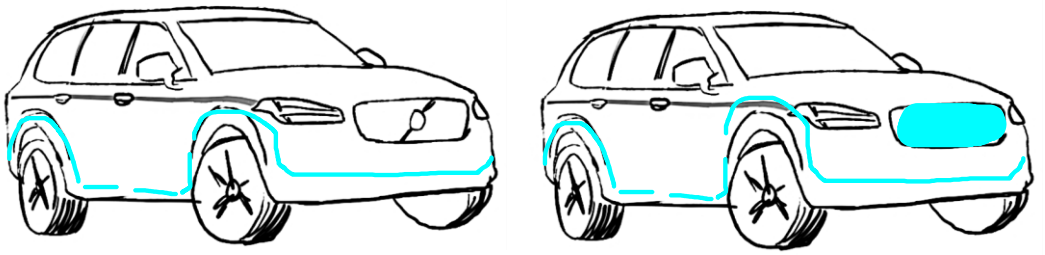
Scale-up:

Crowdsourcing for scaling up evaluation of external interfaces on automated vehicles

Public Report



Author: Azra Habibovic, Daban Rizgary, Mikael Ljung Aust
Date: 2021-09-29
Project within FFI Trafiksäkerhet och automatiserade fordon

FFI Fordonsstrategisk Forskning och Innovation

Contents

1 Sammanfattning

Externa gränssnitt förväntas bidra till säkrare interaktioner mellan automatiserade fordon (AV) och andra trafikanter. Dessa gränssnitt är dock i ett tidigt designstadium och utvärderas i bästa fall med endast ett begränsat antal deltagare. Detta gör att generaliserbarheten av dessa gränssnitt kan ifrågasättas. Samtidigt har online crowdsourcing som erbjuder en större och mer diversifierad deltagarpool inte utforskats tillräckligt. Syftet med denna förstudie är att utforska potentialen av online crowdsourcing för utvärdering av externa gränssnitt i ett tidigt designstadium, för att bidra till välgenomtänkta designbeslut med större potential att kunna tillämpas i större skala. För att skapa en grund för utformningen av externa gränssnittskoncept utvecklade projektet en enhetlig taxonomi och genomförde en systematisk jämförelse av de externa gränssnitten med hjälp av denna. Projektet designade också flera olika gränssnittskoncept och utformade och genomförde två studier via online crowdsourcing MTurk. För att utforska dess validitet replikerades den andra MTurk-studien i en laboratoriemiljö där deltagarna valdes av forskarna. Projektet har utökat kunskap om externa gränssnitt för automatiserade fordon samt skapat en förståelse för hur man utformar och validerar crowdsourcing studier. En övergripande, men preliminär slutsats, är att online crowdsourcing MTurk genererar liknande resultat som motsvarande studie i lab, givet att studien innehåller enkla frågor som inte kräver utförliga svar. Detta är naturligtvis en begränsning och gör online crowdsourcing endast lämpligt för vissa typer av utvärderingar.

2 Abstract

External interfaces are expected to facilitate safer interactions between automated vehicles (AVs) and other road users. These interfaces are in an early design stage and are, at the best, evaluated using only a limited number of participants. This leaves the question of their generalizability open. At the same time, online crowdsourcing that offer a larger and more diverse participant pool is underexplored. The aim of this pre-study is to explore the potential of online crowdsourcing for evaluation of external AV interfaces in an early design stage, to drive better informed design decisions with a greater potential to be applicable on a wider scale. To provide a base for the design of external interface concepts, the project developed a unified taxonomy and conducted a systematic comparison of the external interfaces across various design parameters utilizing the taxonomy. The project also designed several interface concepts and conducted two studies using online crowdsourcing MTurk. To explore its validity, the second MTurk study was replicated in a lab environment where the participants were selected by the experimenters. The project has extended knowledge on external interfaces for automated vehicles as well as created an understanding of how to design crowdsourcing surveys and assess validity of such survey. An overall, but preliminary conclusion, is that the online crowdsourcing MTurk generates similar results as the corresponding lab survey given that the survey does not involve questions that require elaborations. This is of course a limitation, and makes online crowdsourcing suitable only for certain types of evaluation.

3 Background

3.1 The need for trust, acceptance, and safety of automated vehicles

By replacing human drivers, in some or in all driving situations, automated vehicles (AVs) are expected to eliminate issues related to human drivers. Large-scale introduction of such vehicles is thus anticipated to bring many benefits to the society, including improved safety, reduced congestion, lower emissions, higher productivity, and greater access to mobility. However, to reach these benefits, AVs will need to be trusted and to gain societal acceptance. While trust and acceptance could be affected by a range of factors [1], one thing is sure: the ability of AVs to safely and smoothly interact with other road users in their vicinity will play a key role. That is, future AVs may face issues related to interaction with drivers of conventional vehicles as well as with bicyclists and pedestrians. These

interactions have until recently been largely unexplored as the focus in the research community has primarily been on tackling challenges associated with the interactions inside an AV. Under the last 4-5 years, however, a few studies have been conducted on external interactions, indicating a need for additional communication support to warrant safety and acceptance of AVs.

3.2 New interaction patterns between automated vehicles and other road users

When encountering a vehicle today, road users use both vehicle-centric cues (e.g., velocity, deceleration) and driver-centric cues (e.g., eye contact, gesture) to interpret the situation. With the transfer of control from the human driver to the vehicle, the driver-centric cues will no longer be available in the same way. Current research on interactions between AVs and other road users points in two directions: one advocating that motion patterns of AVs are sufficient to communicate the intent of AVs [2, 3], and the other one suggesting that interactions will be affected by the lack of explicit communication with drivers and that additional communication features may be needed [4-7]. Given these contradictory findings, it is important to investigate the role of such features, and the question becomes: what, when and how should the vehicle communicate to other road users to ensure safe interactions?

3.3 Short review of some proposed external AV vehicle interfaces to other road users

In 2012, a group of researchers at MIT suggested a biomimetic interface for AVs. Two years later, RISE and partners, including Volvo Cars, created an interface called AVIP that consists of an outward-facing LED light strip that uses distinct patterns of light to inform surrounding road users about the state of the AV (on/off) and what the AV is about to do. It was guided by the idea that AVs should communicate their intent rather than explicitly inviting people to act. While this idea and our simplistic design have been adopted by some stakeholders (e.g., Ford), others have suggested other solutions: projection of a zebra crossing (Mercedes), a light strip around the vehicle and textual messages in the front windshield such as “after you” (Nissan), vehicle motion direction projected onto the road (Mitsubishi), a robotic hand conveying various signals (Google). In 2018, Volvo Cars showed a concept that signalizes vehicle intent using various modalities around the entire vehicle.

3.4 The challenge of scalability in early stage design assessment

Some of the interfaces above have been evaluated in simple scenarios with a limited number of participants (up to 30). This leaves the question of their generalizability open, and may explain the contradicting research of the need for external communication. Since AVs are to be deployed worldwide, extensive real-world testing would be ideal before an interface is put into general use, due to safety critical implications of miscommunication. However, since such testing is expensive, efficient methods for early stage concept assessment are highly desirable, to test a large number of designs and limit the number of designs for further testing.

In this area, only a limited number of choices is available. One approach is Wizard of Oz (WoZ) based testing, e.g., of how pedestrians might interact with automated vehicles given various external design concepts [6, 8, 2]. This however still requires a significant effort and involve a limited number of participants. In summary, since AVs are not deployed on a larger scale yet, it is challenging to design evaluation experiments that reveal results which are representative of a larger population.

3.5 Possible solution: Online crowdsourcing using Amazon Mechanical Turk (MTurk)

One major premise in this pre-study is that online crowdsourcing services can be used to address the scalability issue described above. Numerous examples of such services exist, some of which focus on simple tasks, functioning as a micro-task marketplace. Of these, Amazon Mechanical Turk (MTurk) is both most well-known and largest in scale. MTurk provides access to a large potential participant pool at modest cost per participant. It has been reported that MTurk has good performance,

especially on social sciences research, since participants are diverse and more representative of a non-college population than traditional samples [9].

The proposed approach in this pilot is to test if MTurk can be applied in a cost-effective manner to identify design concepts appropriate for more detailed development, using a larger and more varied test participant sample than other existing methods. Furthermore, a large number of elements or minor variations can be tested through a series of MTurk runs in a matter of hours, designers should be able to refine concepts more quickly than available resources typically allow. Also, factors that are difficult to *explore during design (e.g., culture, demographic, prior mental models) can be factored in early in the process.*

Of course, using online crowdsourcing is not without challenge. Studies indicate that MTurk ca be a trustworthy data source (e.g., [10]), but a key question that remains for this particular field is whether MTurk results are comparable to studies exploring the same research question in a traditional controlled experiment (lab, test track, or real-world). While Fridman et al. [11] and Li et al. [12] explored interpretation of external vehicle interfaces using Mturk, they did not look into validity compared to a controlled study. This pre-study made a first attempt at addressing this issue.

4 Aim, research questions and method

The aim of this pre-study is to explore the potential of online crowdsourcing methodologies for evaluation of external AV interfaces (eHMI) in an early design stage, to drive better informed design decisions with a greater potential to be applicable on a wider scale. Specifically, the project explored the extent to which the online crowdsourcing platform MTurk is appropriate for conducting studies on external vehicle interfaces, and how well it replicates the corresponding evaluations in a lab experiment where participants are selected by the experimenters. That is, our objective was to develop and apply a methodology in an online MTurk study as well as in a controlled study, to allow comparison between these two approaches.

The overall research question is: *To what extent is the online crowdsourcing platform MTurk suitable for evaluation of external AV interfaces (eHMI) as compared to controlled studies?* The specific research questions (RQ) include:

1. Which design elements of an external AV interface can be evaluated via MTurk?
2. How do we design a study that will work both in MTurk and in a controlled experiment?
3. Which are the best qualitative and quantitative metrics for the studies?
4. Which participant selection criteria is suitable?
5. How to filter out inadequate data generated via MTurk?
6. What are similarities and differences between MTurk results and a controlled experiment?

To answer these questions, the pre-study has utilized the following methods:

- Literature review
- Co-creation workshops for interface design
- Visualization of interfaces in terms of 3D videos
- Workshops for defining survey, incl. metrics
- Online survey and survey in lab

5 Goal

The goal of the pre-study has been to:

- Increase knowledge on evaluation methodologies and design of external AV interfaces.
- Strengthen collaboration between project partners
- Strengthen international collaboration and Swedish competitiveness.

6 Results and goal fulfillment

6.1 Overall results and deliverables

The objective of the pre-study project Scale-up has been reached and it has generated the following results in accordance to the original project plan:

- Knowledge on online large-scale crowdsourcing as an evaluation tool.
- Knowledge on validity of crowdsourced data in the context of external AV interfaces.
- Knowledge on which design features are suitable for communication between automated vehicles and other road users.
- A list of questions and ideas to be further explored in a larger project.
- One bachelor thesis (instead of a master thesis) involving two students.
- Published two scientific papers, started on a third paper.

More specifically, the project has explored validity of online crowdsourcing for evaluation of external AV interfaces (eHMI) in early design phases. First, the project developed a unified taxonomy that allows a systematic comparison of the eHMI across 18 dimensions, and generated a state of the art overview of eHMI design using the taxonomy (Paper 1). This supported design of the eHMI concepts that were designed in co-creation workshops and then illustrated in animated videos. These videos served as basis for the evaluation of the eHMI that was done online using crowdsourcing Mturk (Online Study 1 and Online Study 2), as well as in lab where Online Study 2 was replicated to explore validity of online crowdsourcing (Validity Study). Due to Covid-19, user studies in lab have been delayed and that data are still to be analyzed. This work has strengthen collaboration between the project partners, as well as collaborations on the international arena (e.g., we have published papers together with researchers from the Netherlands and Germany and contributed to standardization work led by ISO), which is in line with the overarching objectives of FFI regarding the Swedish competitiveness and international connected research environments, and cooperation between industry, academia (via thesis work) and institute.

Deliverables in the project are presented in Table 1. Some of them are detailed in the following chapters.

Table 1 Overview of deliverables as specified in the project plan and their current status.

Work package (WP)	Lead (partner)	Deliverable	Status
WP1: Project lead, coordination and dissemination	RISE (VCC)	**D1.1:** Final project report	**Completed.** • The project has been presented at several national and international events, incl. FFI TSAF conference. • Communication and reporting to FFI done as requested. • A final project report compiled, and handed in.
WP2: Design of interface concepts	VCC (RISE)	**D2.1:** A chapter in the final report describing the interface concepts and ideas behind them **D2.2:** Photo/video material capturing the concepts in selected contexts	**Completed.** • Defined and selected relevant scenarios. • Developed taxonomy for a systematic comparison of the eHMI design features in collaboration with researchers from the Netherlands and Germany (Paper 1)

			• Conducted a series of co-creative design workshops. • Generated several eHMI concepts (Paper 2 and Paper 3) • Several eHMI concepts illustrated in form of photos/video.
WP3: Development and execution of MTurk study (online crowdsourcing study)	RISE (VCC)	**D3.1:** A chapter in final report describing evaluation framework, incl. experiment design. **D3.2:** Specially developed HIT **D3.3:** Data generated by completing HIT via MTurk	**Completed.** • Two online crowdsourcing studies (Human Intelligence Tasks, HITs) have been designed and conducted using MTurk. The first one was done in collaboration with Eindhoven University (Paper 2).
WP4: Development and execution of a validation study (traditional controlled experiment)	RISE (VCC)	**D4.1:** A chapter in final report describing the experiment, incl. selection criteria and recruitment of participants **D4.2:** Data generated in the experiment	**Completed.** • A Validation Study in lab (using same material as in the second online study) was designed and conducted.
WP5. Data analysis and assessment of crowdsourcing as a tool for future studies	RISE (VCC)	**D5.1:** A chapter in final report describing methods for ensuring data quality as well as quantitative/qualitative methods used to analyze data from MTurk and validity study. **D5.2:** A chapter in final report describing a) insights on validity of MTurk as evaluation platform and b) insights on suitability of various design features	**Partly completed.** • Taxonomy and state-of-the art analysis completed and published in a journal paper (Paper 1). • Data and insights generated in the first online crowdsourcing study were published in a conference paper (Paper 2). • Data and insights generated in the second online crowdsourcing study and in the validation study are under analysis (been delayed due to Covid-19). These are expected to be published in a paper (Paper 3). In this report, we present initial insights on MTurk as a validation platform.

6.2 Interface concepts (D2.1)

In order to understand which design parameters are commonly used for design of eHMI, the project (in collaboration with researchers from Germany and the Netherlands) developed a unified taxonomy that allows a systematic comparison of the eHMI across 18 dimensions, covering their physical characteristics and communication aspects from the perspective of human factors and human-machine interaction. This taxonomy was then applied to analyze and code 70 eHMI concepts published in scientific papers and various media to portray the state of the art and highlight the relative maturity of different contributions. The results helped us designing the eHMI concepts that were evaluated using online crowdsourcing and lab experiment.

In **Online Study 1** (conducted together with the researchers from the Netherlands), we explored two design features of a light-band eHMI that communicates yielding intent to pedestrians: color (red,

green, cyan) and animation (flashing, pulsing, wiping inwards, wiping outwards, and wiping alternatively inwards as well as outwards). These concepts are exemplified in Figure 1 .

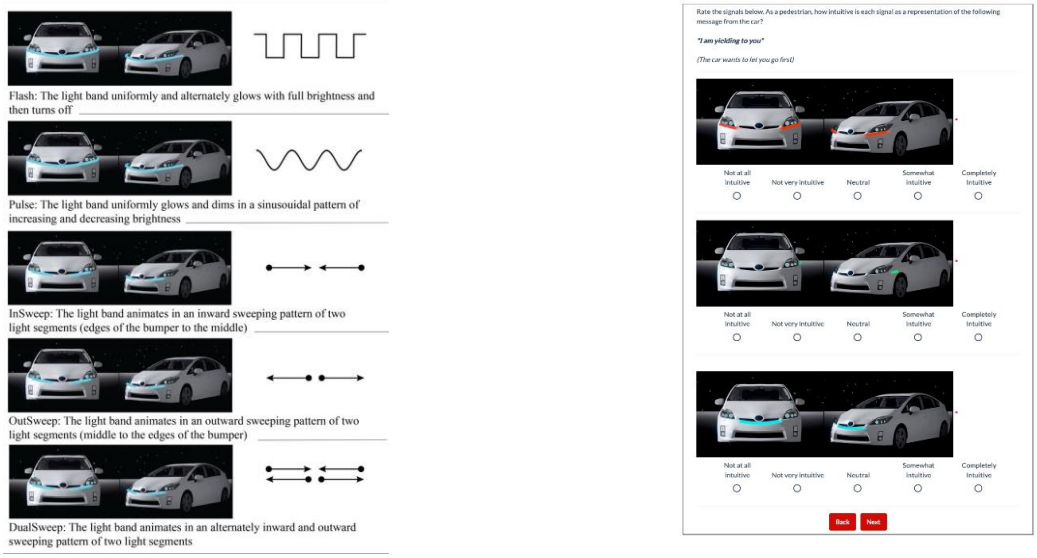


Figure 1 Animation patterns of eHMI concepts (left) and colors (right) designed for Online Study 1. Right hand figure shows also the outline of the survey.

In **Online Study 2**, we explored only one color (cyan) for two light-band eHMI concepts that communicate three messages: “I am in autonomous mode”, “I will yield” , and “I will take off”. We designed 8 different concepts, however, only two of them (Concept 1 and Concept 2) were selected for evaluation. The concepts displayed different animations and placements. The animation and placement of Concept 1 was inspired by the Volvo Cars 360-concept.The animation in Concept 2 was inspired by AVIP that was previously presented by RISE, Volvo Cars and other partners. However, the LED light strip in Concept 2 was partly placed in grill and partly in the lower part of the vehicle. The study is yet to be published in a scientific paper, and thus we omit details on these concepts in this report. However, we show a concept that was used in the training session, see Figure 2.

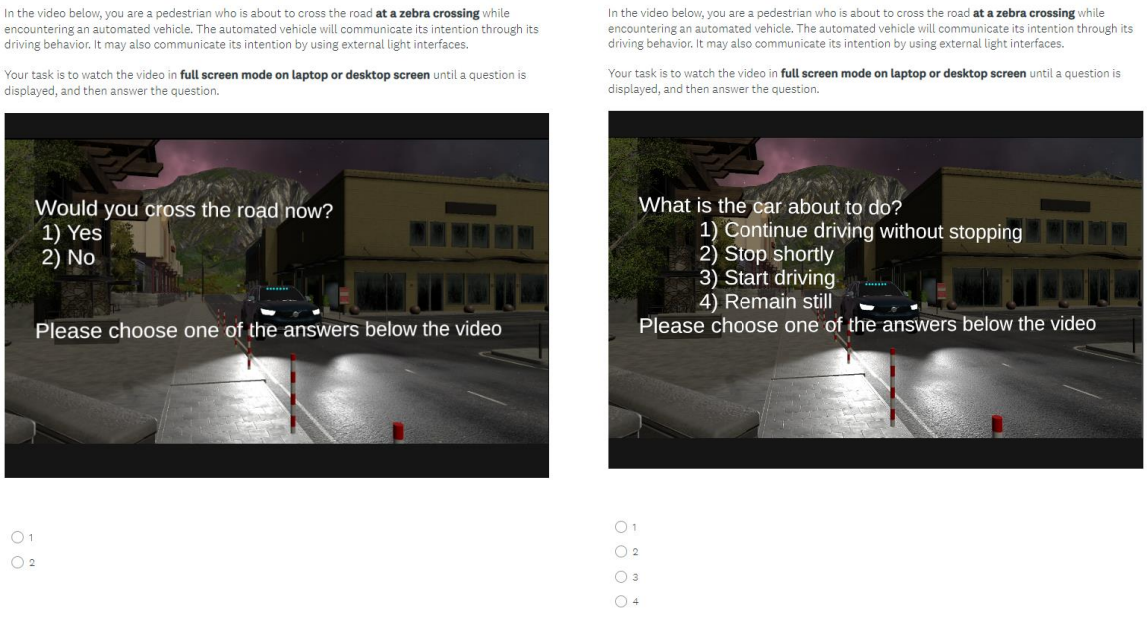


Figure 2 eHMI concept that was used in training session in Online Study 1 and Validation study. The figure shows also the outline of the survey and the two questions that the participants were asked to answer.

6.3 Crowdsourcing studies (D3.1)

Online Study 1 was designed as specified in Table 1, for details see Paper 2.

Table 2 Design of Online Study 1

Research question	Which color and/or animation pattern on a light-band eHMI is the most favorable according to users for a yielding message from an AV towards a pedestrian?
Experiment design	The experiment was conducted in a mixed design with three independent variables (two within-subjects, and one between-subjects), and one dependent variable. The two within-subjects independent variables were: Color with the three levels(green, red, and cyan) and Animation with five different levels (Flash, Pulse, InSweep, OutSweep, DualSweep). The between-subjects independent variable was Message (Intention communication, Instruction). The dependent variable was the user's score of the intuitiveness of a shown eHMI for a yielding message for a car on a Likert scale of 1 to 5, as well as their subjective opinions on what drives this intuitiveness. See Figure 1.
Participants	The experiment sample consisted of 400 participants with 200 in each condition of the between-subjects factor (Intention communication vs. Instruction).
Recruitment of participants	We recruited the participants using Amazon Mechanical Turk (MTurk). In order to ensure the quality of the responses, only MTurk Master Workers were recruited. In order to draw meaningful conclusions and ensure homogeneity and ecological validity of the responses, we recruited participants only from the US.
Validation	We conducted a manipulation check at the end of the survey
Data pre-processing	The data was scrutinized for any obvious discrepancies or inconsistencies. This resulted in omitting responses from 17 participants from the Intention communication condition, and 8 participants from the Instruction condition.
Survey tool	We presented the experiment as a questionnaire on the online survey platform SurveyGizmo and then deployed it through MTurk.
Animation	The eHMI was displayed on a passenger car that was standing still. There was no context (i.e. the background was black). See Figure 1.
Survey outline	Each page of the questionnaire was dedicated to either one color or one animation pattern. Thus the entire questionnaire consisted of 8 pages – 3 pages for the different colors, and 5 pages for the different animations. Foreach of the 3 colors, we showed the 5 animation patterns next to each other on the same page. Conversely, for each of the 5animation patterns, we showed the 3 colors next to each other on the same page. This led each color/animation combination of an eHMI to be graded twice: once from the perspective of its color, and once from the perspective of its animation pattern.
Execution	A pilot study was conducted with 5 people before deployment .showed that the survey took between 8 and 12 minutes to complete.
Time	The survey took between 8 and 12 minutes to conduct.
Incentive	We compensated each worker with 2.00 USD for their participation.
Practicalities	Since we conducted the experiment as a between subjects study, we deployed the Intention communication condition of the study first and allowed it to run until the necessary number of responses (200) was reached. After receiving these 200 responses, we deployed the second condition (Instruction) by excluding any Worker who had already participated in the previous condition.

Online Study 2 was designed as specified in Table 3 (details will be presented in Paper 3 when it is published).

Table 3 Design of Online Study 2

Research question	How is the presence of eHMI affecting a pedestrian's a) situation understanding and b) willingness to cross?
Experiment design	The experiment was designed as an within subject study.
Participants	The experiment sample consisted of 231 participants.
Recruitment of participants	We recruited the participants using Amazon Mechanical Turk (MTurk). In order to ensure the quality of the responses, we recruited either MTurk Master workers, or workers with a HIT approval right above 98%. In order to draw meaningful conclusions and ensure homogeneity and ecological validity of the responses, we recruited participants only from the US and West Europe.
Manipulation check	We conducted a manipulation check by embedding two manipulation check questions in the survey.
Data pre-processing	The data was scrutinized for any obvious discrepancies or inconsistencies. This resulted in omitting responses from 31 participants.
Survey tool	We presented the experiment as a questionnaire on the online survey platform SurveyMonkey and then deployed it through MTurk.
Animation	A 3D animated video was created representing a zebra crossing in an urban environment. The vehicle was either approaching the zebra crossing, or standing still in front of the zebra crossing, when the video started played. At the end of each video, a question was displayed in the video along with the potential answers. To answer the question, the participant needed to select appropriate option displayed under the video, see Figure 2.
Survey outline	The survey was divided into three parts: 1) practice, 2) core questionnaire, 3) experience and background questionnaire. In the first part, each participant experienced two videos showing an eHMI concept (not included in core questionnaire) and was asked to answer the two questions that are asked in the core questionnaire. The core questionnaire included videos of a) three communication concepts (AVIP in Grill, 360, no additional interface), b) two vehicle behaviors (approaching, standstill), c) two scenarios when in standstill (alone, with another pedestrian), d) two questions (would you cross the road now?, what is the car about to do?). Each combination was experienced twice. In addition, two pages with manipulation check questions were shown to each participant. This resulted in 40 pages (or exposures) in the core questionnaire. All pages in the core questionnaire were shown in a random order except for manipulation check pages.
Execution	A pilot study was conducted with 4 people before deployment
Time	The survey took between 30 and 40 minutes to conduct.
Incentive	We compensated each worker with 4 USD for their participation.
Practicalities	All combinations that we wanted to test made the survey long and we needed to focus the evaluation to only two (of totally 8 designed concepts).

6.4 Validation Study (D4.1)

The validation study was conducted at RISE premises at Lindholmen, Gothenburg and took about 1h to complete for each test participant. While the survey that the participants completed was the same as the one in Online Study 2, there were a few differences in the procedure:

- The validation study involved only 16 participants as compared to 231 participants in the online study.
- In validation study, participants were recruited from general public via social media and researchers did not have any track record for these participants when it comes to their experience and ability answering a survey, which was possible to see in MTurk.
- In the validation study, the participants met the test leader (always same test leader), while it was not the case in the online study.
- In the validation study, all participants used same screen and room, while this was not the case for the online study.
- The validation study ended with a debriefing (ca 10 minutes) where the test leader asked the participant to comment on the experience, eHMI concepts and the survey itself. Although the participants in the online study could leave comments at the end of the survey, it was optional and only a few of them did so.
- The participants were compensated in both studies. However, the absolute value of the compensation was different; in the validation study it was 300 SEK and in the online study it was 4 USD.

6.5 Data pre-processing and analysis (D5.1)

Both Online Study 1 and 2 as well as the Validation Study, included questions that helped us to some degree verify the answers in the survey. In the Online Study 1, the participants were subjected to a manipulation check to ensure basic understanding of the task and context, as well as color perception. The questions were displayed at the end of the survey. We verified whether the participants understood two elements: (1) that the responses were to be given from the point of view of a pedestrian (as opposed to cyclists or other drivers), and (2) that the message they were evaluating the eHMI towards was that the car was yielding/letting them cross first (as opposed to the car cruising in automated mode, or starting to drive from rest). In order to control for participants' perception of color without having to ask for medical information, we added another manipulation check where we asked the participants to report the colors of the eHMI they observed in the study.

In Online Study 2, information about color perception was embedded into the background questionnaire. Also, the Online Study 2 embedded two manipulation check questions in the survey itself: 1) asked if the participant hold a driver license and if they gave a different answer as compared to background questionnaire their data was excluded from further analysis, and b) asking a question with two answer options in video while displaying four answer boxes under the video and if a person selected a box that was not among the answer options in the video his/her data was excluded from further analysis. These questions were displayed in the same manner as all other questions (but they were not presented in a random order).

One should, however, note that manipulation check questions do not warrant that the survey was, for instance, answered with full attention or with great honesty. The fact that the participants got paid for participating amplifies the risk for such issues. These risks are applicable to the lab study as well, and in general hard to completely eliminate.

6.6 Validity of MTurk (D5.2)

The comparison of the results from Online Study 2 and Validation Study is yet to be done. Our preliminary conclusion is, however, that the online crowdsourcing MTurk has a great potential to replicate results from a controlled lab survey where the participants are selected by the experimenters given that the survey does not involve questions that require elaborations. This is of course a limitation, and makes the online crowdsourcing suitable only for certain types of evaluation. This is also echoed by the fact that about 25% of the participants in Online Study 2 stated that they would prefer to answer this survey in a research facility with the researchers present.

MTurk has a few options for allowing researchers to ensure a increased quality of data from its workers. One of the functions is that a researcher can use is allow only Master Workers to see and work on their surveys. Another function is that one can disallow workers below a certain percentage of previously approved tasks to complete ones survey. In other words, one can allow for only having workers that have a good track record from previous instances of completing work in MTurk. This is a bit more difficult to control for if one is recruiting participants from general public via social media (as we did in our Validation Study). Having study participants that take the research seriously might, however, only be one of multiple influencing factors that differ between crowdsourcing settings and lab settings.

7 Dissemination and publication

7.1 Dissemination of knowledge and results

Hur har/planeras projektresultatet att användas och spridas?	Markera med X	Kommentar
Öka kunskapen inom området	x	Project members have been involved in several events and workshops where validity of online surveys has been discussed with other professionals. Project members plan to publish a scientific paper based on Scale-up results.
Föras vidare till andra avancerade tekniska utvecklingsprojekt	x	RISE has an ongoing research project with Scania and Halmstad University (FFI eHMI) where knowledge on evaluation methodology and eHMI design gained in Scale-up are utilized. RISE has an ongoing project with Aptiv, Clean Motion, Combitech and Halmstad University (Trv GLAD) where knowledge on evaluation methodology and eHMI design gained in Scale-up are utilized. RISE has an institutional PhD candidate who has been involved in the Scale-up project and knowledge gained here will be utilized in his further studies. VCC has several internal activities in the field of eHMI where Scale-up results will be useful
Föras vidare till produktutvecklingsprojekt		
Introduceras på marknaden		
Användas i utredningar/regelverk/ tillståndsärenden/ politiska beslut	x	RISE is participating in ISO standardization activities on external HMI for automated vehicles RISE have participated in activities on international regulation of external HMI by UNECE

7.2 Publications

The per-study has resulted in the following publications:

- (Paper 1) Debargha Dey, **Azra Habibovic**, Andreas Löcken, Philipp Wintersberger, Bastian Pfleging, Andreas Riener, Marieke Martens, Jacques Terken, Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces,Transportation Research Interdisciplinary Perspectives, Volume 7, 2020, https://doi.org/10.1016/j.trip.2020.100174.

- (Paper 2) Dey, D., **Habibovic, A**., Pfleging, B., Martens, M., & Terken, J. (2020b). Color and animation preferences for a light band eHMI in interactions between automated vehicles and pedestrians. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, HI. doi:10.1145/3313831.3376325.

- (Paper 3): Validating online crowdsourcing in the context of external HMI for automated vehicles. This paper is in progress and is to be submitted.

- A. Karlsson & T. Hedlund., Development of an Intuitive Pedestrian Interaction System for Automated Vehicles. 2019. Bachelors Thesis.

- Poster. Presented at FFI TRAF conference 2019.

- OmAD. The project has been mentioned/described in the newsletter OmAD at several occasions (Link, Link, Link)

- SAFER. The project was associated to SAFER and presented to Road User Behavior group. The information about the project is displayed on SAFER website, in SAFER news and in SAFER annual report.

- Oral presentations at conferences such as Automated Vehicle Symposium 2020, Vision Zero Academy as well as presentations for vehicle industry in Sweden.

8 Conclusions and future research

Our preliminary conclusion based on the studies conducted in the pre-study project Scale-up is that online crowdsourcing might be a suitable evaluation method for external AV interfaces in early design phases given that the questions that are asked in survey do not require elaboration. Our validation study was, however, rather limited and we urge for exploring the validity of online crowdsourcing into more detail. Brief summary of our answers to the specific research questions:

1. *Which design elements of an external AV interface can be evaluated via MTurk?*
   From our studies, we can conclude that it is valuable to evaluate simple design features such as color and animation pattern. Evaluating more complex characteristics that often require elaboration or ability to track the actions of the participant may be less suitable. While we have not finalized the analysis yet, the impression is also that vehicle motion might be somewhat difficult to interpret from videos.
2. *How do we design a study that will work both in MTurk and in a controlled experiment?*
   It is important to simplify questions as much as possible. In the controlled experiment (e.g. in lab) the participants usually have possibility to ask test leader if something is unclear, but that is not the case with large scale online study. This indirectly dictates which design elements of an external AV interface that can be evaluated and what type of questions can be studied. The length of survey is another issue since a participants attention risks to be decreased for longer surveys. Also, we noticed that crowdsourcing surveys that are shorter (up to 10-15 minutes) are more attractive to the participants.
3. *Which are the best qualitative and quantitative metrics for the studies?*

In Online Study 1, the participants were asked to rate the intuitiveness of the communication concepts being displayed to them. While this quantitative metric proved to be useful in this context, we noticed also the importance of being able to motivate answers. Having just rating from a Likert scale is limiting when it comes to the interpretation of the answers. In Online Study 2 and in Validation Study, we used two quantitative metrics: ability to predict a vehicle's intention and willingness to cross. A conclusion is that these questions are complementing each other, and together they provide a deeper understanding for the effect that eHMI concept have on the pedestrian behaviour.

4. *Which participant selection criteria is suitable?*
   This depends on the aim of the study, and how many participants one wants to include in the study. While it might be easier to engage a larger number of participants in online studies as compared to lab studies, the number of potential participants in online studies decreases significantly when criteria such as Master Workers is applied. However, we found this particular criteria important to ensure the quality of the responses. Another criteria that we found important was type of screen being used by the participants. In our pilots, we discovered that viewing the interface on a smartphone limits the visibility of the interface, and it was natural to include only those participants using laptop or desktop screen. Also, depending on the research question, it might be needed to limit the geographical area. In our case, we choose to focus on the USA and West Europe to ensure that the participants are familiar with similar traffic conditions.
5. *How to filter out inadequate data generated via MTurk?*
   A crucial part in this is defining suitable manipulation check questions that enable one to "identify" discrepancies in the data. Such questions need to be thoughtfully selected and adapted to the topic being studied. One should, however, note that manipulation check questions do not warrant that the survey was, for instance, answered with full attention or with great honesty. The fact that the participants get paid for participating amplifies the risk for such issues. These risks are applicable to the lab study as well, and in general hard to completely eliminate.
6. *What are similarities and differences between MTurk results and a controlled experiment?*
   One major difference between MTurk and a controlled experiment is that the participants get opportunity to explain and elaborate their answers after a completed survey, which provides a better understanding of what worked well and what was challenging for them. Interestingly, the participants in both online and lab survey reported that they could stay attentive while completing the survey.

Ideas for future research include:

- Conduct a validation study in lab with a larger sample of participants.
- Conduct a study where lab, online crowdsourcing, and test track results are compared and contrasted.
- Multi-cultural crowdsourcing and validation.
- Compare and contrast different manipulation check questions in the context of external AV interfaces.

9 Participating organizations and contact persons

RISE Research Institutes of Sweden
Contact: Daban Rizgary,
daban.rizgary@ri.se

Volvo Cars
Contact: Mikael Ljung Aust,
mikael.ljung.aust@volvocars.com

10 References

[1] Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum. Factors 39(2):230–253

[2] Rothenbücher D, Li J,Sirkin D, Mok B,Ju W (2016). Ghost driver: a field study investigating the interaction between pedestrians and driverless vehicles. In: 25th IEEE International symposium on robot and human interactive communication.

[3] Dey D, Eggen B, Martens M, Terken J (2017) The impact of vehicle appearance and vehicle behavior on pedestrian interaction with autonomous vehicles. In: Adjunct proceedings of the 9th international ACM conference on automotive user interfaces and interactive vehicular applications (AutomotiveUI'17).

[4] Merat N, Madigan R, Nordhoff S (2016) Human factors, user requirements, and user acceptance of ride-sharing in automated vehicles.

[5] Böckle M-P, Pernestål Brenden A, Klingegård M, Habibovic A, Bout M (2017) SAV2P-exploring the impact of an interface for shared automated vehicles on pedestrians experience. AutomotiveUI.

[6] Lundgren Malmsten V, Habibovic A, Andersson J, Lagström T, Nilsson M, Sirkka A, Fagerlönn J, Fredriksson R, Edgren C, Krupenia S, Saluäär D (2017) Will there be new communication needs when introducing automated vehicles to the urban context? In: Advances in human aspects of transportation.

[7] Habibovic A, Andersson J, Malmsten Lundgren V, Klingegård M, Englund C, Larsson S (2018). Communicating intent of automated vehicles to pedestrians. Frontiers in Psy.

[8] Habibovic A, Andersson J, Nilsson M, Lundgren Malmsten V, Nilsson J (2016) Evaluating interactions with non-existing automated vehicles: three Wizard of Oz approaches. In: Intelligent vehicles symposium (IV), 2016 IEEE

[9] Buhrmester M, Kwang T, Gosling D S (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science 6, 1 (2011), 3–5.

[10] Rand DG (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. J Theor Biol, 299:172-9.

[11] Fridman L, Mehler B, Xia L, Yang Y, Facusse YL, Reimer B (2018). MIT. To Walk or Not to Walk: Crowdsourced Assessment of External Vehicle-to-Pedestrian Displays.

[12] Li Y, Dikmen M, Hussein GT, Wang Y, Burns C. To Cross or Not to Cross: Urgency-Based External Warning Displays on Autonomous Vehicles to Improve Pedestrian Crossing Safety. In: Adjunct proceedings of the 9th international ACM conference on automotive user interfaces and interactive vehicular applications (AutomotiveUI'18).