., Andersson, J., Klingegård, M., Lagström, T., Sirkka, A., Fagerlönn, J., Edgren, C., Fredriksson, R., Krupenia, S., 2018. Communicating intent of automated vehicles to pedestrians. Frontiers in Psychology 9, 1336. doi:10.3389/fpsyg.2018.01336.

Hensch, A.C., Kreißig, I., Beggiato, M., Krems, J.F., 2022. The effect of ehmi malfunctions on younger and elderly pedestrians' trust and acceptance of automated vehicle communication signals. Frontiers in psychology 13, 866475. doi:10.3389/fpsyg.2022.866475.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Hübner, M., Stockmann, J.N.B., Bengler, K., 2025. Crossing the line: Impact of pedestrian group behavior on individual crossing decisions in AV interactions. Transportation Research Part F: Traffic Psychology and Behaviour 109, 921–937. doi:10.1016/j.trf.2025.01.020.

Jones, B., Kenward, M.G., 2003. Design and Analysis of Cross Over Trials. 2 ed., Chapman and Hall/CRC, New York. doi:10.1201/9781420036091.

Kaleefathullah, A.A., Merat, N., Lee, Y.M., Eisma, Y.B., Madigan, R., Garcia, J., De Winter, J., 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, 1070–1085. doi:10.1177/0018720820970751.

Lee, Y.M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., Camara, F., Rothmueller, M., Vendelbo-Larsen, S.A., Rasmussen, P.H., Dietrich, A., Nathanael, D., Portouli, V., Schieben, A., Merat, N., 2021. Road users rarely use explicit communication when interacting in today's traffic: Implications for automated vehicles. Cognition, Technology & Work 23, 367–380. doi:10.1007/s10111-020-00635-y.

Lee, Y.M., Madigan, R., Uzondu, C., Garcia, J., Romano, R., Markkula, G., Merat, N., 2022. Learning to interpret novel ehmi: The effect of vehicle kinematics and ehmi familiarity on pedestrian crossing behavior. Journal of safety research 80, 270–280. doi:10.1016/j.jsr.2021.12.010.

Lee, Y.M., Sidorov, V., Madigan, R., Garcia de Pedro, J., Markkula, G., Merat, N., 2025. Hello, is it me you're stopping for? the effect of external human machine interface familiarity on pedestrians' crossing behaviour in an ambiguous situation. Human Factors 67, 264–279. doi:10.1177/00187208241272070.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22. doi:10.1093/biomet/73.1.13.

Marcus, V., Pitts, G., Motamedi, S., 2025. Impact of eHMI on pedestrians' interactions with level-5 automated driving systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 69, 1250–1254. doi:10.1177/10711813251369865.

Montgomery, D.C., 2017. Design and Analysis of Experiments. 9 ed., Wiley, Hoboken, NJ.

Poulton, E.C., 1973. Unwanted range effects from using within subject experimental designs. Psychological Bulletin 80, 113–121. doi:10.1037/h0034731.

Rasouli, A., Tsotsos, J.K., 2019. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. IEEE Transactions on Intelligent Transportation Systems 21, 900–918. doi:10.1109/TITS.2019.2901817.

Shahi, S., Kumar Debnath, A., Birrell, S., Horan, B., Alatrash, S., Payre, W., 2026. Over time and under pressure: A longitudinal virtual reality study exploring the effects of time pressure and road environment on pedestrian-automated vehicles interactions. Transportation Research Part F: Traffic Psychology and Behaviour 119, 103604. doi:10.1016/j.trf.2026.103604.

Tran, T.T.M., Parker, C., Tomitsch, M., 2021. A review of virtual reality studies on autonomous vehicle–pedestrian interaction. IEEE Transactions on Human-Machine Systems 51, 641–652. doi:10.1109/THMS.2021.3107517.

Tran, T.T.M., Parker, C., Yu, X., Dey, D., Martens, M., Bazilinskyy, P., Tomitsch, M., 2024. Evaluating autonomous vehicle external communication using a multi-pedestrian VR simulator. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, 130:1–130:26. doi:10.1145/3678506.

Williams, E.J., 1949. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Scientific Research A 2, 149–168. doi:10.1071/CH9490149.

de Winter, J., Dodou, D., 2022. External human–machine interfaces: Gimmick or necessity? Transportation Research Interdisciplinary Perspectives 15, 100643. doi:10.1016/j.trip.2022.100643.