A comparison of two methodologies for subjective evaluation of comfort in automated vehicles

Chen Peng[1], Foroogh Hajiseyedjavadi[1 2], and Natasha Merat[1]

1 Institute for Transport Studies, University of Leeds, Leeds, UK. C.Peng@leeds.ac.uk; N.Merat@its.leeds.ac.uk

2 School of Engineering and the Built Environment, Birmingham City University, Birmingham, UK. Foroogh.hajiseyedjavadi@bcu.ac.uk

Abstract

This paper compared two different methodologies, used in two driving simulator studies, for real-time evaluation of comfort imposed by the driving style of different Automated Vehicle (AV) controllers. The first method provided participants with two options for assessing three different AV controllers. Participants rated each controller in terms of whether or not it was comfortable/safe/natural, when it navigated a simulated road. The evaluation was either positive (yes) or negative (no), indicated by pressing one of two buttons on a handset. In the second study, an 11-point Likert-type scale (from -5 to +5) was used to evaluate the extent to which a controller's driving style was "comfortable" and/or "natural", separately. Participants provided this evaluation for three different AV controllers. Here, they were instructed to utter a number from the scale, at designated points during the drive. To understand which method is better for such evaluations, we compared the data collected from the two studies, and investigated the patterns of data obtained for the two methodologies. Results showed that, despite the multiple response options provided by the 11-point scale, a similar pattern was seen to that of the binary method, with more positive responses provided for all controllers. The Likert scale is useful for identifying differences because of the multiple levels of responses. However, allowing people to present their ratings as often as they want, also makes the binary technique useful for such evaluations.

Introduction

One of the factors that contributes to the broad acceptance of Automated Vehicles (AVs), is users' evaluation of comfort of the automated driving style [1], [2]. For automated driving, comfort is more than ensuring acceptable levels of noise, vibrations and temperature etc. of the vehicle, which are aspects also applied to traditional, manually driven, vehicles[3], [4]. For higher levels of automation (SAE Levels 4 and 5) [5], the role of the on-board occupant shifts from an active operator, to a passive user of the vehicle. Here, the user will have less control of the vehicle, and less ability to predict the vehicle's behaviours, which might lead to an uncomfortable ride [6], [7]. How a controller negotiates the road, and whether or not this is the same as how the user handles the vehicles, is also thought to affect comfort [8][9]. Studies in this field have used a range of concepts to describe comfort, including familiar/natural manoeuvres that are likely to fulfil the rider's expectations of AV manoeuvres, and perceived safety, induced by the suitable distance kept with other on-road obstacles [2], [10]. In a number of studies, the description of comfort is very much based on the emphasising the users' subjective state and feelings. For example, comfort in AVs is described as *"a subjective, pleasant state of relaxation given by confidence and an apparently safe vehicle operation, which is achieved by the removal or absence of uneasiness and distress"* (p. 1019) [1], or *"the subjective feeling of pleasantness of driving/riding in a vehicle in the absence of both physiological and psychological stress"* (p. 12) [11]. However, currently, there is no commonly agreed definition of comfort, and there are no widely employed behavioural techniques for measuring this concept.

An AV's controllers can navigate the road in different ways. For example, in terms of lateral control, it can precisely follow the lane centre [8], deviate from the lane centre within an acceptable boundary [8], or adjust its position, based on road-based objects and surrounding features (e.g. high hedges and parked cars) [12]. To understand how the user wants to be driven by an AV, it is important to measure their perceived comfort in different automated driving conditions. Several measurements have been adopted for such studies. The majority of these studies have conducted evaluations after participants complete the whole experimental drive. For example, [13] asked participants to rate their perceived comfort and enjoyment after each drive using a questionnaire composed of 32 items. [6] provided a one-item rating scale after each trial for participants to evaluate if the deceleration and lane-changing manoeuvres of automated vehicles were comfortable. Similarly, [14] instructed participants to evaluate driving behaviours of different driving styles (simulated by a human driver using the Wizard-of-Oz technique) regarding comfort, pleasantness, and safety, separately, after each driving session. These post-hoc ratings provide some insights about comfortable automated driving. However, they are based on the participant's memory of the finished drive, and only depict the experience of the entire session. As several elements influence ride comfort, including the AV's speed, and how it negotiates different road geometries, and road-based obstacles [8], [9], real-time assessments are more informative than post-session evaluations, for capturing users' feedback in this context. Previous studies in this context have used a range of methods for real-time subjective feedback. Examples include a handset control for reporting discomfort [1], vehicle pedals for expressing satisfaction with the AV's speed [15], a rotary knob for measuring perceived vehicle motion, and a slider for assessing perceived pressure from passing through tunnels for train passengers [16]. These tools allow participants to give feedback about more specific manoeuvres, or situations, when the vehicle is operating, rather than after the whole trip. The use of real-time feedback provides knowledge on how the AV should drive in response to more dynamic changes in the road, such as changes in speed for different road geometries. However, a comparison of different methods and scales that assess real-time feedback in AVs, especially regarding driving comfort, is currently lacking.

Taking real-time and post-experiment assessments together, the range of scales used for these evaluations have varied between binary and multi-scale (e.g., 5, 7, 11, 100) levels. A recent between-subject study compared differences in the number of response options (incl. 3, 5, 7 and 11) of the Usability Metric for User Experience (UMUX-LITE) questionnaire, that is used to assess perceived usability of an auto insurance Web site [17]. The same questionnaire items were used for this online survey, while the number of response options varied for different groups of participants. Results suggested that there were no differences between the 5, 7, and 11 response options, whereas weak reliability and correlation was observed between the scales, with the 3-point option. However, this study was used to assess perceived usability of a web page, based on an elaborate standardized questionnaire. To our knowledge, there is currently a distinct lack of studies about the optimal number of response options, and methods, used to evaluate the comfort of an AV controller.

In the present work, we compared two different methodologies, used in two driving simulator studies, conducted as part of the UK-funded HumanDrive project, measuring real-time subjective experience of different highly automated driving styles [8], [9]. For each study, different AV controllers negotiated a range of UK roads, which differed in terms of their geometry, speed limit, and presence of road-based features. Participants were required to rate their ride experience when being driven by each AV controller, using one of two different methods. The two methodologies differed in terms of assessment tools (handset-based versus verbal) and the number of response options (binary versus 11-point).

Method

For study one [8], participants rated three automated driving styles. These were two model-based human-like controllers, at either slow or fast speeds, and a playback of the participant's own manual drive, named Slow, Fast, and Replay, respectively. During the drive, participants used two buttons of an X-box handset, to indicate if they found the controllers comfortable/safe/natural, pressing the right button for Yes, and the left button for No (see Figure 12). Therefore, only one of two buttons was used for evaluating the overall pleasant or unpleasant affect experienced of the three concepts (i.e., comfort, safety and naturalness). This method was used based on the assumption that these three concepts did not differ in terms of reflecting a pleasant drive, as a number of studies have used these concepts interchangeably (e.g., [10], [18], [19]). Participants were asked to provide their response immediately after hearing an auditory beep, which was presented at different road segments, throughout the drives. In addition, they were encouraged to press the two buttons anytime along the drive.

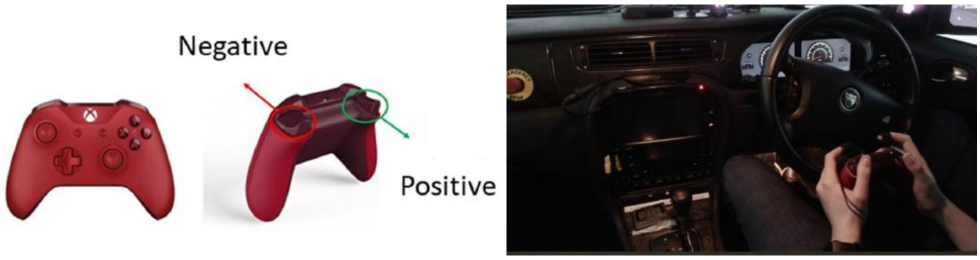

Figure 12. The handset used in Study 1 for rating AV controllers [8]

For study two [9], three automated controllers were evaluated. These included two recorded and replayed from human drivers, and one based on a machine learning (ML) algorithm, named Defensive, Aggressive, and Turner, respectively. Using an 11-point Likert scale (see Figure 13), users provided a verbal response to rate the “comfort” and “naturalness” of the controllers, separately: very comfortable/natural (+5) to very uncomfortable/unnatural (-5). Participants were provided with definitions of the two concepts at the beginning of the study, via an information sheet. They were taught how to use the scale and were able to practice it in a practice drive. A comfortable drive was defined as “*a driving style that does not cause any feeling of uneasiness or discomfort*”, and a natural drive was defined as “*a driving style that is closest to your own driving*”. Due to a lack of clear description of each concept, these definitions were created following an expert group meeting [9]. Participants evaluated each controller in terms of comfort or naturalness by speaking out a value from the 11-point scale, after hearing an auditory beep, which coincided with different road sections (a total of 24 sections), during the drive. They also provided an overall rating for the controller after completing the entire drive, which is not included in the present study.

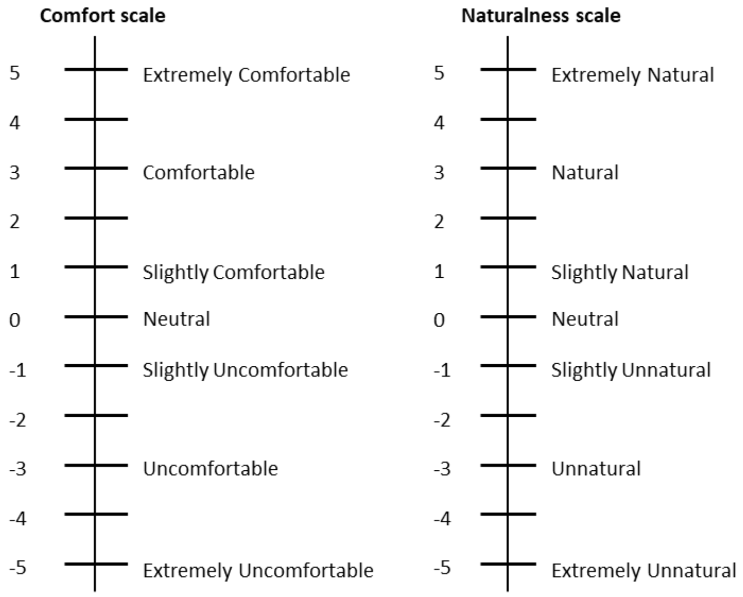

Figure 13. The scale used in Study 2 for AV controller evaluation [9]

It is notable that, as a follow-up of study one, the two concepts were defined and evaluated separately in study two. The aim here was to establish differences in preference between human-like and machine-like AV controllers, and especially whether or not natural manoeuvres are comfortable, as suggested by [2]. Therefore, an 11-point scale was used, to provide more options for the participants on both the positive and negative side, than that used in study one. A summary of the two methods is provided in Table 6.

All participants (24 in study one, and 24 in study two) were recruited, using the University of Leeds Driving Simulator database. All participants provided informed consent to take part in each study. These two studies were approved by the University of Leeds Ethics Committee (LTTRAN-086).

Table 6. Summary of the two methods used in the two simulator studies

	Study 1	Study 2
Questions	"I found the behaviour of the controller safe/ natural/ comfortable"	Rate the driving style in terms of comfort. OR Rate the driving style in terms of naturalness.
Options	Yes or No	-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5
Tools	Press one of two buttons on an X-box handset	Speak out the number
Frequency	Rate after an auditory beep; Free to press buttons throughout the drive.	Rate after an auditory beep during the drive; Rate after the entire drive.

Results

We compared the two methods, by examining the pattern of responses obtained from the two studies.

In study 1, the average number of positive and negative button presses were calculated for each controller [8], as a function shown below. This was calculated for positive and negative assessments, respectively.

$$\begin{aligned} &Average\ button\ presses \\ &= \frac{The\ total\ number\ of\ button\ presses\ of\ a\ participant\ in\ an\ experimental\ condition}{The\ number\ of\ exposures\ to\ this\ experimental\ condition} \end{aligned}$$

In study 2, as described above, evaluations were provided based on an 11-point scale. To compare with study 1, we allocated "positive" for ratings larger than 0, and "negative" for ratings equal to or smaller than 0. For each participant, the number of positive and negative ratings obtained from a total of 24 driving segments were computed. For example, participant 1 gave 23 positive ratings and one negative rating to the the Defensive controller, for the 24 driving segments. This value was then divided by the number of evaluations for each controller, which was 24, for comparable reasons. After that, the means and standard errors of positive and negative ratings across participants were calculated, as illustrated in .

The common pattern

It is worth mentioning that the AV controllers evaluated in the two different studies varied in modelling algorithms. Moreover, as the simulated road geometries were different between the two studies, the way that each controller negotiated the road was different. However, as shown in Figure 3, a common trend in responses was seen for both methods, whereby participants provided more positive evaluations for all controllers, regardless of whether their evaluation was based on a combination of comfort/safety/naturalness (study 1), or "comfort" and "naturalness" separately (study 2).

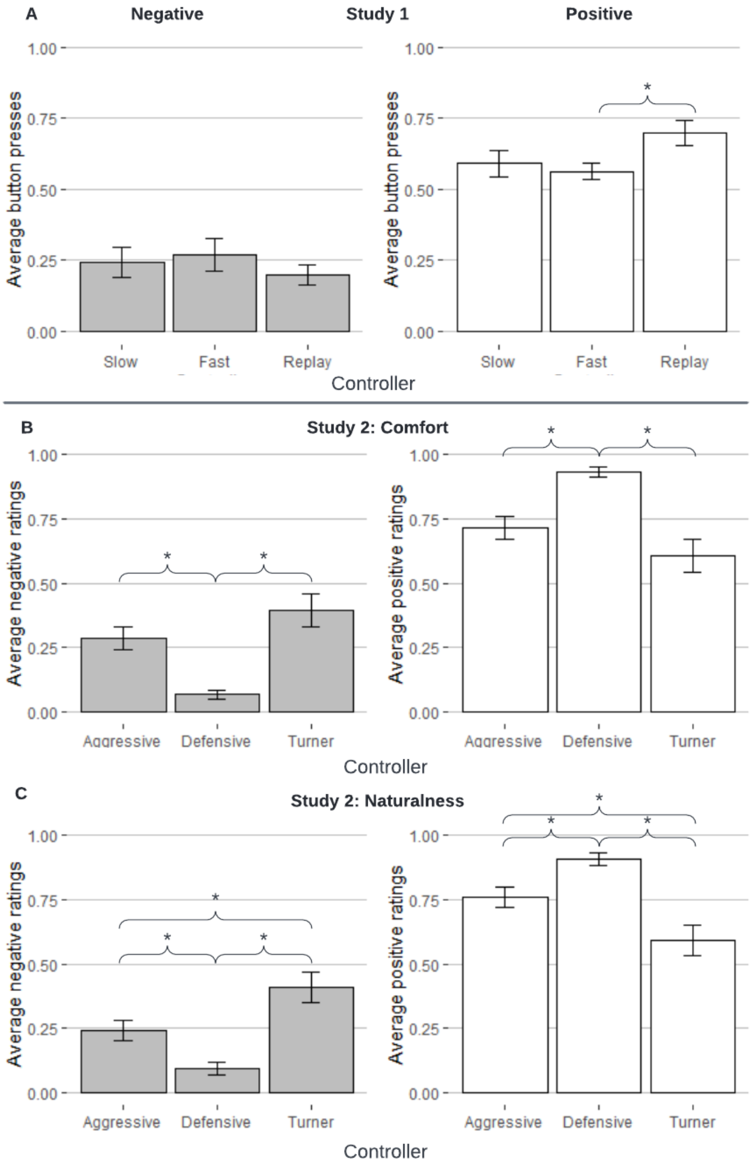

Figure 14. Average negative and positive ratings for controllers, in study 1 (A), and study 2 (B, C). Error bars represent standard error. *p < 0.05.

The Likert scale

With respect to the responses collected from study 2, as shown in Figure 15, the wide range of options provided, based on the Likert scale, seemed not to result in a wider range of responses, with responses densely clustered around the same value for each controller.

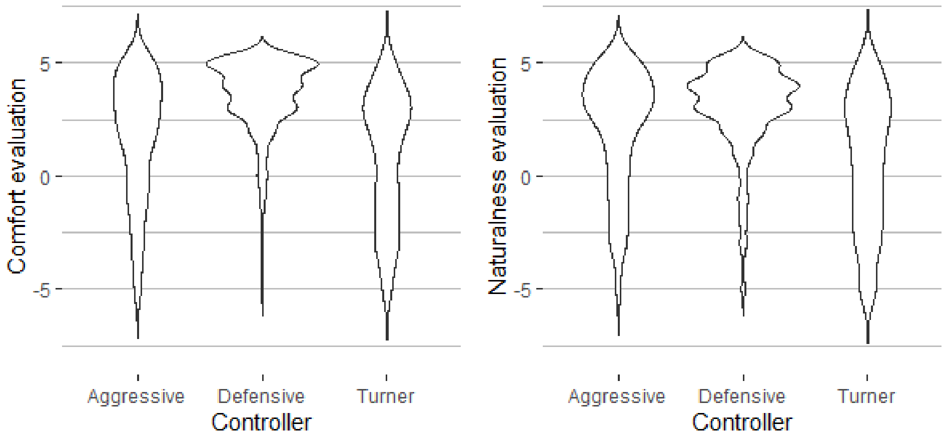

Figure 15. Comfort (left) and naturalness (right) ratings for the three controllers in study 2. The width of the violin plot represents the frequency of a value occurring in the data set. The y-axis shows each level out of the 11-point Likert scale.

Discussion and Conclusion

The results showed that, for both methods, regardless of the number of response options, a similar trend was found, with participants providing more positive, than negative, evaluation for all controllers, in both studies. This finding is in line with that of [13]. Compared to the wide range of response options provided by the Likert scale, a binary method only provides two options for participants to express their evaluation. However, in this study, the results from our binary evaluation technique was more similar to that of the Likert scale, and able to identify some small differences between the three controllers. This similarity between the binary and Likert scales in our study might be due to the number of times that participants were allowed to provide a response, and that this response was allowed during the drive. For example, as a comparison, [17] administered a questionnaire to measure the perceived usability of a website, which was evaluated using a Likert-type scale with a range of response options (3, 5, 7 and 11). Evaluation was provided once, after participants had finished interacting with the website. These authors found that their 3-point scale lacked reliability, compared to those allowing more response options, and suggest a wider range of scales to be used (i.e. minimum of 5). They found no difference between the 5, 7 and 11 response options. Put together, these results suggest that for such subjective evaluations, enabling repeated responses by a binary technique, in real time, might enhance its capabilities for identifying subtle differences between different measures, in a manner similar to that of a Likert scale with more response options. Regarding the Likert scale, our results showed that, for the positive responses, the mode of responses used were the numbers 3 and 4. This is in line with results from many other such studies (e.g., [20]), which show that responses using the Likert-type scale typically cluster at the lower or the upper end [21].

In conclusion, both the handset-based (with binary options) and oral report (with 11 options) methods used in these studies were found to be useful for evaluation of AV controllers during an automated drive. The Likert scale was found to be better than the binary method, in terms of providing more response options. However, if the users of the binary technique are allowed to present their evaluation as often as possible, in real time, the binary method is also found to be useful in this context. To provide more knowledge, future studies may also compare the use of both techniques for measuring exactly the same concept, which was not done in this study.

Acknowledgement

The first author is funded by the European Union Horizon 2020 funded SHAPE-IT project (Grant number: 860410). These two studies mentioned in the present work were part of the HumanDrive project, funded by Innovate UK and the Centre for Connected and Automated Vehicles (Project number TS/P012035/1).

References

[1] Beggiato, M., Hartwich, F., & Krems, J. (2018). Using Smartbands, Pupillometry and Body Motion to Detect Discomfort in Automated Driving. *Frontiers in Human Neuroscience*, *12*(September), 1–12. https://doi.org/10.3389/fnhum.2018.00338

[2] Elbanhawi, M., Simic, M., & Jazar, R. (2015). In the Passenger Seat: Investigating Ride Comfort Measures in Autonomous Cars. *IEEE Intelligent Transportation Systems Magazine*, *7*(3), 4–17. https://doi.org/10.1109/MITS.2015.2405571

[3] da Silva, M. C. G. (2002). Measurements of comfort in vehicles. *Measurement Science and Technology*, *13*(6). https://doi.org/10.1088/0957-0233/13/6/201

[4] Oborne, D. J. (1978). Passenger comfort - an overview. *Applied Ergonomics*, *9*(3), 131–136. https://doi.org/10.1016/0003-6870(78)90002-9

[5] SAE International. (2016). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE International. https://doi.org/https://doi.org/10.4271/J3016_201806

[6] Bellem, H., Klüver, M., Schrauf, M., Schöner, H. P., Hecht, H., & Krems, J. F. (2017). Can We Study Autonomous Driving Comfort in Moving-Base Driving Simulators? A Validation Study. *Human Factors*, *59*(3), 442–456. https://doi.org/10.1177/0018720816682647

[7] Sivak, Micheal, Schoettle, B. (2015). *Motion sickness in self-driving vehicles* (Issue April). https://deepblue.lib.umich.edu/handle/2027.42/111747

[8] Hajiseyedjavadi, F., Romano, R., Paschalidis, E., Wei, C., Solernou, A., Jamson, A. H., Boer, E. R., & Merat, N. (2021). *Effect of Environmental Factors and Individual Differences on Subjective Experience of Human-Like and Conventional Automated Vehicle Controllers [Manuscript submitted for publication]*. https://doi.org/10.13140/RG.2.2.13778.68808

[9] Peng, C., Merat, N., Romano, R., Hajiseyedjavadi, F., Paschalidis, E., Wei, C., Radhakrishnan, V., Solernou, A., Forster, D., & Boer, E. (2021). *Drivers' Evaluation of Different Automated Driving Styles: Is It both Comfortable and Natural? [Manuscript submitted for publication]*.

[10] Summala, H. (2007). Towards Understanding Motivational and Emotional Factors in Driver Behaviour: Comfort Through Satisficin. In *Modelling Driver Behaviour in Automotive Environments: Critical Issues in Driver Interactions with Intelligent Transport Systems* (pp. 189–207). https://doi.org/10.1007/978-1-84628-618-6

[11] Carsten, O., & Martens, M. H. (2018). How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology and Work*, *21*(1), 3–20. https://doi.org/10.1007/s10111-018-0484-0

[12] Wei, C., Romano, R., Merat, N., Wang, Y., Hu, C., Taghavifar, H., Hajiseyedjavadi, F., & Boer, E. R. (2019). Risk-based autonomous vehicle motion control with considering human driver's behaviour. *Transportation Research Part C: Emerging Technologies*, *107*(August), 1–14. https://doi.org/10.1016/j.trc.2019.08.003

[13] Hartwich, F., Beggiato, M., & Krems, J. F. (2018). Driving comfort, enjoyment and acceptance of automated driving–effects of drivers' age and driving style familiarity. *Ergonomics*, *61*(8), 1017–1032. https://doi.org/10.1080/00140139.2018.1441448

[14] Yusof, N. M., Karjanto, J., Terken, J., Delbressine, F., Hassan, M. Z., & Rauterberg, M. (2016). The exploration of autonomous vehicle driving styles: Preferred longitudinal, lateral, and vertical accelerations. *AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Proceedings*, 245–252. https://doi.org/10.1145/3003715.3005455

[15] Lee, J. D., Liu, S. Y., Domeyer, J., & DinparastDjadid, A. (2019). Assessing Drivers' Trust of Automated Vehicle Driving Styles With a Two-Part Mixed Model of Intervention Tendency and Magnitude. *Human Factors*. https://doi.org/10.1177/0018720819880363

[16] Schwanitz, S., Wittkowski, M., Rolny, V., Samel, C., & Basner, M. (2013). Continuous assessments of pressure comfort on a train - A field-laboratory comparison. *Applied Ergonomics*, *44*(1), 11–17. https://doi.org/10.1016/j.apergo.2012.04.004

[17] Lewis, J. R. (2021). Measuring User Experience With 3, 5, 7, or 11 Points: Does It Matter? *Human Factors*, *63*(6), 999–1011. https://doi.org/10.1177/0018720819881312

[18] Basu, C., Yang, Q., Hungerman, D., Singhal, M., & Dragan, A. D. (2017). Do You Want Your Autonomous Car to Drive Like You? *ACM/IEEE International Conference on Human-Robot Interaction, Part F1271*, 417–425. https://doi.org/10.1145/2909824.3020250

[19] Rossner, P., & Bullinger, A. C. (2020). How Do You Want to be Driven? Investigation of Different Highly-Automated Driving Styles on a Highway Scenario. *Advances in Intelligent Systems and Computing*, *964*, 36–43. https://doi.org/10.1007/978-3-030-20503-4_4

[20] Bachman, J. G., & O'Malley, P. M. (1984). Yea-Saying , Nay-Saying , and Going to Extremes : Black-White Differences in Response Styles. *The Public Opinion Quarterly*, *48*(2), 491–509.

[21] Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A bayesian hierarchical approach. *Journal of the American Statistical Association*, *96*(453), 20–31. https://doi.org/10.1198/016214501750332668