Tran-SET

Transportation Consortium of South-Central States

Solving Emerging Transportation Resiliency, Sustainability, and Economic Challenges through the Use of Innovative Materials and Construction Methods: From Research to Implementation

Autonomous Vehicle Communication Strategies Modeled in Virtual Reality

Project No. 20ITSUNM32

Lead University: University of New Mexico

Final Report

October 2021

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

The author would like to thank the Project Review Committee for their insight and guidance, including David Hadwiger, Michael Clamann, and Susan Bogus Halter.

TECHNICAL DOCUMENTATION PAGE

1. Project No.	2. Government Accession No.	3. Recipient's Catalog No.
20ITSUNM03		

4. Title and Subtitle	5. Report Date
Autonomous Vehicle Communication Strategies Modeled in Virtual Reality	Oct. 2021
	6. Performing Organization Code
7. Author(s)	**8. Performing Organization Report No.**
PI: Nicholas Ferenchak https://orcid.org/0000-0002-3766-9205	
9. Performing Organization Name and Address	**10. Work Unit No. (TRAIS)**
Transportation Consortium of South-Central States (Tran-SET) University Transportation Center for Region 6 3319 Patrick F. Taylor Hall, Louisiana State University, Baton Rouge, LA 70803	**11. Contract or Grant No.** 69A3551747106
12. Sponsoring Agency Name and Address	**13. Type of Report and Period Covered**
United States of America Department of Transportation Research and Innovative Technology Administration	Final Research Report Aug. 2020 – Oct. 2021
	14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded and accessible at Tran-SET's website (http://transet.lsu.edu/).

16. Abstract

We sought to better understand how autonomous vehicle (AV) communication strategies impact human road users' perceptions and behaviors. More specifically, we explored the impact of different external human-machine interface (eHMI) designs on understanding, task load, comfort, trust, acceptance, and reaction time. To accomplish this, we created virtual reality (VR) scenarios where human participants interacted with AVs. Participants experienced biking, driving, and pedestrian simulators and were brought back after initial testing to explore acclimation and learning effects. In terms of perceptions, the presence of an eHMI was the strongest predictor of understanding, comfort, trust, and acceptance outcomes in the statistical models when controlling for all other variables. There was a clear divide between text-based eHMIs and non-text eHMIs, with text-based eHMIs reporting better perception scores and the LED Windshield reporting the worst perception scores. There were perception acclimation effects detected (most notable for task load and comfort), but they had less of an impact than the presence of an eHMI. Perception outcomes had weaker relationships with participant characteristics than with AV characteristics. While behavioral outcomes should be interpreted with caution because of low participant sample sizes, behavioral results largely mirrored perception results in that significant reductions in reaction time were observed with the presence of an eHMI (3.69 second reduction), yielding (3.16 second reduction), and acclimation (0.134 second reduction per trial). Results suggest that eHMI design, AV behavior, and acclimation are most impactful in terms of both perceptions and reaction time.

17. Key Words	18. Distribution Statement
Vehicle automation; Communication; Human-automation interaction; Trust in automation; Technology acceptance	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	53	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^oF	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^oC
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^oC	Celsius	1.8C+32	Fahrenheit	^oF
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS, ABBREVIATIONS, AND SYMBOLS

AV	Autonomous Vehicle
CPAP	Continuous Positive Airways Pressure
eHMI	External Human-Machine Interface
IRB	Institutional Review Board
LED	Light-Emitting Diode
NASA	National Aeronautics and Space Administration
NHTSA	National Highway Traffic Safety Administration
OLED	Organic Light-Emitting Diode
PRT	Perception-Reaction Time
SAHS	Sleep Apnoea Hypopnoea Syndrome
SDK	Software Development Kit
SDLP	Standard Deviation of Lane Position
SSQ	Simulator Sickness Questionnaire
TLX	Task Load Index
ULP	Ultra-Low Power
USB	Universal Serial Bus
VR	Virtual Reality
VRSQ	Virtual Reality Sickness Questionnaire

EXECUTIVE SUMMARY

We sought to better understand how autonomous vehicle (AV) communication strategies impact human road users' perceptions and behaviors. More specifically, we explored the impact of different external human-machine interface (eHMI) designs on understanding, task load, comfort, trust, acceptance, and reaction time. To accomplish this, we created virtual reality (VR) scenarios where human participants interacted with AVs. Participants experienced biking, driving, and pedestrian simulators and were brought back after initial testing to explore acclimation and learning effects.

The presence of an eHMI was the strongest predictor of understanding, comfort, trust, and acceptance outcomes in the statistical models when controlling for all other variables. Task load was also improved by eHMIs but did not reach statistical significance (likely because eHMIs took some effort for interpretation).

In terms of which eHMI design was correlated with the best perceptions, there is a clear divide between text-based eHMIs and non-text eHMIs, with text-based eHMIs reporting better perception scores. The Text on Grille and Text on Driver Door eHMIs typically received the best scores while LED Windshield generally received the worst scores. This supports the need for future research on the positioning of the eHMIs as the Text on Driver Door eHMI was on the side of the AV.

There were perception acclimation effects detected, but they had less of an impact than the presence of an eHMI. The largest acclimation effects were in comfort (which increased throughout testing for all eHMI configurations) and task load (which decreased throughout testing for all eHMI configurations), while trust and acceptance were more stable throughout testing (average increases of only 0.1 points and 0.2 points from beginning to end, respectively). This suggests that while task load and comfort improve as participants experience AVs regardless of eHMI configuration, trust and acceptance are much more dependent upon appropriate eHMIs.

All the measured perceptions improved for the driving and pedestrian simulators but stayed relatively stable or even got worse for the biking simulator. However, this may be because of the design of the simulators themselves. The mode of the simulator was non-significant in the understanding, comfort, and acceptance statistical models and was relatively weak in the others, suggesting that mode is not as important as eHMI presence, acclimation, or yielding.

Importantly, perception outcomes had weaker relationships with participant characteristics than with AV characteristics. This suggests that while factors such as age, gender, and preconceived biases toward or against AVs will impact road users' preferences, appropriate eHMI design, AV behavior, and acclimation may be able to help overcome any negative perceptions.

While behavioral outcomes should be interpreted with caution because of low participant sample sizes, behavioral results largely mirrored perception results in that significant reductions in reaction time were observed with the presence of an eHMI (3.69 second reduction), yielding (3.16 second reduction), and acclimation (0.134 second reduction per trial). The LED Windshield had significantly longer reaction times than the other eHMIs and a participants' gender and initial reported trust of AVs had smaller impacts on reaction time. Results suggest that eHMI design, AV behavior, and acclimation are most impactful in terms of both reaction time and perceptions.

1. INTRODUCTION

Autonomous vehicles (AVs) may improve transportation efficiency, safety, equity, and environmental impacts *(1-3)*. Before these benefits can be realized, the technology must be adequately adopted. For at least a few fleet generations, there will likely be a transition period in which both human-driven vehicles and AVs share the road. In addition to human drivers, we can expect pedestrians and bicyclists to continue to share street space with AVs. How will AVs and human users of the road interact and communicate, and how will this impact perception and behavior outcomes?

Human perceptions of AVs are important because trust of these new technologies is integral to their successful integration into the transportation system *(4-8)*. While recent research shows that many people are still hesitant to share the road with AVs *(9)*, improved AV communication has been linked with increased trust and acceptance of the new technology *(10-11)*. External human-machine interfaces (eHMIs) that allow AVs to communicate with human road users may play an important role in that communication *(12)* (Figure 1). How do human drivers', pedestrians', and bicyclists' perceptions of AVs (e.g. understanding, trust, comfort, and acceptance) vary relative to eHMI design?

Figure 1. Examples of AV eHMIs proposed by AV developers.

Understanding behavior in the interaction of humans and AVs is important for two types of outcomes: safety and operations. In terms of safety, the National Highway Traffic Safety Administration (NHTSA) attributes approximately 94% of motor vehicle collisions to human error *(13)*. AVs present an incredible opportunity to improve traffic safety outcomes by reducing or eliminating this human error. However, to realize those benefits, we need to ensure that human road users understand AVs' intentions and actions and can safely interact with the new technology.

In terms of operations, traffic engineers use microsimulation software to model traffic flows on our streets. This software is built upon decades of research into the interactions and behaviors of different system users. But with a new type of user on the street, how will these assumptions change? What if it takes an extra second for a pedestrian to interact with an AV versus a traditional human-driven vehicle? What if human drivers require an extra second of gap acceptance with AVs? While seemingly small on the individual scale, these changes can have outsized implications for the operations of a transportation system *(14)*.

To understand the above research questions, we had human participants interact with AVs in a virtual reality (VR) environment (Figure 2). Participants acted as drivers, pedestrians, and bicyclists. The AVs had a variety of communication eHMIs based on those designed and proposed by AV companies (Figure 1). We measured participant behavior through movement tracking during the VR trials to understand participant perception-reaction time (PRT). We administered a survey after the trials to understand how well participants understood and how much they trusted the different AV communication strategies. This allowed us to understand how to improve perceptions, trust, and acceptance, if/how human behavior changed, and which communication strategy was most effective. As there is currently no standard AV communication strategy, this is a vitally important and timely topic.

Figure 2. Example of VR environment for pedestrians with AV and eHMIs.

In order to test learning effects, we also collected longitudinal repeated measures on participants. Between one to three days after the initial round of testing, participants returned to the lab and experienced the same AV eHMIs again. Comparing initial perceptions and behaviors to those from the second round of testing allowed us to understand how human road users acclimated to the new technology and whether their preferences changed as they became more familiar with the AVs.

AVs may provide many benefits. However, in order to ensure these benefits are realized, we need to make sure that there is understanding, trust, and acceptance of these new technologies and that the new technologies can allow for efficient operations within our transportation systems.

2. OBJECTIVES

The objective of this research was to identify and evaluate eHMI strategies for AV communication with human road users (i.e. human drivers, pedestrians, and bicyclists). We hope that this research will help build our understanding and allow us to work toward the development of a single, uniform AV-human communication strategy.

Testing occurred in fully-immersive three-dimensional VR environments in which participants encountered AVs at typical urban or suburban intersections. There was a single AV in each trial and that AV had a single eHMI (or no eHMI in some instances). Each participant encountered each eHMI design in each of the driving simulator, pedestrian simulator, and bicycle simulator. The paths of the human users and AVs crossed and rights-of-way were ambiguous so that participants were unsure whether or not the AV would yield to them. The AV did not yield to the participant for one trial out of each five trials. In that way, the participants needed to read and understand the eHMIs to understand whether or not they should proceed.

The AVs possessed various eHMIs based on designs currently proposed by AV developers. The lighting, color, sequence, text, and noise interfaces were held consistent to control for those factors. Only changes to the overall eHMI design were studied. Participants returned one to three days after the initial testing to longitudinally examine learning effects.

Outcomes were grouped into two categories: perceptions and behaviors. Perceptions included understanding, trust, comfort, and acceptance. Understanding was measured by pausing the VR experiment upon the participant's interaction with the AV. While paused, the lab technician asked the participant whether they believed the AV would or would not stop for them. This provided us with an accurate representation of whether the participant understood the AV's action, as opposed to waiting until after the simulation was complete, which would bias the participant's answer. After each trial, we further garnered the participant's perceptions through a post-experiment survey. Trust was measured through the Scale of Trust in Automated Systems, a validated and widely-used survey instrument *(15)*. Comfort was measured through another widely-used survey instrument, this one developed to measure the likeability and perceived safety of robots *(16)*. Acceptance was measured through the Assessment of Acceptance of Advanced Transport Telematics, another validated and widely-used survey instrument *(17)*.

Behaviors were measured using video captures of the trials to measure users' body movements and to better understand human users' interaction time while interacting with AVs. We primarily tracked response time, as that was the most direct measurement of the effectiveness of the AV interaction *(18)*.

We also administered the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) to each participant after each trial *(19)*. This measured the mental and physical workload associated with the interaction with the AV. We also administered a VR simulator sickness survey, the Virtual Reality Sickness Questionnaire (VRSQ), to ensure that none of the perceptions or behaviors were biased because of an uncomfortable disposition *(20)*.

3. LITERATURE REVIEW

We cover several topics in this literature review. First, we provide a summary of current research exploring AV communication. After research gaps have been identified, we then discuss outcome metrics that might be measured with this work. Next, we discuss aspects of the experiments including a history of VR simulators, considerations for trial ordering, and possible longitudinal timing of the trials. Finally, we provide a discussion of VR sickness, which proved to be relevant in our research.

3.1. AV Communication

While AV acceptance is important for the new technology, recent research shows that many people are still hesitant of sharing the road with AVs *(9)*. In a recent survey, researchers found that 56% of mode choice decisions favored AVs over regular vehicles *(21)*. In another survey, only 41% of survey respondents were willing to use shared AVs at least once a week at $1/mile and only 15% were willing at $2/mile *(22)*. Another recent survey found that only about 50% of respondents from the United States intend to use AVs *(23)*.

While public acceptance is relatively low, improved AV communication has been linked with increased trust and acceptance *(10, 11)*. AV developers have begun to generate a variety of eHMI concepts to accomplish AV-to-human communication (Figure 1), but little research exists regarding the most effective form of communication *(24-28)*.

Research has found that road users prefer the presence of an eHMI to communicate an AV's intentions over the absence of eHMIs *(12)*. eHMIs have been shown to improve pedestrians' ability to recognize an AV's intentions when the AV is yielding to a pedestrian *(29, 30)*. However, when a pedestrian is interacting with an AV that is not yielding, eHMIs have been found to either not improve or even decrease pedestrians' ability to recognize the AV's intentions *(29, 30)*.

The most effective eHMI configuration is still nebulous. In a crowdsourced survey of photos and videos, pedestrian respondents reported that text messages were clearest while often mistaking lights for sensors *(31)*. In a VR experiment, text messages were preferred by pedestrian participants *(29)* while participants in a video simulation especially preferred large-scale text *(32)*. Text also required no initial learning unlike the other eHMIs tested (front brake lights, Knightrider animation, and an illuminated smiley) *(29)*. Egocentric messages that told the pedestrian to "Walk" (as opposed to allocentric messages such as those that communicate that the AV "Will Stop") were responded to most effectively *(31, 32)*. While text was preferred in the three studies above, the studies did not examine all currently proposed eHMIs and liability, legibility, and technical issues of text interfaces were identified *(31)*. Also, there is conflicting research showing that pedestrians in Europe generally prefer conventional lights to text-based messages *(33)*.

Other researchers have found that lights on AVs may be confusing to pedestrians *(34)*. Participants in a web-based survey largely associated green lights with a moving vehicle and red lights with a slowing down or stopped vehicle *(34)*, while other researchers alternatively found that pedestrians interpreted a green display as permission for them to have the right of way *(31)*. When eHMIs are confusing, research suggests that pedestrians will rely on legacy behaviors such as gap acceptance and braking rather than leverage the information on an external display *(35)*. Since there is generally more pedestrian visibility on the side of the vehicle, other research has identified the

importance of having side-mounted eHMIs, whereas most of the above research assumed front-mounted eHMIs *(36)*.

While text displays appear to be preferred, there are still important questions to be answered. Which eHMI designs are effective in terms of understanding, trust, comfort, and acceptance? Preliminary analysis suggests that understanding, trust, and comfort in similar ways by AV eHMIs, but acceptance follows a different pattern (*37*). To advance the body of knowledge on this topic, we test currently-proposed eHMI designs relative to all the aforementioned perception outcomes. Also, while past research suggests that perceived risk and general attitudes toward AVs vary by gender *(8, 38)*, no existing interactive laboratory-based AV eHMI experiments explore such differences. Furthermore, past research has identified that varying interest in AVs may bias perceptions *(9, 21-23)*. We therefore test the impact of gender and account for preexisting perceptions of AVs in our statistical models, a novel contribution. Also, since no currently published research examines the learning effects of AV eHMIs, we longitudinally test perceptions and behaviors through repeated measures. Since it is necessary to develop a single uniform AV communication strategy that can be readily understood by all pedestrians *(12)*, such holistic research is vitally important.

3.2. Outcome Measurements

As stated previously, outcome measures can generally be categorized into perceptions and behaviors. It is important to note that stated perceptions and intentions do not always align with behaviors (*39-44*).

The first perception measured was understanding. We sought to measure whether the participants understood whether or not the AV would yield to them and what impact the eHMI design had on that understanding. Instead of asking the participants after each trial (at which time they would already know the intentions of the AV), we paused each trial once the participant came into contact with the AV and asked the participant (while they were still in the paused VR environment) whether they understood the AV's intentions. This approach of pausing the VR environment to measure participant understanding has been used by past researchers (*45, 46*).

After the completion of each trial, we further garnered the participant's perceptions through a post-experiment survey. Trust was measured through the Scale of Trust in Automated Systems, a validated and widely-used survey instrument (*15*). This survey was developed for the Air Force Research Laboratory in 1998 to measure the extent to which people trust that automated systems will perform effectively. In developing their survey instrument, the researchers performed a three-phase experiment consisting of a word elicitation study, a questionnaire study, and a paired comparison study to better understand the relationship between trust and distrust. Trust and distrust were found to be opposites, rather than comprising of different concepts. Trust was also found to be similar for human-human trust, human-machine trust, and trust in general. Twelve factors of trust between people and automated systems were identified and compiled into a survey instrument, as detailed in the Methodology section of this report.

Comfort was measured through another widely-used survey instrument, this one developed to measure interactions with robots (*16*). In actuality, the survey instrument was more widely developed to measure the perceived safety of interactions with robots, which consists of the perceptions of danger when interacting with robots and the perceptions of comfort when interacting with robots. The researchers compile several past pieces of research that used the

survey instrument and find that the validity of the questionnaire may be assumed. The researchers note that positive perceptions of safety and comfort are a key requirement if such automated devices are to be accepted. Again, the exact survey instrument used will be detailed in the Methodology section of this report.

Acceptance was measured through the Assessment of Acceptance of Advanced Transport Telematics, another validated and widely-used survey instrument (*17*). The researchers sought to develop a standard methodology of measuring drivers' acceptance of new technologies. The questionnaire consists of two scales: a scale denoting the usefulness of the new technology and a scale denoting the satisfaction of using the new technology. The questionnaire was used in six different studies (sample size of 147) and found to be reliable. Results were sensitive to users' differences in opinion on technology, which we account for in our own study. Again, the exact survey instrument used will be detailed in the Methodology section of this report.

The NASA-TLX is often used in driving simulation research to identify the factors associated with variations in subjective workload (*47-51*). NASA-TLX was developed by Hart and Staveland through a multi-year research program that consisted of more than forty laboratory simulations and is widely used in the field of human factors (*19*). The survey measures subjective workload in terms of mental demand, physical demand, temporal demand, overall performance, effort, and frustration level. Using the NASA-TLX may be useful when understanding the effectiveness of different treatments in terms of these metrics. Again, the exact survey instrument used will be detailed in the Methodology section of this report.

The most applicable behavioral metric for the current study was PRT. PRT can be defined as the time required to perceive, interpret, decide, and initiate a response to some stimulus (*52*). It is a fundamental measurement in transportation engineering, important for both operational analysis of traffic conditions and geometric design of roadways. Past research has shown that PRT may vary with the introduction of autonomous vehicles (*53-55*). We hope to examine gap acceptance in the future when we get to more complex environments with multiple AVs. Understanding the impacts no gap acceptance in future work will be important for determining possible changes to intersection sight distance requirements. Also, we hope to use eye tracking technology in the VR headsets to further understand how sight focus is impacted by eHMI design in future research.

We also hope to eventually measure the impact of eHMIs on safety outcomes for human road users in future research. While we believe that the perceptions and behaviors detailed above will represent a proxy for safety outcomes, understanding safety in an operational environment will constitute important future research. With a nearly 50% increase in pedestrian fatalities across the U.S. over the last decade and concerns over trends in bicyclist safety (*56, 57*), examining safety impacts on vulnerable road users will constitute an important aspect of this future research. We hope to examine collisions as well as proactively consider perceived safety's impact on travel behavior (*58, 59*).

3.3. VR Simulator Reliability and Validity

There exists a wide body of driving simulator research that spans several decades. Much of the early work in driving simulators involved simulators that utilized monitors spanning a 180-degree front field of view synchronized to display a realistic view of a computer-generated road environment. We shall discuss validation and reliability findings from these monitor-based

simulators – for which a much larger body of literature exists – before discussing headset-based VR simulator validation and reliability.

There are two primary measures when studying whether simulators accurately and consistently reflect actual roadway behaviors: reliability and validity. Reliability is a measure of whether the simulator reports consistent results over time. Validity is whether the simulator accurately represents real-world road behaviors. Metrics that are commonly measured and tested (especially for driving simulators) include mean speed, speed variability, lateral position, overall driving performance, and number of driving errors.

Although we found little research testing the reliability of driving simulators, that research has generally confirmed that driving simulators are reliable. In a study measuring standard deviation of lane position (SDLP) for sixty monitor-based driving simulator participants (29 were retested after three months, 31 were retested after a year or more), results suggested that simulator-measured SDLP was a reliable measure for periods ranging from months to years (*60*). Twenty participants in Japan were similarly tested and retested in periods ranging from sixty minutes to one-week intervals (*61*). The researchers found that the simulator-based results were reliable and remained relatively consistent over time. The researchers concluded that driving simulators may serve as a useful tool in tracking the effects of factors such as inebriation, neurologic disorders, and pharmacologic treatments on driving abilities over time.

When examining validity of simulator measurements, there are two forms that can be tested: absolute validity and relative validity. Absolute validity occurs when the values obtained in a simulator match those obtained in a real vehicle in absolute terms (*62*). Proving absolute validity requires the comparison of driving simulator measurements to real-world roadway measurements, typically with instrumented motor vehicles on actual roadways. Relative validity occurs when simulator measurements show the same patterns as real-world driving, although not matching in absolute terms. For instance, relative validity would be established if participants in both simulator and real-world experiments had lower vehicle speeds after a treatment installation, even if the participants in simulator were statistically slower or faster overall.

A systematic review of 44 driving simulator validation studies – all of which compared simulated driving behavior to on-road driving behavior in an actual motor vehicle – found that approximately half of the studies achieved absolute or relative validity, whereas one third produced nonvalid results (*62*). The remaining studies presented both valid and invalid results, suggesting that the simulators in question were not universally valid. The researchers were not able to identify strong patterns in valid versus nonvalid results. Some studies with high fidelity (simulations judged as highly realistic) were found to be valid while others were nonvalid; some studies with low fidelity were similarly were found to be valid while others were nonvalid. Similarly, some studies that measured vehicle speed were found to be valid while others were nonvalid (the same held for studies measuring vehicle positioning and driver error) (*62*). Although they were not able to find consistent patterns in validity versus nonvalidity, the researchers did note that speed and driver errors tended to be higher in the simulators (*63*). This may be because the level of perceived risk is lower in simulators or because of unfamiliarity with the simulated conditions (*64*).

While the papers used in the above systematic review were generally older (1990's and early 2000's), more recent driving simulator validation studies we found generally exhibited validity. For 47 drivers in Australia, there was no statistical difference between the on-road assessment and the driving simulator for mirror checking, left, right and forward observations, speed at

intersections, maintaining speed, obeying traffic lights and stop signs (*65*). Underwood et al. found that when comparing participants' behaviors in three situations (i.e. actual driving on roads, watching video clips, and a driving simulator), there was increased visual scanning by more experienced and especially professional drivers, and earlier eye fixations on hazardous objects for experienced drivers in all three situations (*66*). For 44 participants in Australia, participants decelerated when encountering rumble strips in similar ways in both an instrumented car and a simulator, establishing the relative validity. However, absolute validity was not established as participants generally drove faster in the instrumented car than in the simulator (*67*).

In summary, there appears to be a strong body of literature establishing validity (and to a lesser extent establishing reliability) of driving simulators. However, that validation was not universal, and past researchers were not able to identify determining factors of validity versus nonvalidity. It is important to reiterate that all of the above pieces of research appear to be from monitor-based driving simulators.

The only example we found of validation research for a monitor-based bicycling simulator was from Gustave Eiffel University in France (*68*). 36 participants rode a bicycle simulator for around 600 meters with full control of the handlebar, pedals, and brakes. The participants' subjective evaluation of the realism of the simulator was ranked as a 6.74 out of 10. While the researchers called this measure "subjective validity", we believe it was more a measure of fidelity. Furthermore, the researchers did not measure speed, positioning, or other behavioral measurements.

In the only example that we could find of validation research for a monitor-based pedestrian simulator, 102 children and 74 adults completed simulated road-crossings in both the virtual simulated environment and the identical real environment (*69*). In both the child and adult samples, validity was established via significant correlations between behavior (gap size and wait time) in the virtual and real scenarios. The researchers concluded by stating that the findings suggest pedestrian simulators can be used as a valid tool to understand pedestrian perceptions and behaviors.

We hypothesize that with the increased fidelity and immersivity of headset-based simulators (what we refer to as VR), reliability and validity would be improved over that of monitor-based simulators. While there is a relatively well-established body of research validating monitor-based driving simulators, we were only able to find two pieces of validation research on headset-based VR driving simulators. Results of the first research indicated that perceived speed, distance headway, and physiological data were significantly similar between the driving simulator and the real-world tests (*70*). However, larger decelerations were observed in the simulator. The second driving validation suggested that while there were relatively high levels of validity (especially relative), participants had difficulties holding a given velocity and showed less risk awareness in the simulator (*71*). The researchers also compared simulators that had both static and dynamic bases, finding that participants accepted the dynamic simulations as more realistic than the static simulations.

Regarding headset-based VR bicycling simulator validation, we found one piece of peer-reviewed research. 26 participants – ranging in age from 18 to 35 years – rode an instrumented bicycle on an actual road environment as well as a headset-based VR bicycle simulator (*72*). Absolute validity was established between the two scenarios for average lane position, deviation in lane position, and average passing distance from parked cars. Relative validity was established for the average

speed of cyclists and their speed reduction on approach to intersections. The researchers concluded that while more research is needed for a thorough validation of all behaviors, the evidence suggested that bicyclist behavior can be effectively investigated using headset-based VR.

We also found a single example of headset-based VR pedestrian simulator validation (*73*). Participants experienced equivalent real and virtual environments and performed similar tasks in each, including crossing an intersection and estimating the speed and distance of an approaching vehicle. Results showed no statistical differences between the real and virtual environments for participants' intention to cross, estimation of distances, and perceptions of safety and risk. Statistically significant differences between real and virtual environments were observed in the estimation of speed. The researchers concluded that at lower vehicle speeds (25 mph and lower), headset-based VR can be used as tool to evaluate pedestrian behavior.

3.4. Trial Ordering

The goal of this project is to explore whether various eHMIs have an effect on perception and behavioral outcomes. To do so, simulator participants will encounter a control scenario without any eHMIs and scenarios with varying eHMIs. There are two experimental approaches for accomplishing this goal: within-subject and between-subject. A within-subject experiment has each subject experiences more than one – and typically all – of the possible scenarios (control and all treatments) (*74*). A between-subject experiment has each subject experience only one scenario. Both designs have been found to be legitimate given the context of the question being studied and in terms of the practical implementation of the research study (*74*).

Within-subject analyses have been found to have three primary advantages. First, since each participant will experience each scenario, random assignment of participants does not present an issue. In other words, there is higher risk for between-subject experiments that participant bias may weigh heavier for one particular scenario. This is not an issue for within-subject experiments as any possible participant bias will be spread evenly across all scenarios.

Second, within-subject design often provides significantly more statistical power. Because each participant is exposed to multiple scenarios, trial sample sizes are much higher for within-subject experiments. For between-subject experiments, if each participant is exposed to only one scenario and the researchers desire a similar level of statistical power, the number of participants will need to be greatly increased. This is typically costly in terms of time and money.

Third, within-subject experiments are typically more aligned with the context of the research question and the desired practical implications of the research. In other words, we may very well expect that a participant would experience multiple AVs – and therefore multiple treatments – as they travel around in the real world, and therefore a within-subject design makes sense. The between-subject assumption that each participant will experience one and only one AV does not necessarily align with our expectations of reality.

A primary disadvantage to within-subject design is the demand effect. This occurs when a participant – as they undergo trials in the experiment – begins to interpret the researchers' intentions and changes their behavior accordingly, either consciously or not (*75, 76*). Because participants only experience a single scenario in between-subject experiments, this demand effect is not as significant of an issue for between-subject designs. Within-subject experiments may also lead to learning effects, where a participant becomes acclimated to a scenario after experiencing it

initially. For example, if we were measuring the speed at which a driver participant operates, we may very well expect that a participant's speed would be slowest during the first trial, and subsequently increase as they became accustomed to the environment, regardless of what treatments were present. However, there are methods of combating these learning effects, as we will see below.

We are leaning toward a within-subject design because we understand that our participant sample size will be relatively small. Therefore, exposing each participant to each scenario makes sense for our particular experiment. In addition, the theoretical approach – that each participant will experience multiple AVs in the real world – makes logical sense. We will take steps to minimize demand effects and learning effects.

In our review of pertinent literature, we found abundant examples of researchers using within-subject designs for driving simulator studies (*77-80*). Donmez et al. found that this is appropriate because if a researcher seeks to primarily understand the effect of a between-subject variable (such as age), then a between-subject design is fitting (*81*). However, if a researcher is concerned with the effect of a treatment (as many researchers are, including ourselves), a within-subject design may be best. Accordingly, between-subject driving simulator experiments were more difficult to find (*82, 83*).

While we believe that a within-subject design is appropriate for this study, we mentioned that a primary issue is the demand effect and learning effect that occur as participants experience the scenarios. Counterbalancing the scenarios can be used as a tool to combat these effects. Counterbalancing is a form of trial ordering that is employed to avoid order and sequence effects. Specifically, counterbalancing seeks to ensure that over all the participants tested, each scenario appears in each position (first trial, second trial, third trial, etc.) an equal number of times. In other words, if every participant experienced a certain eHMI during their first simulator trial, the results would be biased for that eHMI because of the novelty of that first trial. That specific eHMI should be experienced during the first trial, second trial, third trial, and fourth trial equally. Counterbalancing minimizes this bias.

In addition to counterbalancing the scenarios, we plan to provide participants with practice trials for each simulator so the participants can become accustomed and habituated to the simulators. This will relieve the original simulator novelty and hopefully reduce bias during the actual experiment trials.

Within-subject driving simulator studies frequently use counterbalancing of their trials. Every within-subject driving simulator study cited above had counterbalanced trials and based on our examination, every other within-subject study we noted throughout this literature review employed counterbalancing (*77-80*). For instance, Godley et al. validated driving simulator results against real world measurements from an instrumented vehicle (*67*). Each participant in the simulator encountered both control and treatment scenarios and each participant in the instrumented vehicle encountered both control and treatment scenarios. To counterbalance, some participants were shown the control scenario first and some participants were shown the treatment scenario first. Similarly, Gershon et al. administered a driving simulation with a within-subject design where every subject drove with and without an interactive cognitive task (ICT) assigned to them (*84*). The with and without trials were counterbalanced and the ICT was found to increase driver alertness.

3.5. Longitudinal Timing

In our experiment, we seek to obtain longitudinal repeated measures on participants to understand whether learning effects play a role in perception and behavior outcomes as participants become acclimated to the AV eHMIs. How much time should elapse between tests? Longitudinal simulator studies exploring roadway or vehicle characteristics were surprisingly rare. Researchers more often counterbalance to account for learning effects instead of retesting over time. In other words, if a participant were to drive through a roadway scenario four times with a unique eHMI each time, we would want to offset the order in which the participants see the treatments because by the fourth trial, participants' behaviors may be biased by acclimation to the scenario or by testing fatigue. Such counterbalancing seeks to negate this acclimation bias, whereas we wish to measure the acclimation bias.

Our main finding in terms of temporal spacing of longitudinal studies was that, in general, researchers should retest participants as often as the researchers would expect the participants to encounter the treatment in the real world. In our experiment, we may reasonably expect that AVs have penetrated the vehicle market and participants would therefore experience AVs often. In this case, we may wish to retest participants on subsequent days. However, if we assume that AVs are still relatively rare, we may want to retest at longer intervals such as subsequent weeks or months.

The most applicable study that we were able to identify longitudinally examined driver behavior changes in terms of alternative vehicle designs. Large et al. used a driving simulator to study drivers' behaviors when in an AV (*85*). Because the researchers were assuming that the trials would approximate a typical commute, they brought participants back for retesting over five successive weekdays. Findings suggested that participants were consistently comfortable and trusting of the new technology and were highly willing to allow the AV to take control so they could pursue their own non-driving activities.

We identified a similar study that used a driving simulator to explore the impact of traffic calming gates on vehicle speeds (*86*). Seventeen participants were brought into the lab and tested on each of five successive days. In each test, all participants were exposed to both a scenario with gates and a scenario without gates for a within-subject design. The researchers found a consistent reduction in vehicle speeds in the presence of the gates, and this speed reduction continued over the five days of testing. This was the only piece of research we found that explored behavior changes over time in response to a roadway design treatment through simulation.

We found two other pieces of research that similarly collected longitudinal repeated measures in simulator settings, although these papers did not report the temporal learning effects. Calvi tested the effectiveness of pavement markings placed on crest vertical curves intended to slow drivers (*87*). He tested participants in a simulator four times for each scenario to examine the learning curve. These four trials were spread over a single session. While the researcher reported that the pavement markings were effective at reducing speeds, he did not report whether this effectiveness varied over time. Similarly, Awan et al. found that pavement markings slowed vehicles around horizontal curves (*88*). The researchers administered two simulator trials in the same session, only using the second trial because of the novelty of the first. These short time intervals appear to function to negate the learning effect rather than measure it.

In addition to the relatively small body of research that has used simulators to test for differences in driving behavior over time relative to a vehicle or roadway design treatment, there is a more

significant body of literature that tests for driving behavior changes relative to factors such as disease progression and treatment. For instance, Turkington et al. used a driving simulator to test 36 participants with severe cases of sleep apnoea hypopnoea syndrome (SAHS) (*89*). Eighteen of the patients underwent treatment with continuous positive airways pressure (CPAP) and eighteen of the participants served as untreated controls. All participants were tested at baseline (before the treatment), at days 1, 3, and 7 of the CPAP trial period, and finally days 1, 3, and 7 after the treatment ended. Differences in driving behavior (improved reaction times and reduced errors) were detected for the treatment group within a week of treatment, and those differences lasted for at least a week.

Similarly, Duchek et al. examined patients with and without Alzheimer's disease to understand the effects of the disease on driving behavior (*90*). Participants drove on an actual road with an instructor observing their behaviors and were brought back in 6-month intervals over several years. Declines in driving performance were noted with disease progression. The large time intervals were appropriate given the temporal progression of the disease.

Simulator reliability testing – where simulator measurements are taken over certain time periods to ensure consistency – had a wide array of retesting intervals ranging from sixty minutes to over a year (*60*, *61*).

In summary, retesting intervals can range from minutes to years depending on the temporal patterns of the exposure, learning effects, or disease effects. For this AV eHMI study, we recommend a retesting interval of a few days because that aligns with the frequency of AV interactions that we would expect in an actual roadway environment.

3.6. VR Sickness

It is also important to monitor simulator sickness while participants are in the VR scenarios (*91-94*). Not only may feelings of sickness alter results and compromise validity, but sickness may put participants at risk of losing their balance and subsequent injury – especially for the bicycle and pedestrian simulators. Several surveys have been created to consistently measure simulator sickness. The one that we typically use was developed by Kim et al. and is called the virtual reality sickness questionnaire (VRSQ) (*20*). It is based on the widely-used simulator sickness questionnaire (SSQ), but modified explicitly for VR (*95*). Specifically, the SSQ consists of 16 items that fall into the three components of nausea, oculomotor, and disorientation. Based on testing of 24 participants in a headset-based VR environment, the researchers found that the items related to nausea were weak compared to those related to oculomotor and disorientation. The VRSQ therefore consists of just nine items related to oculomotor and disorientation. These items are general discomfort, fatigue, eyestrain, difficulty focusing, headache, fullness of head, blurred vision, dizzy (eyes closed), and vertigo. Participants rank their experience of each item on a four-point Likert scale (0 = not at all, 1 = slightly, 2 = moderately, and 3 = very). We plan to administer the VRSQ to our participants throughout the testing to ensure that their behaviors and perceptions are not biased by simulator sickness.

4. METHODOLOGY

We first discuss the VR setup including hardware, software, simulators, and scenario design. We then discuss the participant testing including trial ordering, survey instruments, behavioral measurements, and longitudinal timing. We lastly discuss the statistical data analysis methodologies.

4.1. VR Hardware and Software

For each VR simulator (i.e. pedestrian, bicycle, and driving), we used an HTC Vive Pro VR System (Figure 3). Each system included a headset, two controllers, and two base stations. The headsets' dual-OLED (organic light-emitting diode) displays with resolution of 2880 x 1600 pixels provided high levels of visual fidelity. The base stations were capable of sub-millimeter tracking accuracy. We installed the base stations on light stands at a height of 6.5 feet.

Figure 3. HTC Vive Pro VR System with headset in the middle, two controllers on the bottom, and two base stations on the top.

We used two separate pieces of software to render the scenarios. The pedestrian scenario was built with Unreal Engine 4. Unreal Engine 4 generally provides higher visual fidelity. However, the drawback is there are limited pre-made plug-ins for the software. Given that the pedestrian scenario was the simplest (there were no additional hardware other than the VR system to link to the scenario), Unreal Engine 4 was optimal for the pedestrian scenario.

We used Unity for the bicycle and driving scenarios. In general, this game engine provides less visual fidelity but more functionality. Given that we needed to link a bicycle simulator and a driving simulator into the VR scenario, Unity provided an easier pathway to do so, as detailed below in Section 4.2. Although we utilized two separate game engines, we were able to closely match the scenarios and AV designs between the two engines.

4.2. Simulators

The pedestrian simulator was relatively straightforward. The only piece of hardware was the VR system. We setup the room-scale play area at its maximum allowances of 11.5 feet x 11.5 feet (Figure 4). The pedestrian participants were able to freely walk around the entire play area square.

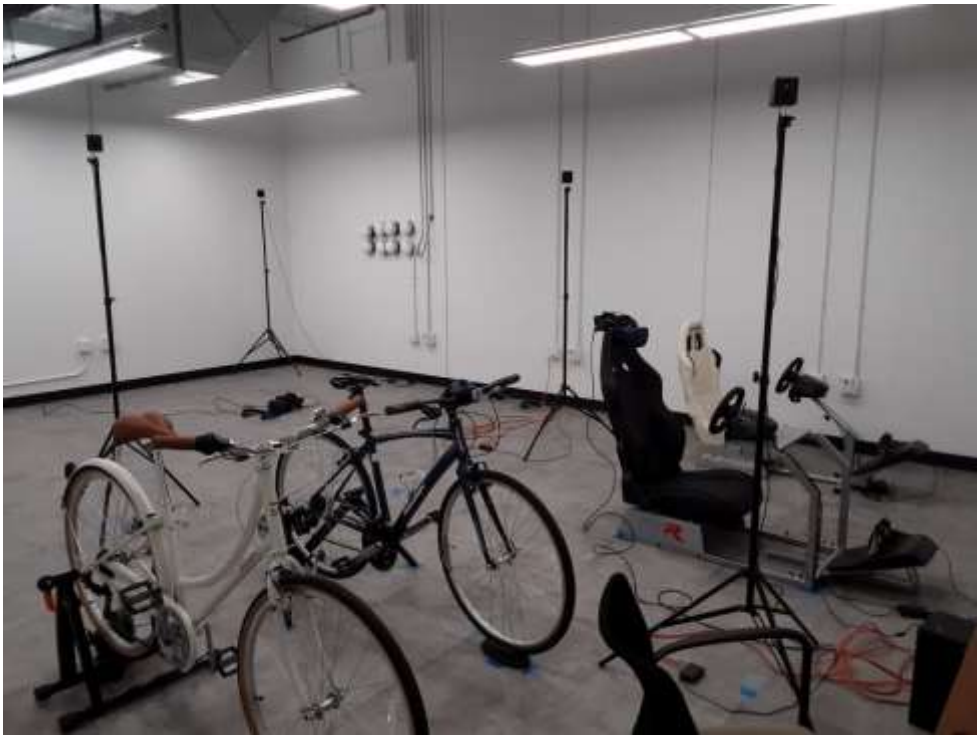

Figure 4. Pedestrian simulator play area in the background behind the bike and driving simulators.

The bicycle simulator consisted of a small women's bike with an adjustable seat (Figure 5). The adjustable height of the seat allowed for participants across a wide range of heights to participate. The rear wheel of the bike was set into a Saris Fluid^2 Trainer. This trainer provided for smooth operation of the bike and resistance was adjusted by shifting the bicycle's gears, providing a realistic riding experience. The front wheel of the bike was placed into a Saris Trainer Climbing Riser Block at its lowest level to keep the bicycle stable.

Figure 5. Bicycle simulator.

The speed of the bicycle was communicated with a Garmin Bike Speed 2 Sensor. This speed sensor was attached to the hub of the bicycle's rear wheel and provided accurate speed readings of the wheel in real-time. The speed sensor's readings were sent to the VR computer through a Garmin Universal Serial Bus (USB) ANT Stick. ANT is an ultra-low power (ULP) wireless protocol that sends information wirelessly from one device to another device. The USB ANT stick was plugged into the computer, received the readings from the speed sensor, and passed the speed readings onto Unity. The speed sensor's readings were interpreted into Unity using the Advanced ANT+ plugin from the Unity Asset Store. In this way, a physical movement of the bicycle's rear wheel was translated into movement of the virtual bicycle in the Unity VR environment.

The direction of the bicycle was communicated with a single VR controller. The controller was attached to the bicycle's handlebar horizontally and facing forward. The virtual handlebar was programmed as an object in Unity and moved about the bike's head tube axis when the VR controller moved. The front wheel of the bicycle in VR was attached to the handlebar object and also moved when the virtual handlebar moved. In this way, a physical movement of the physical bicycle's handlebar was translated into movement of the virtual bicycle's front tire in the Unity VR environment, and thereby initiated a turning movement in VR.

The driving simulator consisted of a Logitech G920 Dual-Motor Feedback Driving Force Racing Wheel (Figure 6). The steering wheel's force feedback provided a highly realistic driving experience. The Logitech racing wheel also included acceleration and brake pedals. The Logitech steering wheel and pedals were connected to Unity through the Logitech Gaming Software Development Kit (SDK) available on the Unity Asset Store. Although the script required some tuning in terms of obtaining a realistic braking and acceleration feel, the SDK provided us with baseline functionality. The steering wheel and pedals were attached to a GTR Simulator GTA Model Racing Seat. The frame length of the racing seat was adjustable so that participants of all heights could comfortably use the driving simulator. The recline of the seat back was also adjustable so that each participant could select a comfortable driving position. We did not incorporate the clutch or shifter as we assumed an automatic transmission would be most familiar to participants. Because we did not have a shifter, the driving simulator was not able to go in reverse.

Figure 6. Driving simulator.

4.3. eHMI Design

The AVs had a variety of eHMIs based on those currently proposed by AV companies (Figure 1). For the non-textual LED (light-emitting diode) windshield strip eHMI design (Figure 7.A), the AV communicated that it was yielding with lights moving in/out laterally from the center of the strip to the outside and then blinking once when the outside lights were illuminated. When the AV was driving without intention to yield, the lights in the center of the LED strip were illuminated continuously. The eHMI design with arrows on the side mirrors (Figure 7.B) was a non-textual display with arrows that blinked when the AV was yielding the right-of-way. The arrows were not illuminated when the AV was driving without intention to yield. A text interface on the grille (Figure 7.C) communicated yielding behavior when arrows illuminated across the interface and the "Please Proceed to Cross" text blinked three times. The interface was not illuminated when the AV was driving without intention to yield. A text interface on the roof (Figure 7.D) communicated yielding behavior when the "Proceed to Cross" text remained solid for two seconds and then blinked twice. The interface was not illuminated when the AV was driving without intention to yield.

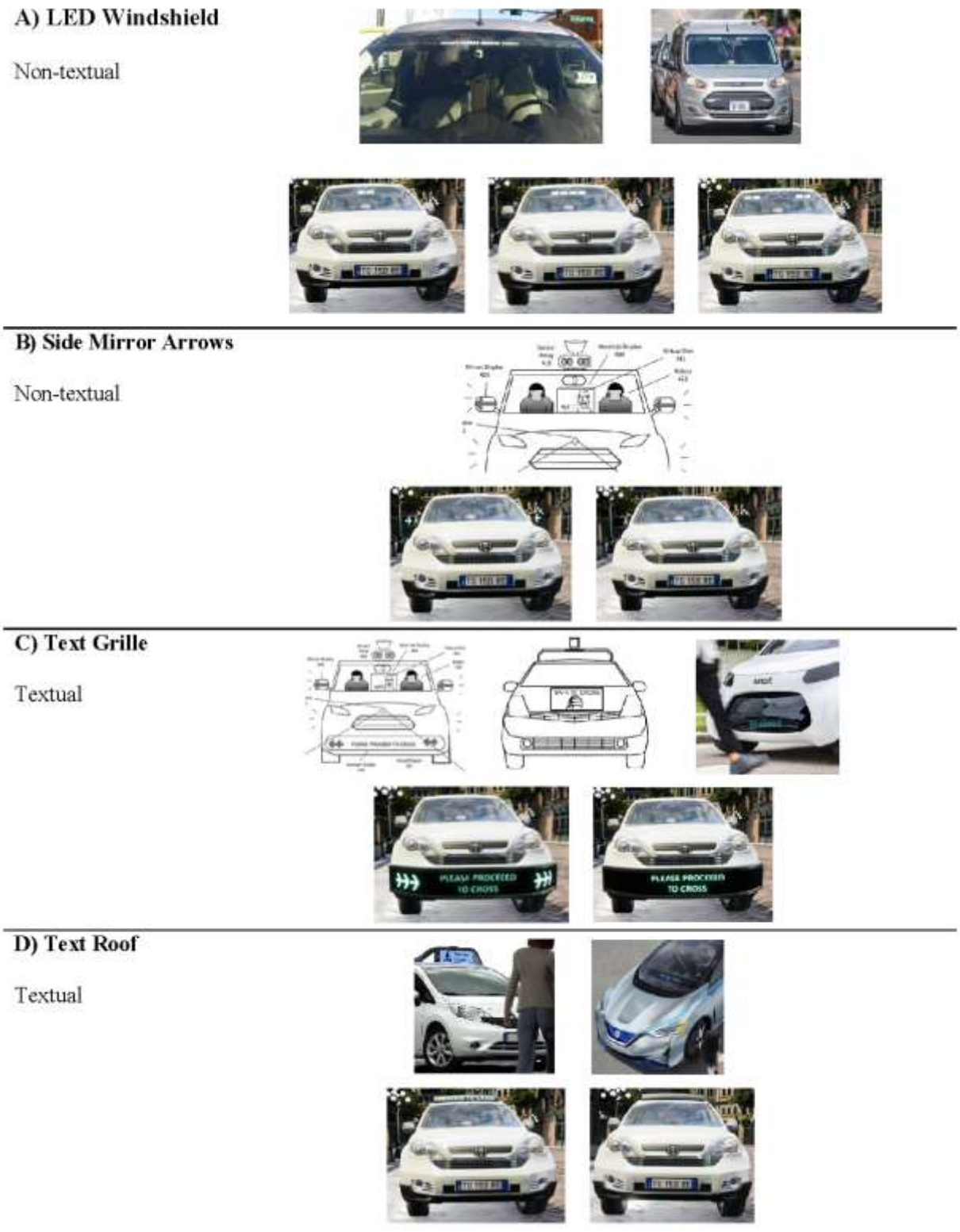

Figure 7. eHMI designs and operations.

After preliminary testing of participants in the driving simulator, we realized that driving participants were back far enough that often they could not clearly decipher the displays positioned on the front of the AV (Figure 8). We therefore also included a side-displayed eHMI for the driving and bicycle simulators. This side-displayed eHMI had a similar layout to the text grille eHMI (Figure 7.C), but was simply placed in a more visible position.

Figure 8. Side-displayed eHMI on driver side door.

All the textual messages were egocentric for the pedestrians (i.e. told the pedestrians what to do in the form of "Proceed to Cross") as opposed to allocentric. The lights on all the displays were the same white color and there were no noises incorporated into the designs. Each AV had one eHMI displayed for each testing trial.

4.4. Scenario Design

We sought to design VR scenarios where the participants would use the eHMI to interpret the AV's actions when interacting with the AV. Therefore, we designed VR scenarios where right-of-way was ambiguous. In other words, if it was clear that the AV had the right-of-way in the interaction, then the participants might not even pay attention to the AV as they would not expect the AV to stop for them. Our scenario needed to be relatively unclear so that the participants were forced to understand the AV when making their decisions.

For the pedestrian scenario, the environment was a midblock crossing in a downtown setting with no adverse weather conditions (Figure 9). The mid-block crossing allowed for a high level of AV-pedestrian interaction because it was not immediately clear who had the right-of-way or whether the AV would yield. Our AV was modeled on a compact sport utility vehicle with no driver or passengers present. This AV design was consistent throughout the different modal simulators. There were no other automobiles or humans present in the environment. The pedestrian could wander freely throughout the pedestrian scenario.

Figure 9. Pedestrian scenario (the gray box in the top left and the black box in the bottom center are not visible in the VR headset).

Each participant appeared in the VR scenario on the sidewalk approximately two feet from the curb. Before entering the scenario, a lab technician informed each participant that they would be interacting with an AV that may or may not yield to them and provided instructions regarding what to expect during the trial. Every trial was administered and supervised by the same lab technician. Once in the VR environment, we first provided a few minutes for each participant to acclimate to the VR environment and proceeded upon the participant's consent. The trial began with an illuminated button appearing vertically from the ground next to the participant at the curb. The participant was instructed to press the button with their hand, in which they held a VR controller. Upon pressing the button at the curb, the AV (which was originally out of sight) began moving toward the participant and another button appeared in the middle of the road. The participant was tasked with pressing the second button, but first had to negotiate with the arriving AV. Once the participant understood whether the AV was yielding to them, they were able to walk out and activate the second button. Upon activation of the second button, the trial was over and the participant was administered a survey. Such task-based experiments have been shown to be effective for exploring road user behavior (*29*, *35*, *96*).

For the biking scenario, the overall layout was similar to the pedestrian scenario. Bicycling participants started each VR trial at the beginning of an off-road bike trail. The bicycle in the VR scenario was a standard road bike. The height of the participant was adjustable in the VR environment so that each participant felt comfortable with their position on the bicycle. The participant was able to steer in a fully immersive fashion since the bicycle was not on a track in the VR scenario. The bicyclist participant would ride about one hundred yards down the trail and arrive at a midblock crossing of a roadway (Figure 10). The bike would automatically stop at the crossing, just in case the participant was not familiar with the use of the brakes. Once the bicycle was within twenty feet of the crossing, the AV was triggered and also approached the crossing on the roadway from the right of the bicyclist. The AV would always stop at the crossing, but as with

the pedestrian scenario, since the crossing was at a midblock location, it was ambiguous whether the AV would yield to the participant or whether it would continue to drive forward. To enforce this ambiguity, we programmed the AV to sometimes yield and sometimes not yield, as further detailed in Section 4.6. The participant was therefore again forced to interpret the AV's eHMI to determine the AV's actions. In the bicycle scenario, we were able to pause the scenario when the bike and AV were both at the crossing and ask the participant whether they understood the AV's intentions. This allowed us to measure understanding before the participant knew exactly how the AV would behave. If we had waited until after the scenario to ask whether the participant understood the AV's intentions, the participant would already have finished the trial and their answer would therefore be biased. The bicycle scenario finished after the bicyclist crossed the roadway and stopped in a parking lot at the end of the trail.

Figure 10. Midblock crossing in the bicycle scenario (the text in the top left and the black box in the bottom center are not visible in the VR headset).

For the driving scenario, the car that the participant used in VR was based on a standard sedan with a generic interior (Figure 11). The height of the participant was again adjustable in the VR scenario so that each participant was able to be comfortably positioned in the virtual vehicle. The steering wheel, tachometer, and speedometer were all linked with the actual behavior of the virtual car to provide a highly realistic driving experience. Again, the driver participant in the car was able to accelerate and brake but was not able to put the car into reverse (which would have been unnecessary in the scenario, anyway). The scenario was fully immersive and the driver could drive anywhere in the scenario that they wished.

Figure 11. Interior of the virtual car in the driving scenario (the text in the top left and the black box in the bottom center are not visible in the VR headset).

To incorporate right-of-way ambiguity into the driving scenario, we based the scenario on a four-way stop-controlled intersection (Figure 12). We programmed the AV to arrive and stop at the intersection at the same time as the participant. In this way, it was unclear whether the AV had seen the participant and was letting them go or whether the AV had not detected the participant and would continue through the intersection. Again, we programmed the AV to sometimes yield and sometimes not yield, as further detailed in Section 4.6. We were able to pause the scenario when both the participant and the AV arrived at the intersection and stopped. At that time, we asked the participant whether they understood the AV's intentions. We then un-paused the scenario and let the participant drive to a parking lot at the end of the street, at which time the trial was completed.

Figure 12. Four-way stop-controlled intersection in the driving scenario (the text in the top left and the black box in the bottom center are not visible in the VR headset).

4.5. Participant Sampling

Because we had human test participants, we received approval for our project from the University of New Mexico's Office of the Institutional Review Board (IRB). This office serves both main and branch campuses of the University of New Mexico and is registered under the IRB Registration Number of IRB00000431. Our specific project was approved through Board Reference Number 09019.

We sought to have a large and representative participant sample across genders, ages, races, ethnicities, and driving, walking, and biking experience. Unfortunately, because of the COVID pandemic, our participant testing was seriously delayed, forcing us to limit our participant sample size. We instead decided to administer a higher number of tests to each of our participants, thereby preserving the overall trial sample size.

We had originally proposed to test 50 participants twice in the pedestrian simulator and twice in the driving simulator for a total of 150-200 trials (assuming some trials might be disrupted by VR sickness). We instead decided that for each simulator, each participant should experience the AV yielding with each of the five eHMIs plus an AV with no eHMIs. We also randomly provided one trial for each simulator where the AV did not yield, making a total of seven trials for each participant and each simulator. Each participant experienced two of the three simulators. We selected only two of the three for the sake of time as participants experienced testing fatigue if they went through all three simulators at once. Furthermore, we would bring participants back to repeat all of the aforementioned tests a second time to examine any learning effects. Each participant would therefore experience seven eHMIs multiplied by two simulators multiplied by two testing times for a total of 28 trials each. To reach our originally proposed trial sample size, we sought to bring ten participants into the lab for a total of 280 trials. Because some of the early participants underwent all three simulators, our total trial size was 310 trials (as further detailed in the Results section). In this way, although our participant sample size was lower than originally anticipated, our trial sample size was significantly larger.

Participants largely came from the University of New Mexico community with many students being represented. We hope to increase our participant sample size now that COVID restrictions on testing have been mostly lifted.

4.6. Trial Ordering

In terms of ordering the simulators, while we were not able to complete a full within-subject design (since most subjects only experienced two of the three simulators) we counterbalanced the simulators to control for any learning effects. For example, we did not want every participant to experience the pedestrian simulator first as this would bias the results for that simulator.

When each participant was introduced to a new simulator, the participant was allowed to have as many practice rounds in the simulator as they needed until they felt comfortable. The AVs in these practice rounds did not have an eHMI so as to not bias the participant. The AV was programmed to randomly yield or not yield throughout these practice rounds so that the participant was not conditioned to expect any one specific behavior.

In terms of the eHMI ordering, primarily because of our small participant sample size, we had a within-subject design where every participant saw each of the five eHMIs (as well as no eHMI) for each simulator, resulting in a total of six eHMI configurations for each simulator. We used a

Latin square experimental design to counterbalance the trials. A Latin square has each eHMI in each row and each column, meaning that each participant sees each eHMI configuration, but never in the same order as any other participant (Table 1). This experiment design ensures that no one eHMI is biased toward the beginning or the end of the learning curve. Since there were six possible eHMI configurations, a complete Latin square experiment design would have had six participants. However, to reach our desired trial sample size, we had ten participants. So while we followed the Latin square experimental design, it was an incomplete Latin square design because we went through approximately one and a half of the Latin squares for the ten participants. As we continue to test participants in the future, we hope to finish our testing on a participant number divisible by six so as to have a complete Latin square design.

Table 1. Latin square experimental design where each letter refers to a different eHMI configuration.

		Trials					
		1	**2**	**3**	**4**	**5**	**6**
Participants	**1**	A	B	F	C	E	D
	2	B	C	A	D	F	E
	3	C	D	B	E	A	F
	4	D	E	C	F	B	A
	5	E	F	D	A	C	B
	6	F	A	E	B	D	C

For the six eHMI configurations noted above, the AV yielded each time. We therefore also added one trial where the AV did not yield in order to amplify the ambiguity of the interaction, thereby forcing participants to interpret the eHMIs. Because the primary focus of the research was on the differences between the eHMIs and not the differences between yielding and not yielding, we decided that this experimental design was appropriate. We would have liked to include more trials where the AV did not yield, but doing so greatly increased experiment fatigue of the participants. Since there was only one non-yielding trial for each participant in each simulator, we could not completely counterbalance the yielding behavior and therefore placed the non-yielding trial randomly within the larger Latin square experimental design. This non-yielding trial was seen as an additional seventh trial and necessarily repeated an eHMI configuration already in the Latin square.

4.7. Survey Instruments

Before the trials began, we had each participant complete a consent form. We then asked some preliminary questions regarding demographics and existing perceptions toward AVs. Specifically, we obtained each participant's age, gender, hours of driving, walking, and biking in a typical week, and whether they had used VR before. We then administered a survey measuring each participant's trust of AVs based on the survey developed by Jian et al. (*15*). This survey took the form of a seven-point Likert scale and was validated by the original researchers (Figure 13). The original questions broadly referred to "the system" since the survey was originally developed to measure trust in automated systems. We therefore changed the wording to ask specifically about AVs. This survey was administered before any of the VR trials began because we wanted to measure pre-existing biases toward or against AVs. The first five questions in the survey are generally negative (asking whether the participant is wary, suspicious, or believes the system is deceptive or

underhanded) while the last seven questions are mirrored and are largely positive (asking whether the participant is confident or trusting).

The survey questions detailed above, as well as the survey questions that were asked during and after the VR trials, were present on an Excel spreadsheet that was visible to the participant. In this way, each participant could both read the survey questions themselves, and the lab technician also read each question to each participant. The participant would verbally answer each question and the lab technician would enter the answer into the spreadsheet. After all the questions were answered for a trial, we hid that row in the spreadsheet so that the no participant could see their previous answers or anyone else's answers.

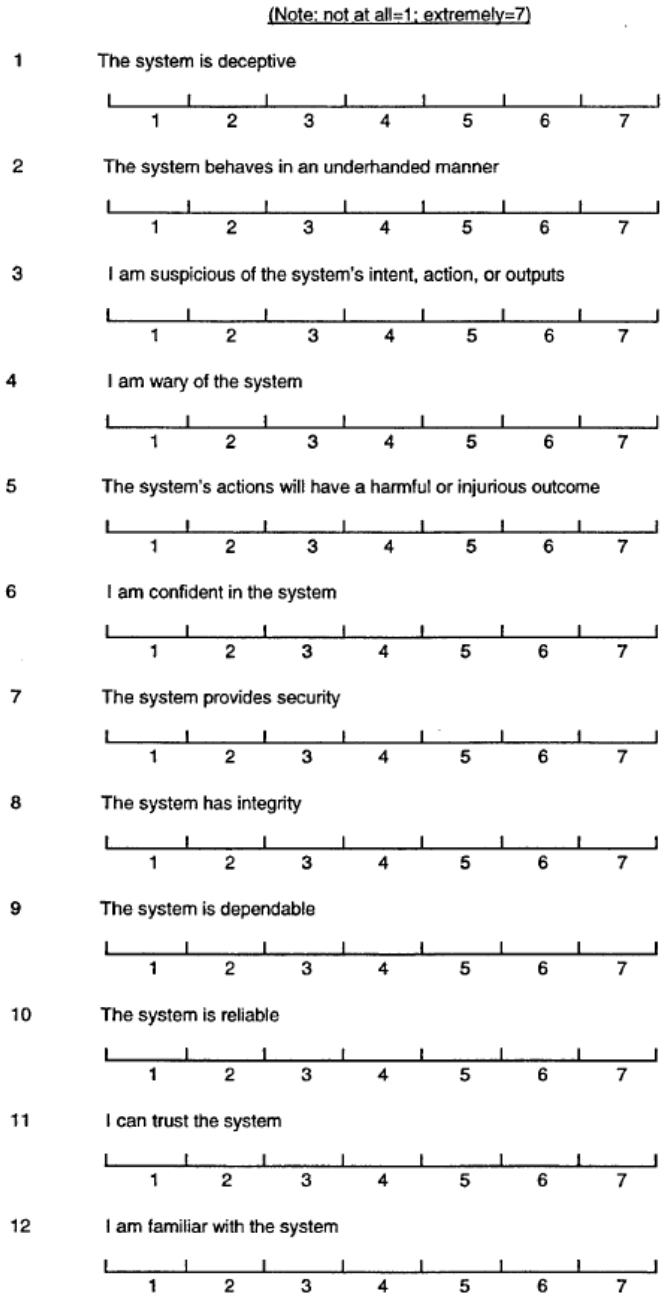

Figure 13. Jian et al. survey that measures participants' trust in automated systems (*15*).

The next questioning occurred during the VR trials. We paused the VR scenario when the participants arrived at the crossings or intersection with the AV and asked the participants whether the AV would or would not yield the right-of-way to them. While the scenario was paused, the participant's view also became blurry so they would not have extra time to examine the eHMI. They were therefore required to make their decision based on their interaction before the pause occurred. The participant would answer verbally whether they believed the AV was going to yield or not, oftentimes with the VR headset still on. This was the most effective way of measuring each participant's understanding of the AV. If we had waited until after each trial finished, the participant would know what the AV's intentions were and their answer would be biased. In that way, measuring understanding during the VR trial provided us with an unbiased assessment.

After each trial, we had each participant answer several surveys that were aimed at measuring trust, comfort, acceptance, and task workload. We first asked each participant the Jian et al. questions again to measure trust of that particular eHMI configuration (*15*) (Figure 13). Next, we measured comfort with the Bartneck et al. survey (*16*) (Figure 14). This survey was originally developed on a five-point Likert scale but we converted to a seven-point Likert scale for consistency with the other surveys. The term "comfort" was also referred to as "perceived safety" by the researchers. The last question is mirrored relative to the first two questions ("quiescent" is on the lower end of the scale whereas "relaxed" and "calm" are on the higher end of the scale). We framed each question as: "During your interaction with the AV, were you anxious (1) or relaxed (7)?"

GODSPEED V: PERCEIVED SAFETY

Please rate your emotional state on these scales:

以下のスケールに基づいてあなたの心の状態を評価してください。

Anxious 不安な	1	2	3	4	5	Relaxed 落ち着いた
Agitated 動揺している	1	2	3	4	5	Calm 冷静な
Quiescent 平穏な	1	2	3	4	5	Surprised 驚いた

Figure 14. Bartneck et al. survey that measures participants' comfort (or perceived safety) in robots (*16*).

We next measured each participant's acceptance of the AV using the Van Der Laan et al. survey (*17*) (Figure 15). This survey was developed to measure drivers' acceptance of new technological transport equipment. We again converted this survey to a seven-point Likert scale for consistency. Items three, six, and eight were mirrored relative to the other items with negative response being lower. We framed each question as: "Was that eHMI configuration useful (1) or useless (7)?"

My judgements of the (...) system are... (please tick a box on every line)		
1	useful	useless
2	pleasant	unpleasant
3	bad	good
4	nice	annoying
5	effective	superfluous
6	irritating	likeable
7	assisting	worthless
8	undesirable	desirable
9	raising alertness	sleep-inducing

Figure 15. Van Der Laan et al. survey that measures participants' acceptance of advanced transport telematics (*17*).

We finally measured the task load of each trial using Hart and Straveland's NASA TLX (*19*) (Figure 16). We used the simplified version of the questions seen in Figure 16. Although the figure shows the scale as ranging from "very low" to "very high", we had participants report their task load responses on a 21-point Likert scale since there were 21 increments on the scale. This was ideal as the participants were often still in the simulators as they answered the questions and the 21-point scale allowed them to easily answer from their position in the simulator.

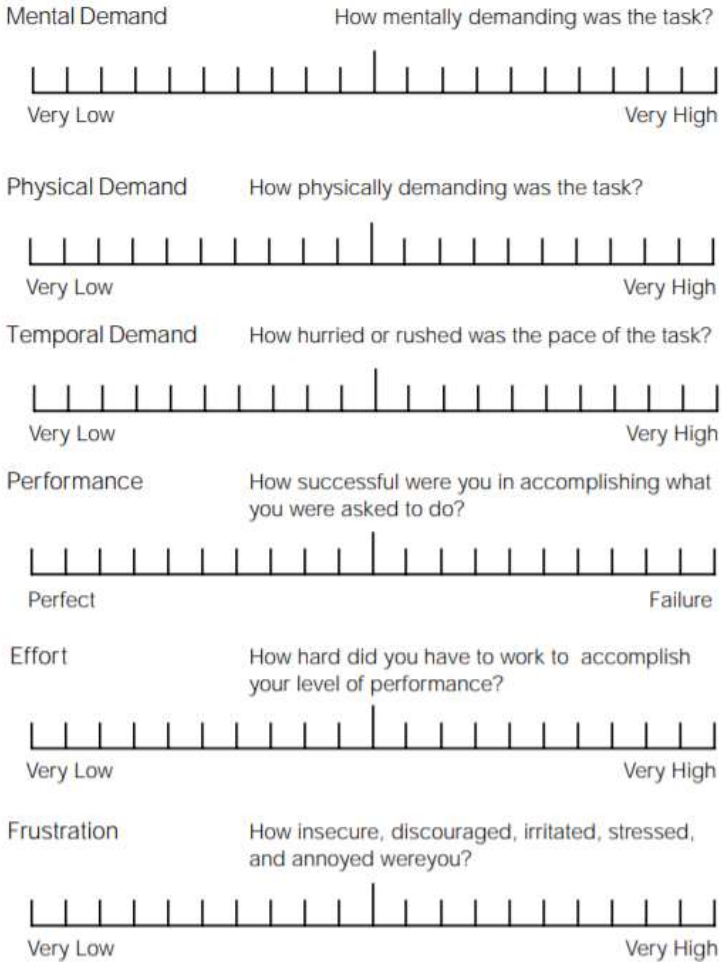

Figure 16. Hart and Straveland survey that measures task load (*19*).

We also allowed participants to provide open-ended comments at the end of the surveys, although we did not use those subjective responses for our data analysis. Upon answering all of the survey questions detailed above, the participant put the VR headset back on and continue with the next trial with another eHMI configuration.

4.8. Behavioral Measurements

For the behavioral measurement of the participants, we measured the timing of the interaction with the AV. We videotaped the sessions with the participants so that we could accurately measure these interactions after the sessions were finished as measuring during the sessions would have likely distracted both the participant and the lab technician. The video camera was setup behind

the participant so that the video captured both the participant and the computer screen that displayed what the participant was seeing inside the VR headset.

All the videos were analyzed by the same lab technician. To the hundredth of a second, we noted when the participant first made visual contact with the AV, when the participant stopped at the crossing, when the AV stopped at the crossing, when the participant answered whether the AV would be yielding, and when the participant continued their movement.

Given the varying layouts of the modal simulators, their reaction time definitions varied. For the pedestrian simulator, the participant was already stopped at the edge of the street and we did not program the pause functionality into the simulator, precluding their answering the understanding question. On the other hand, because the pedestrian was located in close proximity to the edge of the street, they had a much better view of the AV as it proceeded toward the crossing. Therefore, the pedestrian reaction time was defined as the time from the AV stopping to the time that the pedestrian walked into the street. The pedestrian reaction time was relatively short, likely because of the high visibility of the pedestrians' positioning. For the driving and biking simulators, the reaction time was defined as the time from when the AV stopped to the time when the participant answered whether the AV would be yielding or not yielding (while the scenario was paused). The driving and biking reaction times were generally longer than the pedestrian reaction times for two reasons. First, we had to ask the understanding question and the participant had to verbally answer, which likely took a few seconds longer than performing the simple action of stepping into the street. Second, visibility of the AV was lower in the driving and biking simulators compared to the pedestrian simulators. While the pedestrian participants were waiting at the edge of the road and were able to watch the AV as it approached the crossing, the driving and biking participants were occupied with the operation of a vehicle and likely were not able to concentrate on the AV until they had come to a complete stop.

For the above reasons, the reaction times of the participants varied between modes. We therefore did not directly compare reaction times between the different simulators, as further detailed in Section 5.8.

4.9. Longitudinal Timing

We brought participants back between one and three days after their first testing round so that they could complete their second testing round. All VR scenarios, survey instruments, and measurements were exactly the same between the first and second testing rounds. The only exception was that the initial questionnaire was not administered before the second testing round. Anecdotally, the second testing round generally went faster than the first as the participants had acclimated to the testing process, although we did not record information on the overall timing of each round.

4.10. Statistical Analysis

We sought to answer a number of research questions, first focusing on perceptions and then on behaviors. Was there a perception acclimation effect, or did perceptions change relative to time? Did the AV's yielding behavior impact perceptions? Did the presence of an eHMI impact perceptions? Were there differences in perceptions between modes? How did different eHMI designs impact perceptions? Our approach to these perceptions questions was to analyze each one

individually then form statistical models that considered all factors. We explored the same questions for behavioral outcomes, although the smaller sample sizes precluded statistical models.

Since each of the survey instruments (comfort, trust, acceptance, and task load) had up to twelve questions, we first summarized each survey into a single score. We flipped the mirrored questions so that all the questions were on the same scale (with positive responses such as more comfort or more trust being represented by a higher score) and took the average of each response. In this way, we derived a single task load score, a single comfort score, a single trust score, and a single acceptance score.

For the initial explorations before the full statistical models, we took the average scores for each factor that we were examining. For example, when exploring whether the presence of an eHMI impacted comfort, we derived the average comfort score for the 43 trials where there was no eHMI present (and the AV yielded). We compared that number to the average comfort score for the 200 trials where there was an eHMI present (and the AV yielded). This approach is justified because while there may have been some internal bias for individual participants (some participants tended to give higher scores while others gave lower scores), that bias was controlled for by the within-subject trial ordering. In other words, all participants saw all the eHMIs configurations, and those were ordered so no one eHMI configuration would have scored higher or lower than any other.

To test the longitudinal impacts, we divided the trials into four simulator periods. In other words, each participant experienced two simulators during their first session, and each participant again experienced two more simulators when they came back one to three days later for their second session. We therefore divided the results into those four simulator periods to better understand how scores changed over time. This is again warranted because each participant saw each eHMI configuration and a non-yielding trial during each simulator period, so individual biases should not have impacted results.

For the statistical models, we tested the statistical significance of the results with ordinal logistic regressions. This test is appropriate when there is an ordered, non-continuous dependent variable (as we had with our Likert scale output) and there are two or more independent variables. We used the polr command in R's MASS package. We ran a separate model for each outcome variable (i.e. task load, comfort, trust, and acceptance). Survey responses were used for the dependent variables. Independent variables consisted of AV characteristics (presence of an eHMI and yielding behavior), participant characteristics (gender, age, whether they had used VR before, and their reported trust of AVs before the trials began), and trial characteristics (which session the trial was in and the mode of the simulator). We attempted to convert all these independent variables into binary categorical variables so that we could compare results between variables. For instance, our time variable took the form of sessions (either the trial was in the first session or the second session). Participants' reported trust of AVs before the trials was split into the highest 50% of scores and the lowest 50% of scores. Yielding behavior, eHMI presence, gender, and prior VR use were already in binary form. For the understanding outcome variable, since the dependent variable was binary (participants either understood or did not understand the eHMI), we employed a logistic regression model with the same approach detailed above.

5. ANALYSIS AND FINDINGS

5.1. Descriptive Statistics

We were able to test ten participants for a total of 310 trials. This resulted in a total of 9,610 survey questions asked to the participants. As described in Section 4.5, we anticipated 280 trials given our chosen participant sample size. However, for our first participant, we tested all three simulators during each session and added three additional non-yielding trials to each simulator. However, we realized that the number of trials caused testing fatigue and therefore reduced the trial counts to those detailed in Section 4.5 after that initial participant, resulting in our 310 total trials.

Seven of the participants identified as male and three identified as female. The average age of the participants was 27.4 years with a minimum of 20 years and a maximum of 61 years. The average hours of driving in a typical week for the participants was 6.1 hours with a minimum of zero hours and a maximum of 40 hours. The average hours of walking in a typical week for the participants was 5.8 hours with a minimum of one hours and a maximum of ten hours. The average hours of biking in a typical week for the participants was 2.1 hours with a minimum of zero hours and a maximum of five hours. Three participants had used VR before while the other seven had not.

The average score for trust of AVs before the trials began was 4.2 points. Again, this was on a seven-point Likert scale with higher scores correlating with higher levels of trust. The minimum score was 2.3 points and the maximum score was 6.4 points with a standard deviation of 1.2 points.

5.2. Longitudinal Changes in Perceptions

We will break apart each factor (eHMI presence, eHMI design, yielding behavior) longitudinally in the sections below, but we first want to examine the overall impact that acclimation had on the perception outcomes. Understanding of the eHMIs improved over time, but not to a statistically significant degree. In Table 2, we explored all trials where the AV yielded to the participant and where there was an eHMI present. We used a logistic regression because of the dichotomous dependent variable (the participant either understood the AV or they did not) and we used the trial number as the independent variable. The coefficient was positive, signifying that understanding increased as trial number increased (i.e. as participants' exposure to the AVs increased), but this only reached statistical significance at 90% confidence. As will be seen below, this was largely a result of levels of understanding already being relatively high in the early trials.

Table 2. Logistic regression describing relationship between trial number and understanding (all trials had an eHMI and the AV yielded in all trials) (results statistically significant at 95% confidence in bold).

	n	Coefficient	p-value
Understanding	131	0.067	0.076

The other perception outcomes were continuous variables and we therefore derived linear regressions to explore the relationship between trials numbers and perception outcomes. Again, all of the trials represented in Table 3 had eHMIs present and the AV yielded in all the trials. The only perception that reached statistical significance was comfort. The positive coefficient suggests that participants' comfort increased as the trials progressed. The relationships between the other perception variables and trial number were very weak, as we shall see in more detail below.

Table 3. Linear regressions describing relationship between trial number and perception scores (all trials had an eHMI and the AV yielded in all trials) (results statistically significant at 95% confidence in bold).

	n	Coefficient	p-value
Task Load	200	-0.002	0.906
Comfort	200	0.025	**0.002**
Trust	200	-0.001	0.855
Acceptance	200	0.001	0.859

5.3. Yielding Behavior and Perceptions

We next explored the impact of the AV's yielding behavior (either yielded to the participant or did not yield) in all trials where there was an eHMI present. As expected, perceptions were improved when the AV yielded to the participant (Table 4). For all modes, understanding improved from 76% when the AV did not yield to 92% when the AV did yield. Task load decreased when the AV yielded, meaning that the participants reported less mental and physical load.

Table 4. Average perceptions by yielding behavior (all trials had an eHMI present).

		n	Understand	Task Load	Comfort	Trust	Acceptance
Overall	No Yield	57	76%	8.50	4.18	3.35	3.28
	Yield	200	92%	6.50	5.52	4.85	4.73
	Impact of Yielding		15%	-2.00	1.34	1.50	1.45
Bike	No Yield	15	80%	10.82	3.30	3.04	2.69
	Yield	39	85%	7.08	5.45	4.49	4.53
	Impact of Yielding		5%	-3.74	2.15	1.45	1.84
Drive	No Yield	27	74%	8.34	4.11	2.81	2.71
	Yield	93	95%	6.30	5.64	4.99	4.95
	Impact of Yielding		21%	-2.04	1.53	2.18	2.24
Pedestrian	No Yield	15	na	6.44	5.20	4.61	4.88
	Yield	68	na	6.46	5.38	4.87	4.55
	Impact of Yielding		na	0.02	0.18	0.26	-0.33

Because the comfort, trust, and acceptance surveys were on the same seven-point Likert scale, their outcomes can be directly compared. Overall and largely for each mode, comfort received the highest scores, trust fell in the middle, and acceptance was lowest. This order is seen often throughout the results in other sections as well.

When yielding, the driving simulator had the highest comfort, trust, and acceptance scores as well as the lowest task load scores, suggesting that participants had improved perceptions of AVs when they themselves were in a vehicle, possibly because of the protective benefits of the vehicle. The pedestrian simulator appears to be the second preferred simulator (except for comfort), while the bicycle simulator has the lowest perception scores. Future research might work to explore whether this is an accurate reflection of the modal preferences or how much the makeup of the simulators themselves impacts these differences.

The largest changes seen with yielding behavior were for trust and acceptance, which increased by 1.50 points and 1.45 points (respectively) when the AV yielded (Table 2). This may be in part because trust and acceptance were significantly lower than comfort for the non-yielding trials. Comfort was higher than trust or acceptance when the AV did not yield, but comfort scores increased the least in the yielding trials.

Understanding was lower for drivers when the AV did not yield (relative to the bike) and higher for drivers when the AV did yield (relative to the bike). The pedestrian scenario was not programmed with the pause functionality and we were therefore not able to measure understanding directly. Task load, comfort, trust, and acceptance all improved for biking and driving when the AV yielded to the participants.

Some of the pedestrian eHMIs had messages when the AV was not yielding (unlike in the biking and driving simulators), which likely resulted in much more positive perceptions for the non-yielding pedestrian trials relative to the non-yielding biking and driving trials. This suggests that eHMIs that communicate when the AV is not yielding may warrant further investigation.

How did perceptions change over time relative to yielding behavior? Results from Figure 17 represent the average of all modes and only include trials where there was an eHMI present. The x-axis in Figure 17 represents time, with 1 representing the first simulator in the first session, 2 representing the second simulator in the first session, 3 representing the first simulator in the second session, and 4 representing the second simulator in the second session. We use the same x-axis configuration for other longitudinal figures (including Figure 18 and Figure 19).

Results suggest that decreases in task load and increases in comfort, trust, and acceptance were relatively evenly distributed between yielding and non-yielding trials. This suggests that acclimation was occurring both when the AV was yielding and when it was not yielding. Understanding is the outlier perception. Understanding for yielding trials increased from about 80% during the first simulator of the first session to 95% during the last simulator in the last session (however, as we saw in Section 5.2, this change was only statistically significant at 90% confidence). On the other hand, understanding for non-yielding trials barely increased from the beginning of the trials to the end. There was a notable increase in understanding for non-yielding trials at the end of the first session, but understanding generally hovered around 65% for the rest of the trials. Again, this understanding metric was just for the bike and driving simulators (as the pedestrian simulator did not have pause functionality), suggesting that an eHMI message when the AV is not yielding may be beneficial.

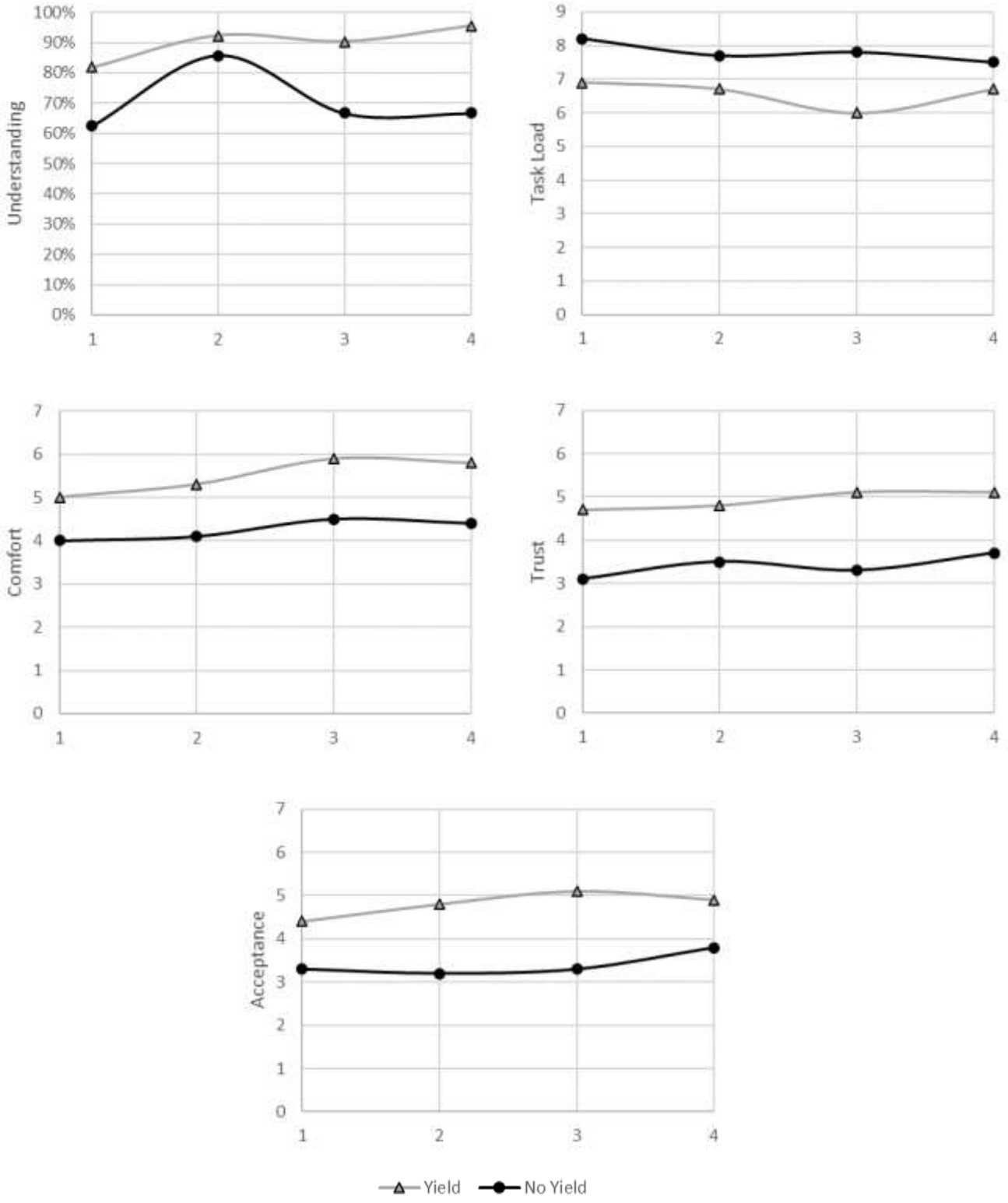

Figure 17. Average perception outcomes for yield versus no yield over time (only trials with an eHMI present) (the x-axis is time with 1=first simulator of first session; 2=second simulator of first session; 3=first simulator of second session; 4=second simulator of second session).

5.4. eHMI Presence and Perceptions

The presence of an eHMI greatly improved perceptions, both overall and for all modes (Table 5). The only exception was an increase in task load for biking which actually increased with the presence of an eHMI. This exception was likely a result of task load being highest for the biking simulator, which required relatively vigorous pedaling.

Table 5. Average perceptions by eHMI presence (the AVs yielded in all trials).

		n	Understand	Task Load	Comfort	Trust	Acceptance
Overall	Without eHMI	43	35%	7.33	4.08	2.75	2.29
	With eHMI	200	92%	6.50	5.52	4.85	4.73
	Impact of eHMI		57%	-0.83	1.44	2.10	2.44
Bike	Without eHMI	8	25%	7.02	3.81	3.04	2.81
	With eHMI	39	85%	7.08	5.45	4.49	4.53
	Impact of eHMI		60%	0.06	1.64	1.45	1.72
Drive	Without eHMI	18	39%	7.30	4.06	2.54	2.03
	With eHMI	93	95%	6.30	5.64	4.99	4.95
	Impact of eHMI		56%	-1.00	1.58	2.45	2.92
Pedestrian	Without eHMI	17	na	7.51	4.24	2.83	2.31
	With eHMI	68	na	6.46	5.38	4.87	4.55
	Impact of eHMI		na	-1.05	1.14	2.04	2.24

Overall, the increases in trust and acceptance were larger (2.10 points and 2.44 points, respectively) than the increase in comfort (1.44 points) (Table 5). This is similar to the yielding behavior findings in Section 5.3 (Table 4). Again, this is likely a result of comfort scores in trials without an eHMI being relatively high (4.08 points for comfort versus 2.75 points for trust and 2.29 points for comfort). While acceptance had the largest response to the eHMIs, trust had the second largest response, and comfort had the smallest response, the original order still held (comfort had the highest score and acceptance the lowest score both without and with an eHMI).

The eHMI introduction had a larger impact on trust, acceptance, and comfort compared to the change in yielding behavior. The introduction of the eHMI resulted in comfort, trust, and acceptance increases of around 1.5-2.5 points versus increases of around 1.0-1.5 points for the yielding behavior. However, the yielding behavior had a larger impact on the task load (a decrease of 0.83 points for the eHMI versus a decrease of 2.00 points for the yielding behavior). This justifies the importance of the inclusion of an eHMI. We explore the longitudinal impacts of eHMIs in Section 5.6.

While results for different modes were largely consistent in direction, they varied in magnitude. Understanding increases were similar for biking and driving. While task load decreases were similar for driving and pedestrians, biking task load actually increased with the introduction of an eHMI. The biking simulator also stands out because of relatively small increases in trust and acceptance with the introduction of an eHMI. Interestingly, this is partly because levels of trust and acceptance were relatively high for biking when there was no eHMI present.

While understanding was low at around 65% when the AV was not yielding, understanding was much lower at 35% when there was no eHMI present. This is likely because as participants got used to seeing a message on the eHMIs, they automatically assumed that the AV would not be yielding to them if there was no eHMI (and therefore no message). This speaks to the importance of making the communication method uniform and widespread.

5.5. Modes and Perceptions

Driving had the highest scores for understanding, comfort, trust, and acceptance as well as the lowest task load score (Table 6). Biking had the lowest scores for understanding, trust, and acceptance as well as the highest task load score. Future work might explore to what extent this relationship is a result of the design of these particular simulators versus actual intrinsic characteristics of these modes. The results seen here make logical sense since biking was likely the most unfamiliar mode to the participants, bicyclists are vulnerable without the protection of a vehicle, and the bicyclists needed to work harder than the pedestrians or drivers. But the biking simulator was also the only simulator built from scratch and the fidelity of the operations may have influenced the results.

Table 6. Average perceptions by mode (the AVs yielded in all trials and all had eHMIs).

	n	Understand	Task Load	Comfort	Trust	Acceptance
Bike	39	85%	7.08	5.45	4.49	4.53
Drive	93	95%	6.30	5.64	4.99	4.95
Pedestrian	68	na	6.46	5.38	4.87	4.55

As seen before, comfort received the highest scores, trust typically fell in the middle, and acceptance was usually lowest. This order largely held throughout the research and although eHMIs may have improved acceptance and trust more than comfort, the order of the scores infrequently changed.

We again examined how the scores changed over time, this time relative to the different modal simulators. We advise interpreting these results in Figure 18 with caution because of low participant sample sizes for each of the data points. For example, the first bicycle data points are the average of all scores for all participants that used the bicycle simulator first in their first session. This was only three participants. If two or three of those participants gave unusually high or low scores, the bicycle trend would be biased. So while the trial sample size is still large, the numbers may be biased because of the small participant sample size. This was not an issue for the other longitudinal analyses because all participants saw all the eHMIs during each simulator, all participants saw a non-yielding AV during each simulator, and all participants saw a non-eHMI AV during each simulator. But not all the participants used the bike simulator for their first simulator.

Over time, biking stands out as an outlier (Figure 18). Biking has the lowest task load in the first longitudinal point and the highest comfort, trust, and acceptance scores. However, biking consistently got worse as participants went through the tests and ended up having the highest task load and lowest comfort, trust, and acceptance scores by the end of the testing. Driving scores improved for comfort and task load but stayed the same for trust and acceptance. However, driving scores for trust and acceptance were already at their highest values at the beginning (scores for trust and acceptance rarely went over 5.0 points). Pedestrian trust and acceptance scores were a bit

worse than driving scores in the beginning but improved throughout and largely matched driving scores at the end, while biking scores were a little lower. Overall, driving and pedestrian scores seem to have improved throughout testing (or at least stayed high) while biking scores got worse.

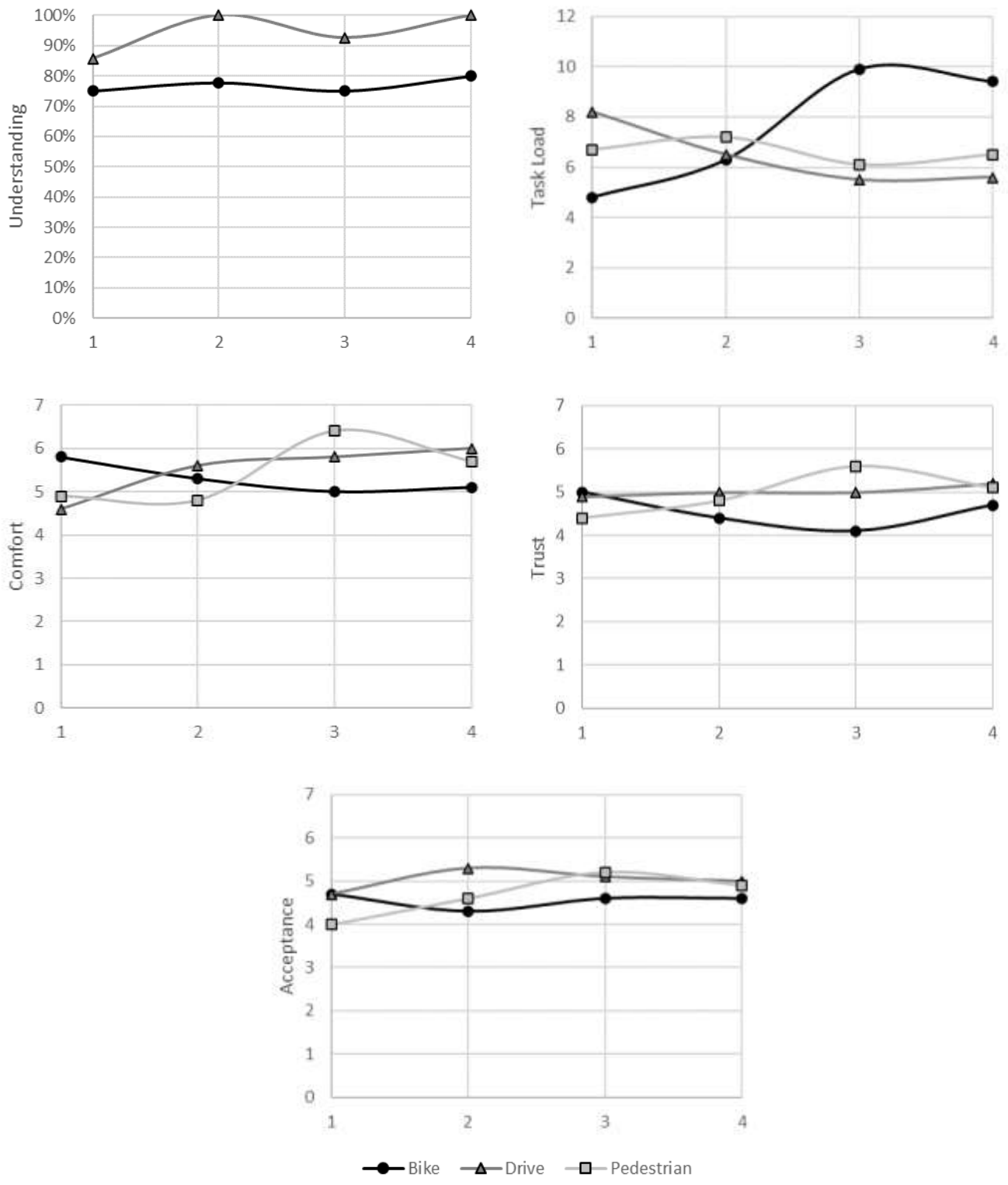

Figure 18. Average perception outcomes for different modes over time (the AVs yielded in all trials and all had eHMIs) (the x-axis is time with 1=first simulator of first session; 2=second simulator of first session; 3=first simulator of second session; 4=second simulator of second session).

5.6. eHMI Design and Perceptions

Which eHMI design performed best in terms of perceptions? The text-based eHMI on the grille (referred to as "Text on Grille") and the text-based eHMI on the driver-side front door (referred to as "Text on Driver Door") received the best perception scores, scoring higher than every other eHMI for understanding, comfort, trust, and acceptance and lower than every other eHMI for task load (Table 7). The text-based eHMI on the roof (referred to as "Text on Roof") generally fell in the middle, while the eHMI consisting of arrows on the side mirrors (referred to as "Side Mirror Arrows") and the eHMI with an LED strip on the windshield (referred to as "LED Windshield") were the worst, with LED Windshield being significantly worse than the others for most categories. The text-based eHMIs performed better than the non-text eHMIs.

Table 7. Average perceptions of eHMI designs (the AVs yielded in all trials; results are average of all modes).

	n	Understand	Task Load	Comfort	Trust	Acceptance
Text on Driver Door	27	93%	6.31	6.02	5.01	6.04
Text on Grille	27	100%	6.14	5.86	5.33	5.37
Text on Roof	25	92%	6.54	5.48	4.92	4.77
Side Mirror Arrows	26	92%	6.62	5.42	4.82	4.66
LED Windshield	27	81%	6.84	5.00	4.23	3.94

Also, it would be reasonable to hypothesize that task load would be higher for text-based eHMIs, but the data suggests that it was actually lower. Participants apparently struggled to interpret the non-text eHMI messages. More testing over longitudinal time periods might close this gap.

Again, comfort generally received the highest scores, trust fell in the middle, and acceptance received the lowest scores. How did these results change over time?

Understanding of the text-based eHMIs was high in the beginning and remained high throughout the testing, generally remaining in the 90%-100% range (Figure 19). Understanding of LED Windshield was particularly low in the beginning at around 60%, but increased to about 80% by the second simulator of the first session. Understanding of the Side Mirror Arrows was higher than the LED Windshield by lower than the text-based eHMIs. Understanding of the AVs with no eHMI present actually decreased by the second simulator of the first session as people started to expect an eHMI. Overall, text-based eHMIs performed better (average of 97% understanding) than non-text eHMIs (average of 86% understanding).

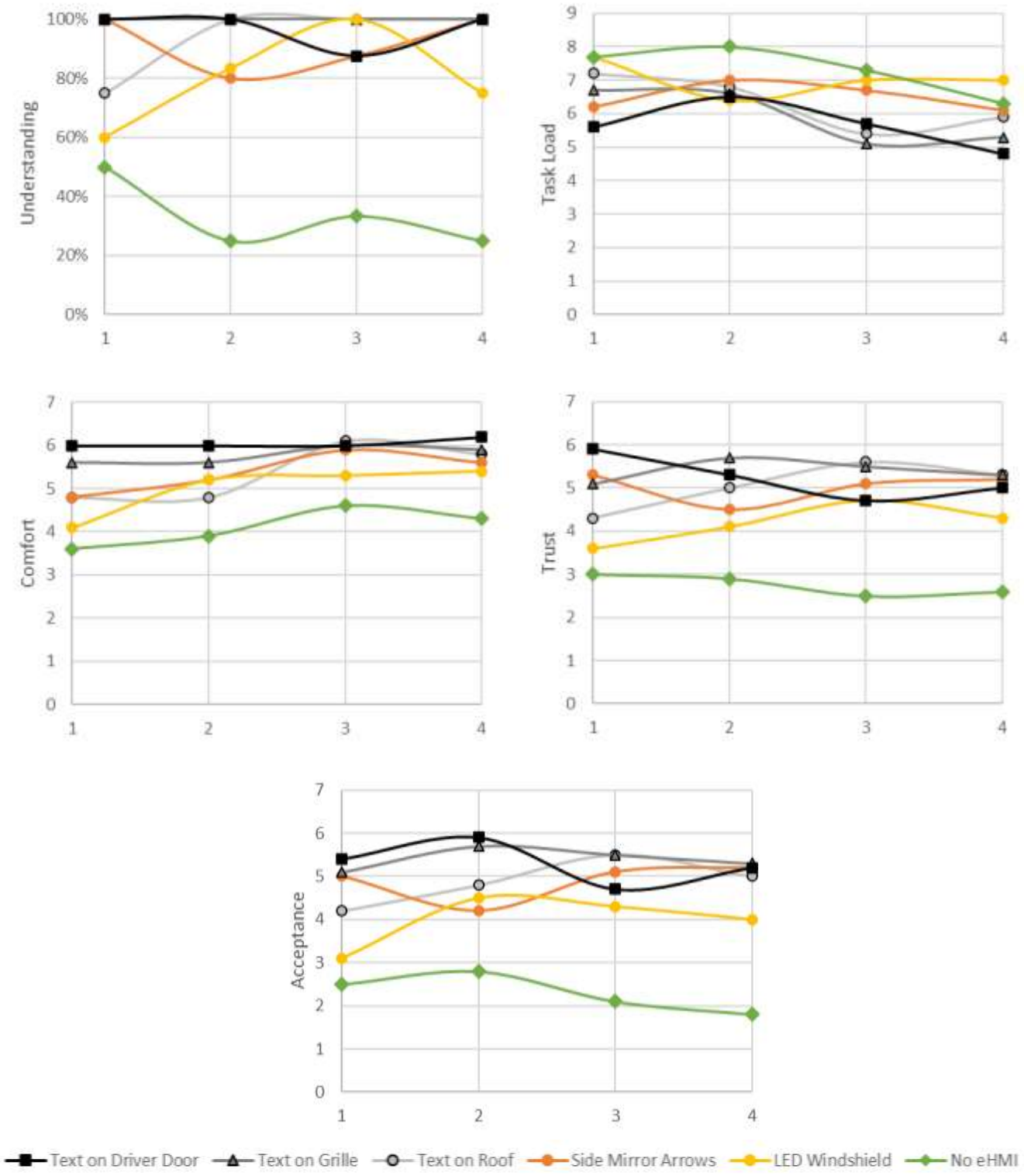

Figure 19. Average perception outcomes for different eHMI designs over time (the AVs yielded in all trials; results are average of all modes) (the x-axis is time with 1=first simulator of first session; 2=second simulator of first session; 3=first simulator of second session; 4=second simulator of second session).

A very interesting pattern emerges when examining task load in Figure 19. During the first simulator of the first session, the eHMIs are spread across a relatively wide range with task load scores of between 5.5 points and 7.5 points, with each eHMI having their own unique value and little clustering occurring. During the second simulator of the first session, task load scores for each eHMI cluster closely between 6.5 points and 7.0 points (except for AVs with no eHMI which remain high, as expected). During the first simulator of the second session, a clear distinction forms where the text-based eHMIs receive lower task load scores (clustered between 5.0 points and 5.5 points) and non-text eHMIs receive higher scores (clustered between 6.5 points and 7.0 points). This text distinction remains until the end of testing, although it is weakened in the final simulator. In summary, task load scores decreased for all eHMIs from beginning to end, but the decreases were strongest for the text-based eHMIs and AVs with no eHMIs. The non-text eHMIs stayed relatively stable from beginning to end, with the LED Windshield actually having an even higher task load score than AVs with no eHMI at the end.

The same pattern of clustering and text versus non-text differentiation occurs for comfort, except that complete clustering does not occur until the first simulator of the second session (one simulator later than with task load). Again, the text-based eHMIs begin with the highest comfort ratings (4.5 points to 6.0 points), the non-text eHMIs begin with lower scores (4.0 points to 4.5 points), and the AVs with no eHMI begin with the lowest scores (about 3.5 points). By the third simulator, all eHMI designs have concentrated at 6.0 points, except for LED Windshield which is just above 5.0 points and AVs with no eHMI at 4.5 points. After the clustering, there is again a slight differentiation during the fourth simulator where text-based eHMIs are higher and non-text eHMIs are lower. Overall, comfort scores largely improved over time for the non-text eHMIs and for AVs with no eHMI, while the text-based eHMIs were already high-scoring at the beginning. All comfort scores were higher at the end than at the beginning. Overall, comfort scores increased 0.7 points from beginning to end.

The same basic pattern again emerges for trust. The scores are highly variable in the beginning, ranging from 3.0 points to 6.0 points. Clustering again begins by the second simulator but is not fully complete until the fourth simulator (this full clustering is again pushed back one simulator later than the previous metric, or one simulator later than for comfort). In the final simulator, the eHMIs are strongly clustered at 5.0 points for trust (as opposed to 6.0 points for comfort), except for LED Windshield at just above 4.0 points and AVs with no eHMI at 2.5 points. Interestingly, while comfort increased throughout, trust scores are lower at the end than at the beginning for Side Mirror Arrows, Text on Driver Door, and no eHMI and LED Windshield experienced a drop in trust during the last simulator. These findings suggest that participants' comfort with AVs increased in general with acclimation relatively independent of the eHMI, but the participants were more dependent on appropriate eHMIs to improve their trust. Overall, trust scores increased only 0.1 points from beginning to end, suggesting that trust was more difficult to increase than comfort.

The same pattern is again apparent with acceptance. The range of scores is similarly wide as the trust scores, but slightly lower (2.5 points to 5.5 points). As with trust scores, clustering of acceptance scores begins by the second simulator but is not complete until the fourth simulator. The acceptance scores cluster strongly at 5.0 points by the final simulator, just as the trust scores did. Again, LED Windshield is lower than the other eHMIs at 4.0 points and AVs with no eHMI were below 2.0 points. Overall, acceptance scores increased 0.2 points from beginning to end, suggesting that both trust and acceptance were more difficult to increase than comfort.

5.7. Statistical Models for Perceptions

The above results provide in-depth examinations of each variable individually. To complete our analysis, we wanted to enter all variables into a statistical model so we could better understand the relationship between the variables. The results in Table 8 for task load, comfort, trust, and acceptance are from ordinal logistic regression models, which are optimal when there is an ordered, non-continuous dependent variable (as we had with our Likert scale outputs) and there are two or more independent variables. The results in Table 8 for understanding are from a logistic regression as the dependent variable was dichotomous (the participant either understood or did not understand).

The session variable refers to which session (first or second) the results were from, and is therefore a binary proxy for time. eHMI presence is also binary as there was either an eHMI present (signified as 1) or there was no eHMI present (0). The yield variable is also binary as the AV either yielded (1) or did not yield (0). The pre-trust variable was derived from the trust survey administered to each participant before the trials began and was a measure of participants' preconceived biases toward or against AVs. For the sake of comparison, we converted the pre-trust score into a binary variable with the lowest half of scores being represented by 0 and the highest half of scores being represented by 1. The gender variable was also binary with females being represented by 0 and males being represented by 1. The "Use VR Before" variable was also binary and represented whether participants had used VR before (1) or had not used VR before (0). The mode variable had three categories with biking being represented by 0, driving represented by 1, and pedestrian represented by 2. Age was a continuous variable. Therefore the session, eHMI presence, yield, pre-trust, gender, and Use VR Before variables were binary while the mode had three categories and age was a continuous variable.

Table 8. Ordinal regression results for all trials (results statistically significant at 95% confidence in bold).

	Understanding		**Task Load**		**Comfort**		**Trust**		**Acceptance**	
	Coef.	p-value	Coef.	p-value	Coef.	p-value	Coef.	p-value	Coef.	p-value
Session	0.2213	0.5938	-0.8614	**<0.0001**	1.0842	**<0.0001**	0.1913	0.3398	0.1528	0.4484
eHMI Presence	2.5021	**<0.0001**	-0.4881	0.0939	1.7914	**<0.0001**	2.6551	**<0.0001**	2.9500	**<0.0001**
Yield	1.0190	**0.0390**	-0.7779	**0.0041**	1.2278	**<0.0001**	1.5749	**<0.0001**	1.5195	**<0.0001**
Pre-Trust	-0.4675	0.4218	-3.5488	**<0.0001**	0.4580	0.0767	1.1237	**<0.0001**	1.0460	**<0.0001**
Gender	0.2869	0.6355	0.1203	0.6634	-0.3389	0.2589	-0.9618	**0.0017**	-1.3681	**<0.0001**
Use VR Before	-0.4194	0.4166	1.8453	**<0.0001**	0.1550	0.5346	-0.5353	**0.0311**	-0.5647	**0.0222**
Mode	0.5543	0.2478	-0.5914	**<0.0001**	0.1397	0.3804	0.4050	**0.0103**	0.2726	0.0834
Age	0.0241	0.2933	0.0893	**<0.0001**	-0.0305	**0.0077**	0.0039	0.7230	0.0056	0.6133

One of the most important conclusions from Table 8 is the importance of the eHMIs. The presence of an eHMI had statistically significant relationships with understanding, comfort, trust, and acceptance. The positive direction of these relationships signifies that the presence of an eHMI improved these outcomes. Because the first six variables were binary and their coefficients can be compared, results suggest that the presence of an eHMI had the strongest relationship with understanding, comfort, trust, and acceptance.

The second strongest variable was whether or not the AV yielded to the participant. Yielding had a statistically significant and positive impact on understanding, comfort, trust, and acceptance and a statistically significant and negative impact on task load (all in the expected direction).

What impact did acclimation have? Interestingly, the session variable only reached statistical significance with the task load and comfort outcomes, where they were in the expected direction (task load decreased with time and comfort increased with time). This largely aligns with the findings from the regressions in Section 5.2 that found only comfort had a statistically significant relationship with trial number.

What was the relationship with participant characteristics such as age, gender, and preconceived ideas of trust? Importantly, the participant characteristics had weaker relationships with outcomes than the AV characteristics had. However, statistically significant results were largely in the expected direction for age and pre-trust. Task load increased with increasing age and comfort decreased with increasing age while trust and acceptance were higher with more trusting participants and task load was lower. Gender was opposite of the expected direction, with females reporting more trust and acceptance. We advise interpreting the "Use VR Before" variable with caution because there were only three participants that had used VR before, so those results were likely biased by that small participant sample size.

Overall, whether the AV had an eHMI present and whether the AV yielded were the strongest factors correlated with outcomes. The acclimation factor was relatively weak and most impacted task load and comfort while participant characteristics appear to be more important for trust and acceptance.

5.8. Behavioral Outcomes

We provide preliminary findings for behavioral outcomes below, but we advise using caution in the interpretation of the behavioral results as the sample sizes were low. We ended up with 128 trials with a recorded reaction time, compared to our 310 trails overall. This was a result of technical difficulties in our recording device and poor positioning of the device in some of our early sessions. Furthermore, the reaction time for each modal simulator varied because of the different layouts of the scenarios (as detailed in Section 4.8). Because each modal simulator varied and the biking simulator had an especially small sample size (only eight trials were successfully recorded because of early technical issues with the recording device and poor positioning), we removed the biking simulator reaction time data from our analysis. We had 60 trials from the driving simulator and 60 trials from the pedestrian simulator with recorded reaction times that we used for our analysis.

In Table 9, we analyzed trials where the AV yielded and an eHMI was present. There were a total of 82 such trials with a recorded reaction time. Males had shorter reaction times than females and participants that reported high initial levels of trust in AVs also recorded shorter reaction times (Table 9). Reaction times in trials where the participant understood the AVs intentions were shorter than trials where the participant did not understand the AV's intentions.

Table 9. Reaction time in relation to participant characteristics (the AVs yielded in all trials and all AVs had eHMIs; results are average of driving and pedestrian trials).

	n	Reaction Time (sec.)	Standard Deviation (sec.)
Female	43	3.32	4.31
Male	39	2.26	2.30
Low Initial Trust	41	3.46	4.37
High Initial Trust	41	2.17	2.26
Understood	49	4.74	3.30
Did Not Understand	10	8.10	5.63

Reaction times were shorter in trials where the AV yielded right-of-way to the participant (Table 10). This is likely because the eHMI did not display any message when it was not yielding in the driving simulator, which may have confused the participants. Reaction times were also shorter in trials where there was an eHMI present. As detailed above, the pedestrian reaction times were shorter, likely a result of the layout of the scenarios themselves. The Text on Grille eHMI had the shortest reaction time. This aligns with the perception outcomes as the Text on Grille eHMI also had some of the highest perception scores. The Text on Roof and Side Mirror Arrows eHMIs also reported relatively short reaction times. Of note is the Text on Driver Door eHMIs relatively long reaction time, which was a result of this eHMI only being available for the driving and biking simulators. When compared to other results just from the driving simulator, the Text on Driver Door reaction time was similar to the reaction times for the other text-based eHMIs.

Table 10. Reaction time in relation to AV and eHMI characteristics (unless otherwise noted, the AVs yielded in all trials, all AVs had eHMIs, and results are average of driving and pedestrian trials).

	n	Reaction Time (sec.)	Standard Deviation (sec.)
No Yield (w/ an eHMI)	16	5.98	3.43
Yield (w/ an eHMI)	82	2.82	3.54
Without eHMI (yielded)	19	6.51	4.23
With eHMI (yielded)	82	2.82	3.54
Drive	42	4.47	3.87
Pedestrian	40	1.08	2.01
Text on Driver Door	8	3.40	1.76
Text on Grille	18	1.59	2.01
Text on Roof	18	2.34	1.97
Side Mirror Arrows	19	1.92	1.75
LED Windshield	19	5.09	5.87
No eHMI	19	6.51	4.23

A standard perception-reaction time assumption used today in the transportation engineering field is 2.5 seconds. Several of the eHMIs' reaction times hover around or below this value (Table 10). However, results suggest that if the LED Windshield is used on new AV technologies or even no eHMI at all, reaction times may increase significantly. This change could have important implications when considering traffic operations in the future.

Based on linear regression results, reaction times decreased as participants went through more trials (Table 11). The coefficient can be interpreted as reaction time decreasing by 0.134 seconds for each additional trial, and results were statistically significant at greater than 99% confidence. This is evidence that the AV and eHMI acclimation noted in the perception results was also present in the participants' behaviors.

Table 11. Linear regression describing relationship between trial number and reaction time (results statistically significant at 95% confidence in bold).

	n	Coefficient	p-value
Reaction Time	120	-0.134	**0.006**

Again, this acclimation primarily took place in the eHMIs that had the worst outcomes in the early trials. In the first simulator of the first session, AVs with the LED Windshield eHMI and those with no eHMI had reaction times of around 9.00-10.50 seconds, significantly longer than any of the other eHMIs. The reaction times of those eHMI configurations quickly decreased as participants became acclimated. The LED Windshield had reaction times of around 3.50 seconds by the end of the testing, which was only slightly longer than the three best-performing eHMIs (Text on Grille, Text on Roof, and Side Mirror Arrows), which all concentrated around 1.50 seconds. AVs with no eHMIs retained significantly longer reaction times, remaining at around 6.50 seconds at the end of testing. The three best-performing eHMIs (Text on Grille, Text on Roof, and Side Mirror Arrows), which all concentrated around 1.50 seconds at the end of testing, all began at around 1.50-2.50 seconds, meaning that participants understood them immediately and there was very little acclimation needed.

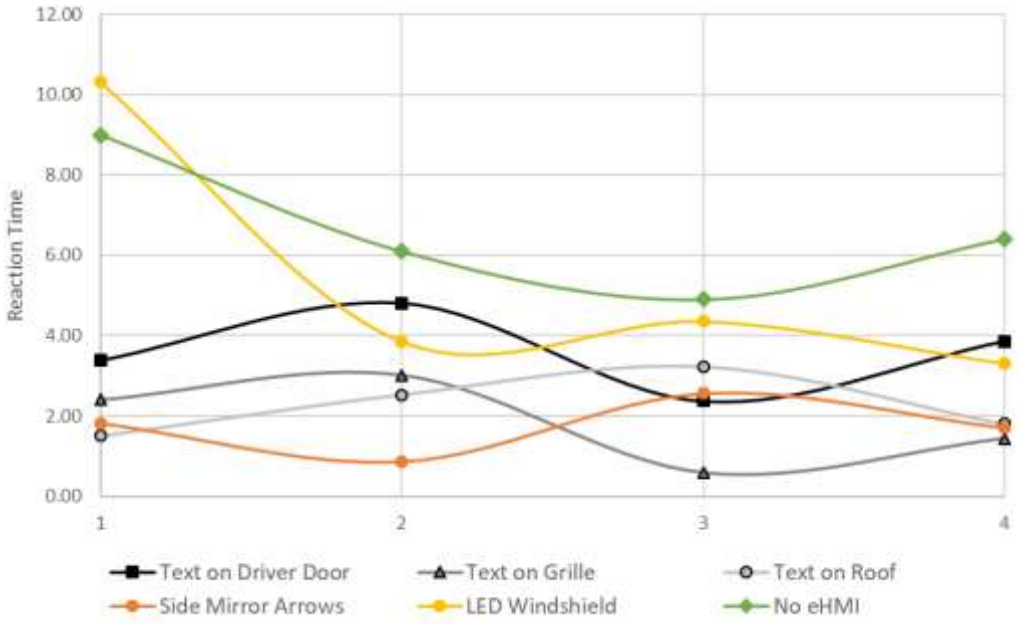

Figure 20. Average reaction time for different eHMI designs over time (the AVs yielded in all trials; results are average of driving and pedestrian) (the x-axis is time with 1=first simulator of first session; 2=second simulator of first session; 3=first simulator of second session; 4=second simulator of second session).

Again, while many of the behavioral results above provide interesting insight, more data collection is needed to provide more confidence in the results. While the current work helped us to optimize our data collection methodologies and identified interesting trends, more data is needed before conclusions can be drawn about absolute values of reaction times with AVs.

6. CONCLUSIONS

The presence of an eHMI significantly improved perceptions. The presence of an eHMI was the strongest predictor of understanding, comfort, trust, and acceptance outcomes in the statistical models when controlling for all other variables. Task load was also improved by eHMIs but did not reach statistical significance (likely because eHMIs took some effort for interpretation). Acclimation effects noted throughout the testing support the importance of eHMIs because while the perceptions of AVs with eHMIs largely improved over time, the understanding, trust, and acceptance of AVs without eHMIs actually got worse by the end of the testing.

In terms of which eHMI design was correlated with the best perceptions, there is a clear divide between text-based eHMIs and non-text eHMIs, with text-based eHMIs reporting better perception scores. The Text on Grille and Text on Driver Door eHMIs typically received the best scores while LED Windshield generally received the worst scores. This supports the need for future research on the positioning of the eHMIs as the Text on Driver Door eHMI was on the side of the AV. Longitudinally, improvements in perceptions were largely seen in the second tier of eHMIs (typically Text on Roof and Side Mirror Arrows). These second tier eHMIs started with lower scores that then improved throughout testing, whereas the favored eHMIs (Text on Grille and Text on Driver Door) had high scores from the beginning that largely stayed the same.

There were perception acclimation effects detected, but they had less of an impact than the presence of an eHMI. The largest acclimation effects were in comfort (which increased throughout testing for all eHMI configurations) and task load (which decreased throughout testing for all eHMI configurations). Trust and acceptance were more stable throughout testing (average increases of only 0.1 points and 0.2 points from beginning to end, respectively), with LED Windshield and AVs without an eHMI actually seeing decreases in trust and acceptance. Accordingly, comfort was the only perception that had a statistically significant relationship with trial number while task load and comfort were the only perceptions that had statistically significant relationships with session number. This suggests that while task load and comfort improve as participants experience AVs regardless of eHMI configuration, trust and acceptance are much more dependent upon appropriate eHMIs.

As expected, participants reported improved perceptions when the AV yielded to them. However, perceptions improved similarly over time for both yielding and non-yielding AVs.

All the measured perceptions improved for the driving and pedestrian simulators but stayed relatively stable or even got worse for the biking simulator. However, this may be because of the design of the simulators themselves. The mode of the simulator was non-significant in the understanding, comfort, and acceptance statistical models and was relatively weak in the others, suggesting that mode is not as important as eHMI presence, acclimation, or yielding.

Importantly, perception outcomes had weaker relationships with participant characteristics than with AV characteristics. This suggests that while factors such as age, gender, and preconceived biases toward or against AVs will impact road users' preferences, appropriate eHMI design, AV behavior, and acclimation may be able to help overcome any negative perceptions.

While behavioral outcomes should be interpreted with caution because of low participant sample sizes, overall reaction times with an eHMI averaged 2.82 seconds which was close to a standard assumption of 2.50 seconds used in transportation engineering. Behavioral results largely mirrored

perception results in that significant reductions in reaction time were observed with the presence of an eHMI (3.69 second reduction), yielding (3.16 second reduction), and acclimation (0.134 second reduction per trial). The LED Windshield had significantly longer reaction times than the other eHMIs and a participants' gender and initial reported trust of AVs had smaller impacts on reaction time. Results suggest that eHMI design, AV behavior, and acclimation are most impactful in terms of reaction time.

REFERENCES

1. Anderson, J.M., N. Kalra, K.D. Stanley, P. Sorensen, C. Samaras, and T.A. Oluwatola. *Autonomous vehicle technology: How to best realize its social benefits*. Santa Monica, CA: RAND Corporation, 2014.

2. Eugensson, A., M. Brännström, D. Frasher, M. Rothoff, S. Solyom, and A. Robertsson. Environmental, safety legal and societal implications of autonomous driving systems. *International Technical Conference on the Enhanced Safety of Vehicles*. 2013. 334.

3. Fagnant, D.J. and K. Kockelman. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 2015. 77: 167-181.

4. Adnan, N., S.M. Nordin, M.A. Bahruddin, and M. Ali. How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. *Transportation Research Part A: Policy and Practice*, 2018. 118: 819-836.

5. Basu, C. and M. Singhal. Trust dynamics in human autonomous vehicle interaction: A review of trust models. *2016 AAAI Spring Symposium Series*, 2016.

6. Choi, J.K. and Y.G. Ji. Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 2015. 31(10): 692-702.

7. Ekman, F., M. Johansson, and J. Sochor. Creating appropriate trust for autonomous vehicle systems: A framework for HMI Design. *IEEE Transactions on Human-Machine Systems*, 2018. 48(1): 95-101.

8. Hulse, L.M., H. Xie, and E.R. Galea. Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age. *Safety Science*, 2018. 102: 1-13.

9. Becker, F. and K.W. Axhausen. Literature review on surveys investigating the acceptance of automated vehicles. *Transportation*, 2017. 44(6): 1293-1306.

10. Böckle, M.P., A.P. Brenden, M. Klingegård, A. Habibovic, and M. Bout. SAV2P: Exploring the impact of an interface for shared automated vehicles on pedestrians' experience. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, 2017. 136-140.

11. Habibovic, A., V.M. Lundgren, J. Andersson, M. Klingegård, T. Lagström, A. Sirkka, J. Fagerlönn, et al. Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology*, 2018. 9: 1336.

12. Habibovic, A., J. Andersson, V.M. Lundgren, M. Klingegård, C. Englund, and S. Larsson. External vehicle interfaces for communication with other road users? *Road Vehicle Automation*, 2019. 5: 91-102.

13. NHTSA. *Traffic Safety Facts: Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. DOT HS 812 115. 2015.

14. Azevedo C.L., K. Marczuk, S. Raveau, H. Soh, M. Adnan, K. Basak, H. Loganathan, N. Deshmunkh, D.H. Lee, E. Frazzoli, and M. Ben-Akiva. Microsimulation of demand and supply of autonomous mobility on demand. *Transportation Research Record*, 2016. 2564(1): 21-30.

15. Jian J.Y., A.M. Bisantz, C.G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 2000. 4(1): 53-71.

16. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics. 2009 Jan;1(1):71-81.

17. Van Der Laan JD, Heino A, De Waard D. A simple procedure for the assessment of acceptance of advanced transport telematics. Transportation Research Part C: Emerging Technologies. 1997 Feb 1;5(1):1-0.

18. Burns C.G., L. Oliveira, P. Thomas, S. Iyer, and S. Birrell. Pedestrian decision-making responses to external human-machine interface designs for autonomous vehicles. *IEEE Intelligent Vehicles Symposium*, 2019. 4: 70-75.

19. Hart S.G. and L.E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 1988. 52: 139-183).

20. Kim H.K., J. Park, Y. Choi, and M. Choe. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics*, 2018. 69: 66-73.

21. Haboucha, C.J., R. Ishaq, and Y. Shiftan. User preferences regarding autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 2017. 78: 37-49.

22. Bansal, P., K.M. Kockelman, and A. Singh. Assessing public opinions of and interest in new vehicle technologies: An Austin perspective. *Transportation Research Part C: Emerging Technologies*, 2016. 67: 1-14.

23. Zmud, J., I.N. Sener, and J. Wagner. *Consumer acceptance and travel behavior: impacts of automated vehicles*. No. PRC 15-49 F. Texas A&M Transportation Institute, 2016.

24. Ford Media Center. *Ford, Virginia Tech Go Undercover to Develop Signals that Enable Autonomous Vehicles to Communicate with People*. 2017.

25. Grimm, D.K., R.J. Kiefer, L.S. Angell, R.K. Deering, and C.A. Green. *Vehicle to entity communication*. U.S. Patent 8,253,589, issued August 28, 2012.

26. Hillis, W.D., K.I. Williams, T.A. Tombrello, J.W. Sarrett, L.W. Khanlian, A.L. Kaehler, and R. Howe. *Communication between autonomous vehicle and external observers*. U.S. Patent 9,475,422, issued October 25, 2016.

27. Sweeney, M., T. Pilarski, W.P. Ross, and C. Liu. *Light output system for a self-driving vehicle*. U.S. Patent 10,160,378, issued December 25, 2018.

28. Urmson, C.P., I.J. Mahon, D.A. Dolgov, and J. Zhu. *Pedestrian notifications*. U.S. Patent 9,196,164, issued November 24, 2015.

29. De Clercq, K., A. Dietrich, J.P.N. Velasco, J. de Winter, and R. Happee. External human-machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. *Human Factors*, 2019. 61(8): 1353-1370.

30. Weber, F., R. Chadowitz, K. Schmidt, J. Messerschmidt, and T. Fuest. Crossing the Street Across the Globe: A Study on the Effects of eHMI on Pedestrians in the US, Germany and China. *International Conference on Human-Computer Interaction*, 2019. 515-530.

31. Bazilinskyy, P., D. Dodou, and J.D. Winter. Survey on eHMI concepts: The effect of text, color, and perspective. *Transportation research part F: traffic psychology and behavior*, 2019. 67: 175-194.

32. Ackermann, C., M. Beggiato, S. Schubert, and J.F. Krems. An experimental study to investigate design and assessment criteria: What is important for communication between pedestrians and automated vehicles? *Applied Ergonomics*, 2019. 75: 272-282.

33. Merat, N., T. Louw, R. Madigan, M. Wilbrink, and A. Schieben. What externally presented information do VRUs require when interacting with fully Automated Road Transport Systems in shared space? *Accident Analysis & Prevention*, 2018. 118: 244-252.

34. Zhang, J., E. Vinkhuyzen, and M. Cefkin. Evaluation of an autonomous vehicle external communication system concept: a survey study. *International conference on applied human factors and ergonomics*, 2017. 650-661.

35. Clamann, M., M. Aubert, and M.L. Cummings. *Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles*. No. 17-02119. 2017.

36. Troel-Madec, M., L. Boissieux, S. Borkoswki, D. Vaufreydaz, J. Alaimo, S. Chatagnon, and A. Spalanzani. eHMI positioning for autonomous vehicle/pedestrians interaction. *Proceedings of the 31st Conference on l'Interaction Homme-Machine: Adjunct*, 2019. 1-8.

37. Ferenchak, N.N and S. Shafique. Pedestrians' Perceptions of Autonomous Vehicle External Human-Machine Interfaces. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 2021. 8(3): 034501.

38. Balakrishnan, S., S. Moridpour, and R. Tay. Sociodemographic Influences on Injury Severity in Truck-Vulnerable Road User Crashes. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2019. 5(4): 04019015.

39. Chandon, P., V.G. Morwitz, and W.J. Reinartz. Do intentions really predict behavior? Self-generated validity effects in survey research. *Journal of Marketing*, 2005. 69(2): 1-14

40. Feldman, J.M. and J.G. Lynch. Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, 1988. 73(3): 421.

41. Laumann, K. and M.R. Skogstad. Challenge to Collect Empirical Data for Human Reliability Analysis—Illustrated by the Difficulties in Collecting Empirical Data on the Performance-Shaping Factor Complexity. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 2020. 6(1).

42. Lee, E., M.Y. Hu, and R.S. Toh. Are consumer survey results distorted? Systematic impact of behavioral frequency and duration on survey response errors. *Journal of Marketing Research*, 2000. 37(1): 125-133

43. Nicholson, A. Road safety: risk management perspective. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2020. 6(1): 04019017

44. Tzioutziou, A. and Y. Xenidis. The Impact of Weighting Methods and Behavioral Attitudes on the Weighting Process in Decision-Making. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 2020. 6(1).

45. Grivokostopoulou F., I. Perikos, and I. Hatzilygeroudis. An innovative educational environment based on virtual reality and gamification for learning search algorithms. *IEEE Eighth International Conference on Technology for Education (T4E)*, 2016. 8: 110-115.

46. Thomson J.A., A.K. Tolmie, H.C. Foot, K.M. Whelan, P. Sarvary, and S. Morrison. Influence of virtual reality training on the roadside crossing judgments of child pedestrians. *Journal of Experimental Psychology: Applied*, 2005. 11(3): 175.

47. Calvi, A., F. D'Amico, L. Ciampoli, and C. Ferrante. Evaluation of driving performance after a transition from automated to manual control: a driving simulator study. *Transportation Research Procedia*, 2020. 45: 755-762.

48. Palinko, O., A. Kun, A. Shyrokov, and P. Heeman. Estimating cognitive load using remote eye tracking in a driving simulator. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 2010. 141-144.

49. Saupp, L., B. Schwarz, M. Schwalm, and L. Eckstein. Evaluation of active Drivesticks as alternative controls in a high fidelity driving simulator. *Automotive and Engine Technology*, 2019. 4(1): 37-44.

50. Shakouri, M., L. Ikuma, F. Aghazadeh, and I. Nahmens. Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: The case of highway work zones. *International Journal of Industrial Ergonomics*, 2018. 66: 136-145.

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

52. Lerner, N.D. Brake perception-reaction times of older and younger drivers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1993. 37(2): 206-210.

53. Ye, X., X. Wang, S. Liu, and A.P. Tarko. Feasibility study of highway alignment design controls for autonomous vehicles. *Accident Analysis & Prevention*, 2021. 159: 106252.

54. Rouchitsas, A. and H. Alm. External human–machine interfaces for autonomous vehicle-to-pedestrian communication: a review of empirical work. *Frontiers in Psychology*, 2019. 10: 2757.

55. Petzoldt, T., K. Schleinitz, and R. Banse. Potential safety effects of a frontal brake light for motor vehicles. *IET Intelligent Transport Systems*, 2018. 12(6): 449-453.

56. Ferenchak, N.N. and W.E. Marshall. Is bicycling getting safer? Bicycle fatality rates (1985–2017) using four exposure metrics. *Transportation research interdisciplinary perspectives*, 2020. 8: 100219.

57. Ferenchak, N.N. and M.G. Abadi. Nighttime pedestrian fatalities: a comprehensive examination of infrastructure, user, vehicle, and situational factors. *Journal of Safety Research*, 2021.

58. Ferenchak, N.N. and W.E. Marshall. Quantifying suppressed child pedestrian and bicycle trips. *Travel Behaviour and Society*, 2020. 20: 91-103.

59. Ferenchak, N.N. and W.E. Marshall. Suppressed child pedestrian and bicycle trips as an indicator of safety: Adopting a proactive safety approach. *Transportation research part A: policy and practice*, 2019. 124: 128-144.

60. Marcotte, T., E. Roberts, T. Rosenthal, R. Heaton, H. Bentley, and I. Grant. Test-Retest Reliability of Standard Deviation of Lane Position as Assessed on a PC-Based Driving Simulator. *Proceedings of the 2nd International Driving Symposium on Human Factors in Driver Assessment*, 2005. Training and Vehicle Design: Driving Assessment, 2.

61. Iwata, M., K. Iwamoto, I. Kitajima, T. Nogi, K. Onishi, Y. Kajiyama, I. Nishino, M. Ando, and N. Ozaki. Validity and reliability of a driving simulator for evaluating the influence of medicinal drugs on driving performance. *Psychopharmacology*, 2021. 238(3): 775-86.

62. Wynne, R., V. Beanland, and P. Salmon. Systematic review of driving simulator validation studies. *Safety Science*, 2019. 117: 138-151.

63. Shechtman, O., S. Classen, K. Awadzi, and W. Mann. Comparison of driving errors between on-the-road and simulated driving assessment: a validation study. *Traffic Injury Prevention*, 2009. 10(4): 379-385.

64. Ranney, T. (2011). *Psychological Fidelity: Perception of Risk*.

65. Meuleners, L. and M. Fraser. A validation study of driving errors using a driving simulator. *Transportation Research Part F-traffic Psychology and Behaviour*, 2015. 29: 14-21.

66. Underwood, G., D. Crundall, and P. Chapman. Driving simulator validation with hazard perception. *Transportation Research Part F-traffic Psychology and Behaviour*, 2011. 14(6): 435-446.

67. Godley, S., T. Triggs, and B. Fildes. Driving simulator validation for speed research. *Accident Analysis & Prevention*, 2002. 34(5): 589-600.

68. Shoman, M. and H. Imine. Bicycle Simulator Improvement and Validation. *IEEE Access*, 2021. 9: 55063-55076.

69. Schwebel, D., J. Gaines, and J. Severson. Validation of virtual reality as a tool to understand and prevent child pedestrian injury. *Accident Analysis & Prevention*, 2008. 40(4): 1394-1400.

70. Hirata, T., T. Yai, and T. Takagawa. Development of the Driving Simulation System MOVIC-T4 and Its Validation Using Field Driving Data. *Tsinghua Science & Technology*, 2007. 12(2): 141-150.

71. Hartfiel, B. and R. Stark. Validity of primary driving tasks in head-mounted display-based driving simulators. *Virtual Reality*, 2021. 1-15.

72. O'Hern, S., J. Oxley, and M. Stevenson. Validation of a bicycle simulator for road safety research. *Accident Analysis & Prevention*, 2017. 100: 53-58.

73. Bhagavathula, R., B. Williams, J. Owens, and R. Gibbons. *Virtual Reality as a Tool to Evaluate Pedestrian Safety*, 2020.

74. Charness, G., U. Gneezy, U., and M. Kuhn. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, 2012. 81(1): 1-8.

75. Rosenthal, R. *Experimenter effects in behavioral research*, 1966.

76. White, A. The Influence of Experimenter Motivation, Attitudes, and Methods of Handling Subjects on Psi Test Results. *Handbook of Parapsychology*, 1977. 273-301.

77. Figueira, A. and A. Larocca. Analysis of the factors influencing overtaking in two-lane highways: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2020. 69: 38-48.

78. Jackson, J. and R. Blackman. A driving-simulator test of Wilde's risk homeostasis theory. *Journal of Applied Psychology*, 1994. 79(6): 950-958.

79. Ogourtsova, T., M. Kalaba, I. Gelinas, N. Korner-Bitensky, and M. Ware. Cannabis use and driving-related performance in young recreational users: a within-subject randomized clinical trial. *CMAJ Open*, 2018. 6(4).

80. Reinolsmann, N., W. Alhajyaseen, T. Brijs, A. Pirdavani, Q. Hussain, and K. Brijs. Investigating the impact of dynamic merge control strategies on driving behavior on rural and urban expressways – A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2019. 65: 469-484.

81. Donmez, B., L. Boyle, and J. Lee. Driving Simulator Experiments: Power for Repeated Measures vs. Completely Randomized Design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2006. 50(21): 2336-2339.

82. Frank, L., J. Casali, and W. Wierwille. Effects of Visual Display and Motion System Delays on Operator Performance and Uneasiness in a Driving Simulator. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1987. 31(5): 492-496.

83. Lee, Y.C. Measuring Drivers' Frustration in a Driving Simulator. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2010. 54(19): 1531-1535.

84. Gershon, P., A. Ronen, T. Oron-Gilad, and D. Shinar. The effects of an interactive cognitive task (ICT) in suppressing fatigue symptoms in driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2009. 12(1): 21-28.

85. Large, D., G. Burnett, A. Morris, A. Muthumani, and R. Matthias. A Longitudinal Simulator Study to Explore Drivers' Behaviour During Highly-Automated Driving. *International Conference on Applied Human Factors and Ergonomics*, 2017. 583-594.

86. Ariën, C., K. Brijs, T. Brijs, W. Ceulemans, G. Vanroelen, E. Jongen, S. Daniels, and G. Wets. Does the effect of traffic calming measures endure over time? A simulator study on the influence of gates. *Transportation Research Part F: Traffic Psychology and Behaviour,* 2014. 22: 63-75.

87. Calvi, A. Investigating the effectiveness of perceptual treatments on a crest vertical curve: a driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2018. 58: 1074-1086.

88. Awan, H., A. Pirdavani, A. Houben, S. Westhof, M. Adnan, and T. Brijs. Impact of perceptual countermeasures on driving behavior at curves using driving simulator. *Traffic Injury Prevention,* 2019. 20(1): 93-99.

89. Turkington, P., M. Sircar, D. Saralaya, and M. Elliott. Time course of changes in driving simulator performance with and without treatment in patients with sleep apnoea hypopnoea syndrome. *Thorax,* 2004. 59(1): 56-59.

90. Duchek, J., D. Carr, L. Hunt, C. Roe, C. Xiong, K. Shah, and J. Morris. Longitudinal driving performance in early-stage dementia of the Alzheimer type. *Journal of the American Geriatrics Society*, 2003. 51(10): 1342-1347.

91. Chang, E., H. Kim, and B. Yoo. Virtual Reality Sickness: A Review of Causes and Measurements. *International Journal of Human-computer Interaction,* 2020. 36(17): 1658-1682.

92. Dużmańska, N., P. Strojny, and A. Strojny. Can Simulator Sickness Be Avoided? A Review on Temporal Aspects of Simulator Sickness. *Frontiers in Psychology*, 2018. 9: 2132.

93. Geršak, G., H. Lu, and J. Guna. Effect of VR technology matureness on VR sickness. *Multimedia Tools and Applications*, 2020. 79(21): 14491-14507.

94. Schultheis, M., J. Rebimbas, R. Mourant, and S. Millis. Examining the usability of a virtual reality driving simulator. *Assistive Technology,* 2007. 19(1): 1-8.

95. Kennedy, R., N. Lane, K. Berbaum, and M. Lilienthal. Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 1993. 3(3): 203-220.

96. Deb, S., L.J. Strawderman, and D.W. Carruth. Investigating pedestrian suggestions for external features on fully autonomous vehicles: A virtual reality experiment. *Transportation research part F: traffic psychology and behavior,* 2018. 59: 135-149