Ministry of Infrastructure
and Water Management

Opportunities and bariers for measuring traffic safety indicators based on vehicle sensor data

Master Thesis
2022

Steven Jansen
Transport, Infrastructure and Logistics

TU Delft

Opportunities and barriers for measuring traffic safety indicators based on vehicle sensor data

a Delphi study

By

Steven Jansen

in partial fulfilment of the requirements for the degree of

Master of Science
in Transport, Infrastructure & Logistics

at the Delft University of Technology,
to be defended publicly on 31-10-2022.

Thesis committee:	Prof. dr. B. van Arem (Chair)	TU Delft
	Dr. E. Papadimitriou	TU Delft
	Dr.ir. H. Farah,	TU Delft
	Dr. N. Schaap	Min. I&W
	LL.M. M. van Leeuwen	Min. I&W

Summary
Vehicle automation in the form of Advanced Driver Assistance Systems (ADAS) is seen as a promising way to increase traffic safety. Several types of ADAS will even become mandatory in all new vehicles in the European Union (EU) from 2024 onwards. However, the size of the effects of ADAS on traffic safety in practice is unknown. This has several causes, from methodological differences in measuring effects, to the lack of focus on the effects of combining multiple ADAS, to unknowns about how consumers use the systems in practice. The uncertainty about the effects of ADAS on traffic safety is a problem for both policymakers and consumers. For the former not just because it is the government's responsibility to approve and deal with these systems on the roads, but also because it is seen as a potential policy instrument to increase traffic safety. For the latter because (perceived) safety is an important factor in buying and using a vehicle.
At the same time, ADAS may provide an opportunity. Vehicles equipped with ADAS have several sensors onboard to facilitate these systems and are often connected with built-in SIM cards. The wealth of data collected by these systems may allow to measure traffic safety in a more pro-active way, by measuring traffic conflicts (interactions that are dangerous but do not necessarily result in an accident). This would solve several of the problems with existing ways of measuring traffic safety with reactive indicators (accidents and resulting fatalities and severe injuries). This research will focus on safety at network level as an aggregated statistic as this can help to assess the development of traffic safety over time, and with that the effects of future policy interventions. Measuring traffic safety with proactive indicators would allow for more and faster acquired knowledge than the current reactive indicators, which could ultimately improve traffic safety policies and traffic safety.

Therefore, the objective of this research is to explore if data collected by vehicles equipped with ADAS can be used for measuring traffic safety at network level. In order to reach this goal, a literature review is conducted followed by expert input in the form of a Delphi study. The research focusses on two aspects: potential types of indicators for measuring traffic safety and on the feasibility of using vehicle sensor data to measure these indicators.

Literature review
Traffic safety is a complex concept that can be measured in many ways and at different levels. The two main parties involved in measuring traffic safety are the government and (academic) researchers. The literature review discusses a variety of indicators used by these two parties to various ends. The Dutch government uses fatalities and severe injuries as a measure for traffic safety at network level and is in the process of implementing the Dutch Safety Performance Indicators (SPIs) which focus on risks involved in traffic. Various indicators to measure traffic conflicts have been developed and used in the academic community. Some of these are specifically focussed on the safety evaluation of vehicles equipped with automation. An overview of all of these types of indicators is presented in the literature review, including strengths and weakness of individual types of indicators.
In order to assess if any of these indicators can be measured based on vehicle sensor data, it is first necessary to discuss if and how collecting and using vehicle sensor data is even possible at a large scale. The literature review shows that there is already legislation in place that regulates access to various types of vehicle data, such as accident data (via eCall) and Safety Related Traffic Information (SRTI). OEMs own and have access to more data produced by vehicles, such as the vehicle sensor data that is the focus of this research. Several theoretical models are discussed to collect and share vehicle sensor data, of which the extended vehicle (ExVe)/neutral server model is preferred by the OEMs. This makes it the most likely model to be used in any future effort to use vehicle sensor data to measure traffic safety, given the importance of OEMs as stakeholders. In this model, vehicle sensor data is transmitted encrypted to dedicated servers of the OEM which can then make (processed) data available to third parties. This can be done directly or indirectly through a neutral server operated by a (consortium of) independent parties. Pilots such as the Proof-of-Concept Data for Road Safety show that the ExVe/neutral server model is not just a theoretical model but can effectively be applied in practice as

well. However, various potential barriers are identified that could prevent the large-scale collection and usage of vehicle sensor data.

Methodology

Given the novelty of the idea of using vehicle sensor data to measure traffic safety, no consensus has been reached on what types of indicators should be used and whether it would be feasible to measure these indicators on a large scale with vehicle sensor data. Therefore, this is asked to experts in the form of a Delphi study.

A Delphi study is a survey under specifically selected experts, making it fundamentally different from a regular survey. Where a regular survey has the goal to generalise results of a representative sample to a larger population, a Delphi survey aims to reach consensus among experts. The experts remain anonymous to each other to prevent influencing of the results. Next to this, the iterative character and controlled feedback provided to experts in the second round are important aspects of a Delphi study.

The Delphi survey consisted of questions that required experts to give a score on a 7-point Likert scale and provided room for an explanation. The Delphi survey was conducted over two rounds which asked the same questions. The second round included additional information at each question: a histogram showing the distribution of scores given in round 1 and arguments for these scores as given by the experts, both in favour and against. This allows the experts to re-evaluate their initial response, which should give a better result.

27 experts were invited, with backgrounds ranging from government to academia and industry. The goal of this was to ensure a heterogenous panel, which usually provides better results as varying perspectives are included. 16 experts completed the first round and 11 the second, with all experts scoring well enough on a set of criteria developed to establish a minimum level of expertise.

A selection of the four most promising types of indicators as identified in the literature review were evaluated by the experts on four criteria (validity, reliability, sensitivity, and understandability), as well as a fifth indicator based on the answers of multiple experts on the question if any other relevant potential indicator was not discussed in the first round. The types of indicators included are the Dutch SPIs, Proximity based SMoS such as Time-to-Collision (TTC), Kinematic based SMoS like swerving or strong acceleration, Engagement of ADAS such as Forward Collision Warning (FCW). Driver distraction measured by the Driver Distraction and Attention Warning (DDAW) was added in the second round.

Additionally, the experts rated nine potential barriers for the collection and usage of vehicle sensor data to measure traffic safety. The barriers included are technical feasibility, legal feasibility, willingness of OEMS, suppliers, and service providers, cybersecurity, and going from pilot to reality. Again, one additional potential barrier was added in the second round at the suggestion of several experts, being the willingness of people.

Results

The experts that filled in the survey were committed, as is shown by the high percentage (74%) of questions where the optional explanations were given in full sentences. In this Delphi study, consensus is measured as a certain level of agreement measured by two metrics where consensus is reached if both metrics reach a certain threshold based on academic literature. After two rounds, the experts reached consensus on 16 out of 20 questions related to the types of indicators, and on 5 out of 10 in questions related to potential barriers for using vehicle sensor data in practice. The topics that do not reach consensus are treated more cautiously in the remainder of the research. Further analysis is conducted to look into the cause of disagreement. Table 1 shows the results after two rounds on the questions related to the types of indicators with the median score and in between brackets the minimum and maximum scores given. A higher score means the experts rated a type of indicator better on a criterium.

Table 1 Overview of the median (minimum, and maximum) scores given by experts per type of indicator per criteria (on a 7-point scale; strongly disagree to strongly agree) where an Asterix (*) means no consensus is reached

	Validity	Reliability	Sensitivity	Understandability
Dutch SPIs	5 (3 to 6)	6 (5 to 7)	5 (2 to 7)	6 (5 to 7)
Proximity based SMoS	6 (4 to 7)	5 (2 to 6)	5 (4 to 6)	5 (1 to 6)
Kinematic based SMoS	6 (4 to 6)	5* (3 to 7)	5 (4 to 6)	5* (2 to 6)
Engagement of ADAS	6 (3 to 7)	5* (2 to 7)	5 (3 to 7)	5 (4 to 7)
Driver distraction	5 (4 to 7)	4* (2 to 6)	4,5 (4 to 6)	5 (4 to 7)

The results regarding the size of potential barriers for collecting and using vehicle sensor data in practice are summarised in a series of boxplots (figure 1). Each boxplot shows the median score given by the experts (the black dot), the middle 50% of scores (the blue box), and the entire range (the vertical line). The figure therefore shows both the size of a barrier as estimated by the experts, as well as the degree of agreement among the experts.

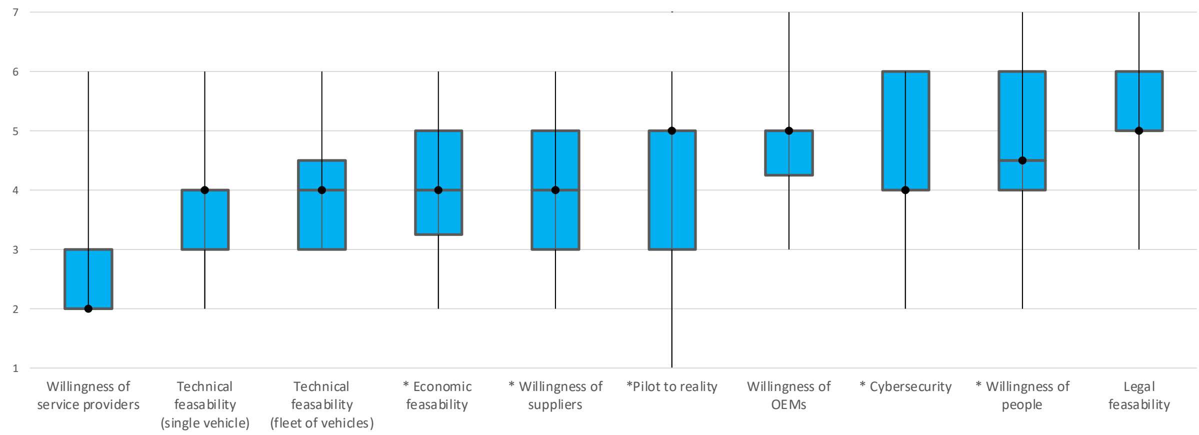


Figure 1 Boxplots potential barriers, where the black dot denotes the median score and the Asterix (*) means no consensus. Scale: 1=No barrier at all, 2=Weak barrier, 3=Somewhat weak barrier, 4=Moderate barrier, 5=Somewhat strong barrier, 6=Strong barrier, 7=insurmountable barrier

Discussion

Most types of indicators score well on most criteria, and limited differences are found between the types of indicators evaluated by the experts. This could indicate that these types of indicators would be suitable indicators to measure traffic safety at network level. However, the limited differences may also be found due to methodological errors or limitations regarding the way questions are asked or the experts involved. Further research should be undertaken to confirm these results.

The four most important barriers for collecting and using vehicle sensor data in practice are legal feasibility, willingness of people, cybersecurity, and willingness of OEMs. Legal feasibility is the largest barrier as the legal framework dictates the room for using vehicle sensor data. Uncertainty exists around upcoming EU regulations. If legal analysis of these regulations turns out positive, this research indicates that there are no other insurmountable barriers for using vehicle sensor data in practice. At the same time, this research also indicates that some of the barriers are sizable and thus form a starting point for further research and discussion on these specific topics.

It is important to note that no single indicator can tell the entire story of safety. Current standard practice of measuring traffic safety at network level is to use severe injuries and fatalities as indicators. Any of the types of indicator discussed in this research is aimed to supplement this current practice to help deal with their weaknesses.

It is important to note that the types of indicators discussed in this research only measure traffic safety from the point of view of motorised traffic, and only of the subset of the vehicles that have sufficient sensors and connectivity to collect and transmit the required data. Vulnerable road users and older vehicles are not included. Further research should focus on how well this subset of vehicles represents the traffic safety of all vehicles and further discuss on what types of accidents these types of indicators measure, and which it does not.

Conclusion and recommendations

This research explored the idea of using data collected by vehicles to measure traffic safety at network level. It has shown that if certain conditions are met and various barriers are addressed, it should be possible to collect and use vehicle sensor data to measure traffic safety. The main condition is that that upcoming legislation will provide sufficient legal room. This research has proven that, while it is still a novel idea surrounded by various issues, there is at this point enough of a path forward towards collecting and using vehicle sensor data to measure traffic safety to warrant further research and discussions on this topic. This research provides a starting point for this further research into the suitability of the types of indicators discussed in this research and on various current unknowns and the wider context of these types of indicators. Research could focus on the relationship between these types of indicators and specific types of accidents, the validity of this relationship, and the thresholds that could or should be used to signal traffic conflicts at different road types.

Additionally, research and discussions with OEMs should be initiated to address the barriers identified in this research. Future research should also focus on the differences between vehicle sensors and ADAS of different OEMs in order to understand its effects on the indicators that they measure. Lastly, the issue of measuring only the subset of the entire fleet of vehicle in the Netherlands that is the connected and equipped with sensors should be addressed

Preface

My grandfather once told me: “a problem has two sides, a beginning and an end”. I kept this in my mind during the process of working on this thesis. Whenever I was stuck, I tried to figure out what the particular problem was that I was facing (the beginning) and what I tried to achieve here (the end). If know you these two things, he told me, all you have to do is figure out is how to get from one to the other.

The end product is this thesis, *Opportunities and barriers for measuring traffic safety indicators based on vehicle sensor data: a Delphi study*, which is written in fulfilment of the degree of Master of Science in Transport, Infrastructure & Logistics. Writing this thesis has been a learning experience unlike all educational experiences before.

I want to thank Nina and Maxime for the opportunity to write my thesis at the ministry of Infrastructure & Water Management. Their support and extensive knowledge made this a better thesis. By inviting me to all kinds of meetings and events the ministry they have given me valuable experiences, conversations and insights that I could not have gotten from any academic literature, as well as a sometimes welcome change from my own laptop and work.

My gratitude also goes to Eleonora, Haneen and Bart. Thank you for challenging me, for your patience and for your guidance in writing this thesis. Thank you as well for the many friendly reminders to be as specific as possible. I have tried to take your advice in this thesis and will continue to do so in all my work to come.

Lastly, I want to thank my family, friends, roommates and Sanne in particular for their support. Without you this thesis would not have come to this end.

But the end has been reached. Not just of all the little problems encountered in this thesis, but of this thesis as a whole and with that my time at the TU Delft. At the same time this is a beginning, one I very much look forward to.

Steven Jansen
15-10-2022
Delft

Table of Contents

List of figures

List of tables

1. Introduction

1972 was the worst year in terms of traffic fatalities in the Netherlands, with 3264 fatalities (SWOV, 2020b). Ever since then, the number of traffic fatalities has been falling steadily. In fact, the declining trend was so strong that in 2008 the target for 2020 was lowered to a maximum of 500 traffic fatalities per year (Ministerie van Verkeer en Waterstaat, 2009). On the long term, the Dutch government has, together with the European Union (EU) as a whole, adopted the Vision Zero; with as goal to reach zero traffic fatalities by 2050 (European Commission, 2011; Ministerie van Infrastructuur en Watermanagemant, 2018). In the Netherlands, this Vision Zero is translated into the "Sustainable Safety" vision (Schagen & Aarts, 2018).
However, after reaching an all-time low of 570 traffic fatalities in both 2013 and 2014, the decline has halted and in 2020, 610 people died in traffic instead of the goal of 500 (L. Aarts et al., 2021). It is expected that without the COVID-19 pandemic and the accompanying measures this number would have been even higher (L. Aarts et al., 2021), as in 2019 661 traffic fatalities were recorded (SWOV, 2020c). For severe injuries, the target for 2020 is even further off with 19.700 severe injuries against a target of 10.600 (L. Aarts et al., 2021).

To increase traffic safety and reach the target of Vision Zero, the Dutch Ministry of Infrastructure and Water Management has identified nine policy areas, ranging from infrastructure to driver behaviour to technological developments (Ministerie van Infrastructuur en Watermanagemant, 2018). The European Commission (EC) identifies similar focus areas with infrastructure, vehicles and road use (European Commission, 2019). One specific development that both see as a way to increase traffic safety are Advanced Driver Assistance Systems (ADAS). Both also aim to promote the penetration and use of such systems. This can for example be seen in the fact that the Dutch Ministry of Infrastructure and Water Management has initiated the *ADAS Alliantie*, an alliance of a wide range of public and private institutions that aims to promote the safe use of ADAS (ADAS Alliantie, 2019). Figure 1 shows the penetration rate of several ADAS in the Netherlands.

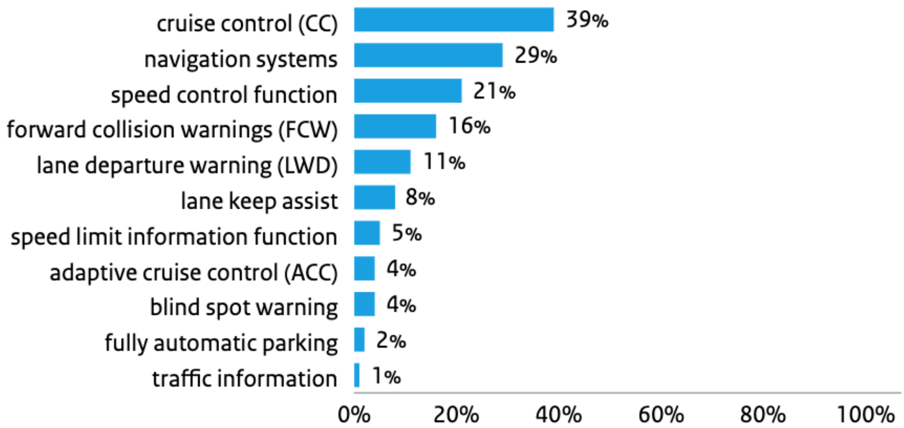


Figure 1 Penetration rate of several ADAS in the Netherlands on 01-12-18 (Tillema et al., 2020)

On an EU-wide level, the European Commission (EC) has revised its General Safety Regulation to mandate a variety of ADAS in order to be granted EU type-approval for new type of vehicles by July 2022 (Regulation EU 2019/2144, 2019). By July 2024, every newly produced vehicle must have these ADAS (Regulation EU 2019/2144, 2019). As a result, the penetration rate of ADAS, or at the least the penetration rate of these specific ADAS, will increase. This is expected to increase traffic safety. The EC estimates that these ADAS will have saved thousands of lives and ten thousands of severe injuries on EU roads by 2030 (European Commission, 2019).

The reasoning behind this is that the human driver is seen as the main contributing factor responsible for the majority of the accidents and that ADAS can help reduce these error. The ADAS takes away a single and relatively easy task from the driver and executes this in a more consistent way and at a higher performance level (Onderzoeksraad voor Veiligheid, 2019). In research for the US National Highway Traffic Safety Administration (NHTSA), Singh (2015) concluded that for 94% of almost 2.2 million accidents driver errors were the critical factor. Of these human errors, recognition error (41%) was the largest type of error, followed by decision error (33%) and performance error (11%). The fact that individual ADAS can increase traffic safety, or at the least do not negatively affect traffic safety, is also confirmed by for example Vlakveld (2019) in a literature review for the SWOV or in meta-analysis such as Wang et al. (2020).

1.1 ADAS as a challenge

Individual ADAS may be expected to increase traffic safety, but uncertainty about the size of the effects of ADAS on traffic safety remain. A large variety in effectiveness of specific ADAS is found in different studies, as is noted by both Wang et al. (2020) and Vlakveld (2019). For example, in a meta-analysis by Wang et al. (2020), one study found that Forward Collision Warning (FCW) could reduce the number of rear-end crashes by 41% while another study only found an effect of 12%. Research on the safety effects of other types of ADAS show similar ranges. Vlakveld (2019) even found sufficient conflicting studies on the safety effectiveness of Adaptive Cruise Control (ACC) that he cannot conclude anything about the effect of ACC. The reason why these large variances exist among different studies is threefold, according to Wang et al. (2020). Studies use different assessment methods, under different experimental conditions, and with varying driver conditions which can thus result in different effects found.

This is not just seen as a concern in the academic world. ERTRAC, an European public-private partnership between the EC and experts from both the industry and the academic world, also identify the lack of standardised testing methods for the safety effects of automated systems that can account for real-world scenarios with varying circumstances as one of the main challenges (ERTRAC, 2019).

The importance of looking at real-world scenarios is amplified by two more factors: the effects of the combination of various ADAS is seldomly studied, and even if all the technology works well in theory, people will still have to use it and use it correctly in practice.

Firstly, Wang et al. (2020) notes that the combination of various ADAS is seldom studied in the academic literature, even though increasingly more vehicles combine several ADAS. This could be a problem because of what Victor et al. (2018) calls the irony of automation. The irony is that introducing more automation to increase safety might lead to more unsafe situations. This can occur if people start to pay less attention to traffic due to having to carry out less tasks which could result in a slower reaction time in a situation which does require action from the driver. This risk is also highlighted by the Dutch Safety Board which points out that this will mainly be a problem in the current ‘hybrid situation’ in which both man and machine control the vehicle (Onderzoeksraad voor Veiligheid, 2019). Noy et al. (2018) add to this that partial automation in vehicles will change the dynamics of the vehicle under varying circumstances. This will require the driver to develop more skills to cope with these new situations but, due to the automation, fewer learning opportunities will occur. This irony of automation can change driving behaviour and is not accounted for in all safety assessment methods which can lead to an overestimation of the safety effects of ADAS in reality (Sohrabi et al., 2021).

Secondly, even if both individual ADAS and the combination of ADAS significantly increase traffic safety, and if these ADAS have penetrated the market in large numbers, consumers still need to use these systems and do this correctly in order to achieve the safety effects. Over a quarter of the consumers that have vehicles equipped with ACC do not or rarely use it while for Lane Keeping Assist (LKA) this is about a third (Boelhouwer et al., 2020). Harms et al. (2020) reach different conclusions based on a survey among Dutch business drivers. They compared the self-reported presence and usage of ADAS with the actual presence of ADAS according to the vehicle specifications. As can be seen in figure 2 below, almost all drivers that are aware of the ADAS in their vehicle use it but, especially with LDW and ACC, the majority of the drivers are not aware of the presence of these ADAS in their vehicle.

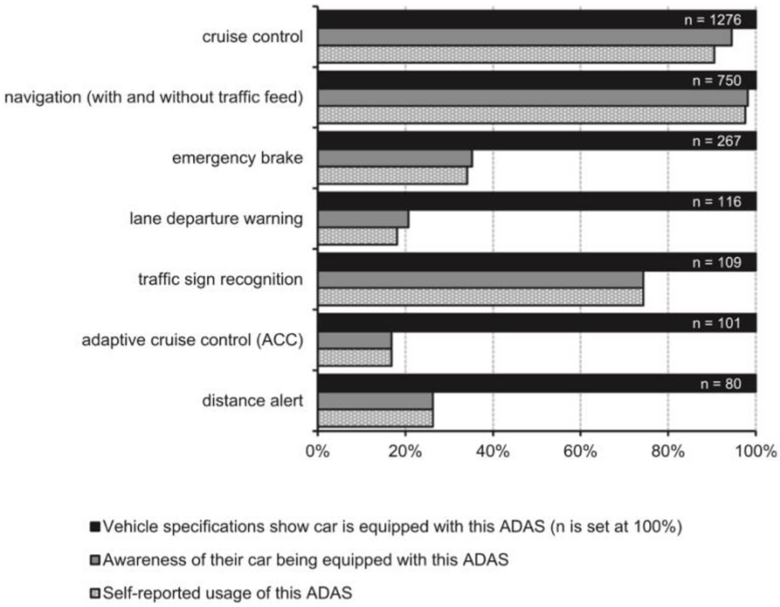


Figure 2 Awareness and use of ADAS (Harms et al., 2020)

These types of considerations are usually not taken into account by research into safety effectiveness of ADAS as most studies assume a 100% market penetration (Sohrabi et al., 2021). This risk is also recognised on a policy level, by the Dutch government in their strategic plans on traffic safety towards 2030 (Ministerie van Infrastructuur en Watermanagemant, 2018)

This uncertainty about the actual size of safety effects of ADAS, is a problem for two reasons. Firstly, this is a problem for policymakers. Policymakers and their executive authorities represent the society and as such, should act in its interest. This is not only the case because of the fatalities , injuries and suffering on a personal level that result from traffic unsafety but also for its associated costs for society as a whole (Ministerie van Infrastructuur en Water Managemant, 2018). In 2018, these costs are estimated to be 17 billion euros in the Netherlands, or well over 2% or its GDP (SWOV, 2020a).
Secondly, the (perceived) safety of ADAS is an important factor in the intent of consumers to use such systems (Sener et al., 2019). In a study by Sener et al. (2019), concerns about safety were the second most important reason for consumers to not want to use ADAS, behind the lack of trust in these systems. Less trust of consumers in the abilities and safety effects of ADAS could hinder the adoption of ADAS, as Sohrabi et al. (2021) argues. This could in turn lead to less safety benefits of these systems. Additionally, Blumenthal et al. (2020) also point at the importance of communicating with the public about safety of such systems because it is both a technical and a political issue due to the potential risks involved for the public.

1.2 Safety evaluation for driving automation systems

To help solve this problem, more insight is needed in how well vehicles with driving automation systems perform. A wide variety of safety evaluation methods for driving automation systems exist, each with their own advantages and disadvantages. An important distinction can be made between the several types of evaluation methods. Some methods (road test data analysis, Field Operational Test (FOT), Naturalistic Driving Study (NDS)) are based on real-world environments with others are based on simulated scenarios or even simulated environments (test track tests, driving simulator, traffic simulation, failure risk assessment). According to Vlakveld (2019), the first category is more valid while the second category can yield more reliable results. These methods are discussed more in-depth in appendix B.1.

In the case of ADAS and traffic safety, it is important to look at the real-world application of ADAS because of several uncertainties around this topic. Several methods, mostly in the second category named above, cannot explicitly deal with mixed-traffic issues (Sohrabi et al., 2021). Mixed-traffic issues are issues that may arise when (partially) automated vehicles are introduced into a network with human driven vehicles (HDVs). This could result in more heterogeneity in several driving elements such as speed and headway which in turn can lead to more potential conflicts (Virdi et al., 2019). Additionally, as discussed earlier, the irony of automation and the actual usage and penetration rates of ADAS will determine the actual effects of ADAS on safety.

This uncertainty about the influence of the irony of automation can be researched with driving test studies in which vehicles are outfitted with a variety of sensors and are driven on public roads. The most valid of these is the naturalistic driving study (NDS) because it allows to collect data on how “drivers typically drive in the wild’’ (Fridman et al., 2019), as opposed to a Field Operational Test (FOT) which follows a strict experimental design. However, these types of studies usually have a relatively short time span and are held with a small group of cars due to high costs (Vlakveld, 2019). As a result, accidents may not happen, and not enough data can be gathered.

This is also highlighted by Sohrabi et al. (2021) who claims that more data is required to draw reliable conclusions on the safety of driving automation systems because existing road tests are limited. Kalra & Paddock (2016) have calculated the number of miles that need to be driven in order to demonstrate that failure rates of vehicles with driving automation systems are lower than failure rates of vehicles driven by humans. The results can be seen in figure 3. If vehicles with driving automation systems would be 20% better than human drivers, it would still take 11 billion miles to statistically prove this. This shows that there is a large need for more data in the safety evaluation of vehicles with driving automation systems.

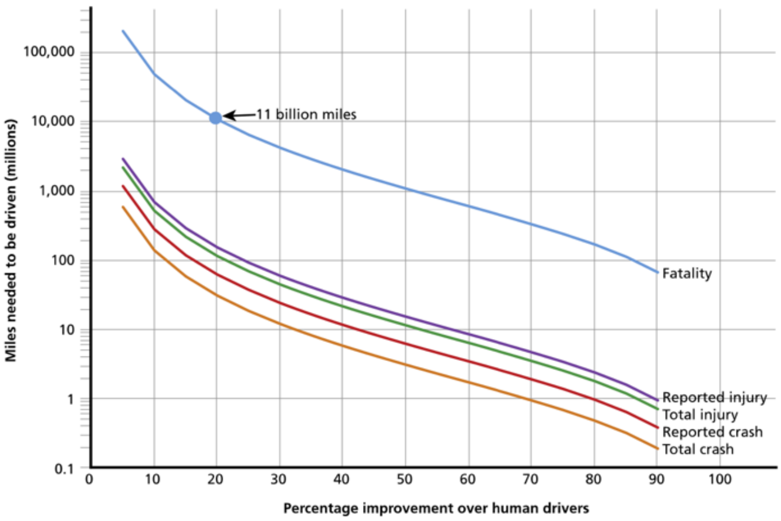


Figure 3 Miles needed to demonstrate with 95% confidence that failure rates of vehicles with driving automation are lower than human failure rates (Kalra & Paddock, 2016)

Currently, data used in safety evaluation methods vary, both in the type of data and the way of collecting it. It can include crash statistics, observational data from fixed point sensors, on-board data collection with smartphones or with research-designated sensors. All of these have their problems as is discussed in Appendix B.2.
Retrofitting vehicles with research-designated sensors is a flexible method that can detailed datasets. It however also comes at a high financial cost. This poses a problem as this data is necessary as is shown in this review. As Talal et al. (p25, 2020) concludes about DAS: "providing a low-cost, reliable and easy-to-implement system is a tremendous step towards research advancement".

1.3 ADAS as an opportunity

However, as much as ADAS may be a challenge in the sense that the size of the effects on traffic safety are unknown, it may also be a solution. For ADAS to work, vehicles are becoming increasingly smarter and equipped with more and better sensors like GPS, radar, cameras or even LiDAR (Ackermann et al., 2019). The market penetration rate of these systems will only increase, especially once several ADAS become mandatory in the EU in new types of vehicles in 2022 and in all new vehicles from 2024 (Regulation EU 2019/2144, 2019). The question can be raised if these already present sensors could be used to collect data for research purposes, in order to supplement data collected through current methods. This data can take many forms, it can be data produced by ADAS, but also by other parts of the vehicle actuator input, speed and acceleration, or GPS data. Some examples from various perspectives are discussed below that use different types of data produced by vehicles.

In the academic world, studies are carried into the using vehicle data to measure traffic safety. For example, Jang et al. (2020) uses data gathered by forward collision warning (FCW) to measure several surrogate measures of safety (SMoS) such as the time-to-collision (TTC) and the crash potential index (CPI). These are then used to identify the effects of forward hazardous situation warning (FHSW). This study did use a separate system to collect and transmit the data from the ADAS. Xie et al. (2019) does something similar by using real-world data from connected vehicles to identify high risk locations. They do this based on a newly developed SMoS: time-to-collision with disturbance (TTCD).
According to interviews with members of the ADAS development team from a Swedish leading automotive company by Orlovska et al. (2020), car manufacturers already use data from ADAS for internal verification and validation of these ADAS. However, once the vehicles enter the market, little interest is given to the follow-up of ADAS performance. This is because the manufacturers assume the ADAS to be validated and verified sufficiently within their specified limitations (Orlovska et al., 2020).
One last example is found in which a government uses data collected by vehicles with the aim to increase safety. Since the end of 2021, vehicles in the Netherlands will automatically share data with road authorities to help detect damages to the infrastructure or icy roads in the winter (Minsterie van Infrastructuur en Waterstaat, 2021). This can help road authorities to maintain their roads more efficiently and ultimately improve traffic safety (Minsterie van Infrastructuur en Waterstaat, 2021).

This shows that new and emerging data sources like vehicle sensor data could potentially help create advances in safety assessment methodology. This trend is shown in figure 4 below by Mannering & Bhat (2014). It shows that with an expanding data frontier due to new data sources, new methodological opportunities can emerge to leverage this new data.

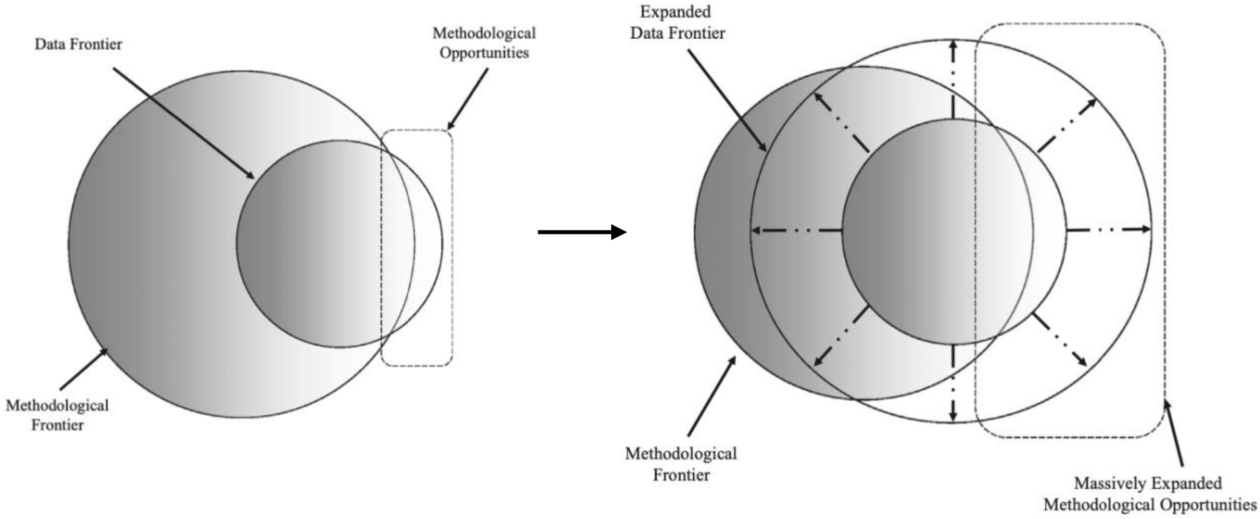


Figure 4 State of methodological research with emerging data sources (Mannering & Bhat, 2014)

1.4 Literature gaps and research question

So, based on this overview of academic literature, the following literature gaps can be defined:

1. There is no agreement in scientific literature on the estimates of the safety benefits of ADAS in practice.
2. The effect of a combination of various ADAS on traffic safety is seldomly studied in the academic literature.
3. Current data collection methods for the safety evaluation of ADAS either lack sufficient depth and reliability or are too expensive, showing the need for a more cost-effective, reliable, and easy data collection system.

These literature gaps can potentially be filled in using new methodologies that have become possible due to the larger availability of data, produced, and collected by vehicle sensors. The question can be raised if the safety benefits of ADAS in practice can be proven this additional data. After all, if in an ideal world all vehicles would be able to collect and transmit data on driving behaviour, it would be possible to conduct what is essentially a Naturalistic Driving Study on an unprecedented scale.
Proving the safety benefits of ADAS in practice implies the need for a baseline or a comparison to vehicles without ADAS. However, comparing vehicles with automation to fully human driven vehicles (HDVs) is highly difficult and usually unfair, as will be discussed in the literature review (section 2.2).

It may however be possible to measure the safety performance of vehicles with ADAS on its own, so without comparing it to HDVs. This essentially amounts to measuring traffic safety which is an essential part of monitoring the effects of traffic safety policy (Ministerie van Infrastructuur en Watermanagemant, 2018; SWOV, 2005). Additionally, it is also useful for estimating current (specific) aspects of traffic safety and for comparison with other countries (ETSC, 2001). However, current practice on measuring traffic safety does have its problems for which measuring traffic safety based on vehicle sensor data could potentially be a good solution as will be shown below.
Traffic safety policies used to be based on a reactive approach of “black sports”, locations with a relatively high number of accidents (Ministerie van Infrastructuur en Watermanagemant, 2018). By now, most of these locations have been made safer which in turn means that the accidents that are still happening are more spread out over the network (Ministerie van Infrastructuur en Watermanagemant, 2018). Therefore, monitoring traffic safety at network level has become more important.
Measuring traffic safety at network level is currently being done with severe injuries and fatalities as indicators. As will be discussed in section 2.2 of the literature review, these indicators are useful but not without their problems. Accidents, and especially severe accidents, are rare. By using these as

indicators for traffic safety, any policy is reactive and may result in underlaying causes to remain unaffected (Ministerie van Infrastructuur en Watermanagemant, 2018).
A more pro-active approach of measuring traffic safety is to measure dangerous situations that nearly became an accident: traffic conflicts or critical events (see section 2.2 of the literature review). These events are much less rare and therefore allow for a more pro-active approach of measuring traffic safety. Measuring these critical events requires data that was previously difficult or even impossible to collect but this could change completely with the advance of vehicles full of sensors that gather large amounts of data.

1.4.1 Research objective and relevance

The objective of this research will be to explore if data collected by vehicles equipped with ADAS can be used for measuring traffic safety at network level. It will focus on two main aspects: indicators for measuring traffic safety at network level based on vehicle sensor data, and on the feasibility of using vehicle sensor data to do so in practice. So, firstly it aims to look at what could be measured and secondly at the question if it even is possible to use vehicle sensor data to do just that in practice.

This research has a clear societal relevance because it can contribute to improving the way in which traffic safety is measured. This can help to better evaluate traffic safety policies, decrease the number of accidents, reach the goal of Vision Zero and ultimately save lives.
It also is scientifically relevant because it helps contribute to the body of knowledge of traffic safety evaluation. Current methods and data sources are lacking in various ways, as is shown in the literature review. This research could help explore if data collected by vehicles could help overcome some of the currently existing issues and ultimately improve the way in which traffic safety is measured currently and might in the future contribute to the evaluation of the safety benefits of vehicles equipped with automation.

1.4.2 Research question

The identified literature gaps in combination with the research objective results in the following research question:

(How) can data from sensors of vehicles equipped with ADAS be used to measure traffic safety at network level?

The wide term vehicle sensor data is used on purpose to avoid missing potentially valuable data sources within the vehicle. In this research, vehicle sensor data refers to all data produced by sensors in a vehicle. This can take various forms like data produced by ADAS, but also by other parts of the vehicle such as actuator input, speed and acceleration, or GPS data.

With measuring traffic safety at network level is meant any indicator that measures traffic safety independent of a specific location but instead the performance on the network as a whole.

In order to answer this question, several sub-questions need to be answered first.

1. **How is traffic safety measured currently?**
   Before new ways to measure traffic safety based on vehicle sensor data can be developed, it is first necessary to understand how traffic safety is measured currently. Measuring traffic safety is done using indicators which differ for specific use cases where the indicator used also depends on the goal.
   Two important and distinctively different groups that measure traffic safety exist: governments and researchers. These two groups may look and measure traffic safety from a different perspective because of different goals and because of practical limitations. What could be a useful and practical way to evaluate traffic safety with a small number of vehicles in a controlled environment for a researcher may be very different for the whole of the Netherlands by the ministry of Infrastructure and Water Management. However, both perspectives can help to identify potential indicators based on vehicle sensor data that can be used to measure traffic safety at network level. Therefore, this sub question is split in two:

   a. *What indicators are currently used by the Dutch government to measure traffic safety?*

   Firstly, it will give an overview of different ways in which traffic safety is currently defined and how this is measured. This concerns ways in which general traffic safety is measured in practice at network level.

   b. *What indicators to measure traffic safety are proposed in the academic world?*

   Secondly, it will look at ways researchers and the academic world measure traffic safety. This is often more focused on evaluating traffic safety at individual vehicle level in specific circumstances.

2. **What are possible scenarios for using vehicle sensor data to measure traffic safety in practice?**
   In order to look at how vehicle sensor data could be used to measure traffic safety, it is also necessary to look at the data these vehicles produce and if and how this data can be extracted from vehicles so it can be processed into useable information.

3. **What are feasible and suitable indicators based on vehicle sensor data to measure traffic safety at network level?**
   This research question applies the knowledge from the first research question about indicators for measuring traffic safety to the context of the second research question: using vehicle sensor data to do so.

4. **What barriers exist for the collection and usage of vehicle sensor data to measure traffic safety in practice by the Dutch Ministry of Infrastructure and Water Management?**
   Using vehicle sensor data to measure traffic safety in practice is a new concept which has not been applied on a large scale in the Netherlands. A variety of barriers can be thought of that have prevented this, such as privacy regulations and technical feasibility. This research question will look at such barriers and assess to what extent they exist.

Figure 5 below gives a graphical overview of the logic and design of this research. The first two research questions (1a and b, 2) will result in potential indicators based on vehicle sensor data to measure traffic safety. The input for this will be a review of relevant literature of both scientific and policy related documents. These potential indicators based on vehicle sensor data safety will then be evaluated by experts in the form of a Delphi study (to be introduced below) to assess how feasible and suitable these metrics are for measuring traffic at network level. This will answer research question 3. To answer research question 4, expert knowledge will also be used to assess potential barriers for the collection and usage of vehicle sensor data to measure traffic safety.

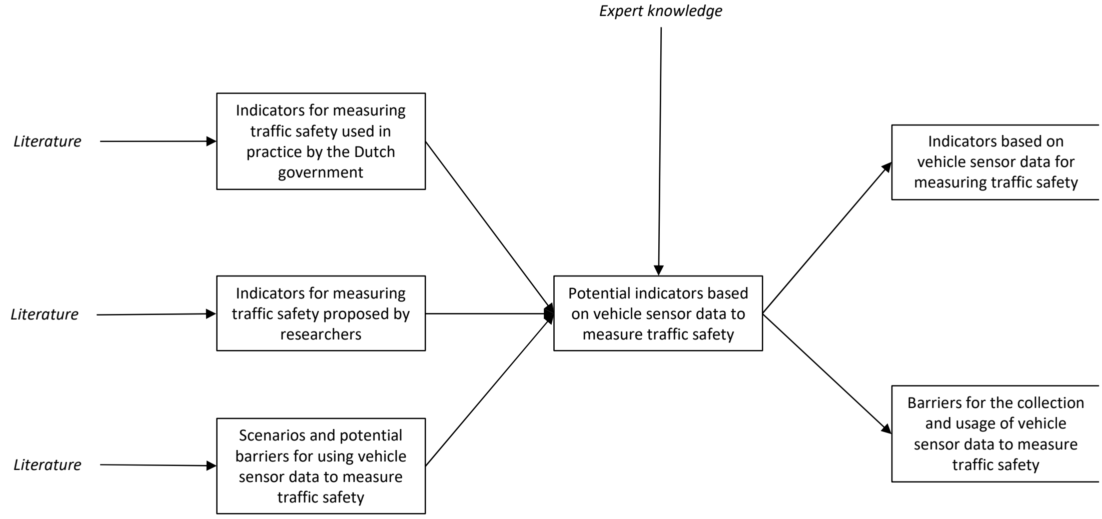


Figure 5 Graphical overview of the research

1.4.3 Scope

There is a wide range of ADAS with a large variety of functions that also differ amongst different car manufacturers. For this research, the ADAS in table 1 will be included. The choice for these five ADAS is based on the EU 2019/2144 regulation which will make these ADAS mandatory by 2022 in new type of vehicles and for all new vehicles by 2024. As a result, the penetration rate of these systems will rise sharply, as can be seen in figure 6. Figure 6 shows the penetration rate of emergency lane-keeping system (called LKA in the figure) which will rise steadily to approximately 50% in 2030 (MuConsult, 2021).

Table 1 Overview of ADAS included in the research (based on ADAS Alliantie, 2019; Regulation (EU) 2019/2144, 2019)

ADAS	Description
Intelligent Speed Assistance (ISA)	ISA gives feedback to the driver to help adhere to the maximum speed
Driver drowsiness and attention warning (DDAW)	This system asses the driver's alertness through vehicle systems analysis and warns the driver if needed
Advanced driver distraction warning	This system warns the driver when the driver is distracted
Advanced Emergency Breaking (AEB)	AEB is a further development of FCW and automatically applies the brakes when it detects a potential crash
Emergency lane-keeping system	This ADAS keeps the vehicle within the boundaries of its lane when a lane departure occurs and a collision in imminent

EU 2019/2144 includes other ADAS such as the alcohol interlock installation facilitation or reversing detection. These are excluded in this research because these do not directly impact the actual driving of the vehicle. Furthermore, the research will focus on Dutch roads.

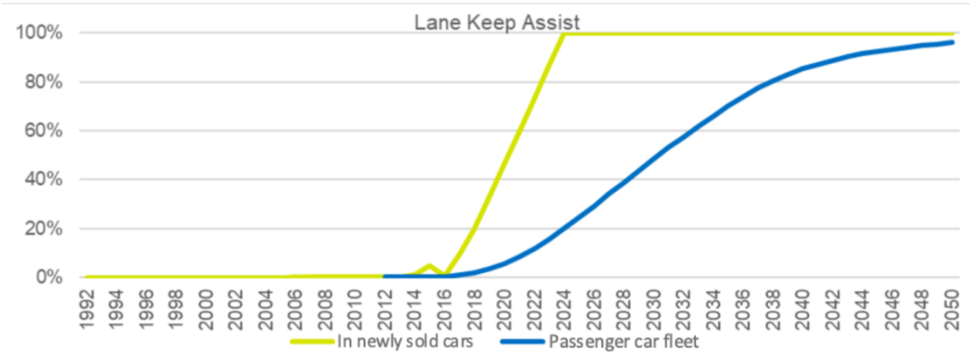


Figure 6 Expected rise in penetration rate of LKA in newly sold cars and in the passenger car fleet (MuConsult, 2021)

1.5 Overview of methodology

As is shown in figure 5, both literature review and expert judgement will be used in this research. The first two research questions will be answered with a review of relevant literature of both scientific and policy related documents.

Literature review

Scientific literature will mainly be collected via ScienceDirect and Scopus, as well as Google Scholar. To find relevant papers keywords like measuring traffic safety, at network level, vehicle sensor data, vehicle data, and combinations of these will be used. When specific indicators are mentioned in papers, more literature will be looked up in these databases on these specific indicators such as Time-to-Collision (TTC) or more in general on Surrogate Measures of Safety (SMoS).

Next to scientific and peer-reviewed papers, policy documents are a valuable source of information for this research. The Dutch government, in all her forms, plays and important role in traffic safety. This includes traffic safety policies as well as measuring traffic safety to monitor the effects of the policies, but also in a wider role that sets the frameworks and laws, for example on data protection. Policy documents that will be used as sources in this thesis are therefore policy documents from the Dutch government or from the European Commission or research and policy evaluations carried out by for example the Dutch Safety Board (*onderzoeksraad voor Veiligheid, OvV*) or the national scientific institute for road safety research (SWOV). These are found through their respective websites, by searching on Google, and on suggestions made by employees of the department Smart Mobility of the Ministry of Infrastructure and Water Management.

Additionally, one interview will be conducted to gain additional insight into an ongoing pilot project that uses vehicle sensor data to assess road quality, slipperiness due to ice, and traffic safety. The pilot is still ongoing at the time of this research, so no written evaluation was available but given the high relevance of this pilot it is included in this way.

From the literature review, two intermediate results are derived:

1. A list of potential indicators based on vehicle sensor data for measuring traffic safety at network level.
2. A list of potential barriers for the collection and usage of vehicle sensor data in practice.

The question about how suitable these potential indicators are for measuring traffic safety on a network level and the question on the size of the potential barriers of using vehicle sensor data to do so are then asked to experts in the form of a Delphi study. Additional literature research will be conducted into the workings of the Delphi method to ensure an as rigorously conducted Delphi survey as possible.

Expert judgement is needed because using vehicle sensor data to measure traffic safety is a relatively new concept so limited information is available in the literature. The Delphi method allows a researcher to leverage the knowledge of experts to assess new technologies without the need for extensive data which would be needed for example in a simulation study.

Delphi Method

Skulmoski et al. (2007, p2) describe the Delphi methods as “an iterative process used to collect and distil the judgments of experts using a series of questionnaires interspersed with feedback.” Simply put, experts are invited to respond to a series of questions. The researcher then combines and summarises the responses and again asks the experts to respond to the same questions in order to reach consensus (Belton et al., 2019). The Delphi method has its origin in US military research in the 1950s in which it was used as a forecasting method (Okoli & Pawlowski, 2004). From there, it has been used in, and adapted to, a wide variety of fields such as healthcare and quality-indicator development (Belton et al., 2019), policy related questions (de Loë et al., 2016) and in transport (Melander, 2018). The common denominator is that the Delphi research is about the future. Hsu & Sandford (2007, p1) put it eloquently: “Common surveys try to identify “what is,” whereas the Delphi technique attempts to address “what could/should be”.” It works well when there is incomplete knowledge about a problem and the goal is to improve the understanding of this problem, opportunity or solution (Skulmoski et al., 2007).

That is the case in this research. As is shown in this introduction, measuring traffic safety based on vehicle sensor data could be a solution to existing problems with monitoring traffic safety at network level and a step forward in the safety evaluation of vehicles equipped with automation. But how this would work in practice and what would be useful indicators is not known.

Delphi method is chosen over other methods that use the input of expert like interviews because Delphi method can combine qualitative and quantitative questions in a structured way. This then allows for a fair comparison of different indicators and potential barriers for measuring traffic safety based on vehicle sensor data.

How the Delphi method works in detail and how it is applied in this research is discussed in chapter 3 Methodology and operationalisation of the Delphi Study.

1.6 Outline of the study

This study consists of six chapters in total. This introduction is followed by a literature review that aims to answer research questions 1a, 1b, and 2. It concludes with a list of potential indicators based on vehicle sensor data for measuring traffic safety at network level and with a list of potential barriers for doing so in practice. This is input for chapter 3: Methodology and operationalisation of the Delphi Study which explains how a Delphi survey works in general and how it is applied in this study.

The results are discussed in chapter 4, followed by the discussion in chapter 5 and finally the conclusion in chapter 6.

2. Literature review

Traffic safety lies at the core of this research and forms the start of this literature review it in the form of RQ1: How is traffic safety measured currently? The first part of the literature review focusses on traffic safety (section 2.1) and on measuring it with indicators used by both the Dutch government (section 2.2) and by academic researchers (section 2.3), as these are the main actors involved in measuring traffic safety. The second part of the literature research focusses on vehicle sensor data and its practical application, aiming to answer RQ2: What are possible scenarios for using vehicle sensor data to measure traffic safety in practice? The research focusses on types of vehicle sensor data (section 2.4) and suggestions in academic literature on how this data can be shared (section 2.5). It also looks at the practice by discussing pilots that involve using and sharing vehicle sensor data (section 2.6) and which lessons are learned from this by identifying potential barriers for collecting and using vehicle sensor data (section 2.7).

2.1 Traffic safety and ways to measuring it

Traffic safety is a broad concept with many aspects that are not easily defined. This section will show this by introducing two frameworks on traffic safety and potential opportunities for safety measures. It also introduces indicators and the difference between lagging and leading indicators.

2.1.1 Traffic safety as a concept

Traffic safety, or even safety in general, does not have a consensus definition (Fraade-Blanar et al., 2018). The exact definition often depends on the context in which it is used or on the specific situation it is applied to. However, these definitions are usually centred around one concept: lack of harm (Fraade-Blanar et al., 2018). So, traffic safety is defined from the opposite of safety, unsafety or risk as Kulmala notes (2010). He for example defines it as “the expected number of fatally or otherwise injured persons of an entity in a unit of time” (p1360, 2010). In this definintion, an entity can refer to a variety of aspects of traffic ranging from a specific road section or a certain type of junctions to (groups of) drivers or vehicles (Kulmala, 2010).

This already shows that traffic safety consitsts of several aspects. Kulmala (2010) works this out in his safety assesment framework (figure 7). Based on earlier research by Thulin & Nilsson (1994), he defines three dimensions of traffic safety: exposure, crash risk and consequence. Exposure is the size of potential exposure to accidents which, depending on the intended purpose, can be measured in kilometers traveled per driver, vehicle or in total. Crash risk is the risk of being in an accident per million person kilometer while consequence is the probability of (serious) injury or death as a result of the crash.

Alongside the three dimensions of safety, Kulmala (2010) also included the three levels of road user decision making (strategic, tactical, and operational) and the type of effect (engineering and behavioural). He then identified nine different types of safety measures can be taken that cover all three of these aspects. This shows that traffic safety has a variety of aspects with many potential safety measures that impact these through a complex system.

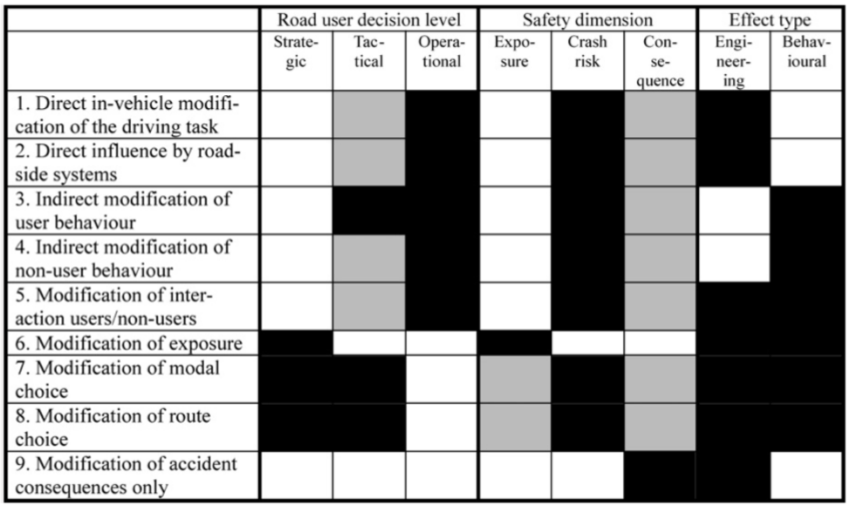


Figure 7 The safety assessment framework of Kulmala (2010) by road user decision level, safety dimension and safety effect type. Black colour indicates that the mechanism typically focuses on that aspect, grey means relevance but no focus on the aspect

That traffic safety is complex and consists of a variety of factors on which safety measures work can also be seen in the Haddon Matrix (table 2). The Haddon Matrix is a framework in which each cell is a possible opportunity for safety measures. It consists of the three stages of a crash – pre-crash, crash, and post-crash – and of three factors that can interact at each stage of a crash: the human, the machine, and the environment (Fraade-Blanar et al., 2018; Peden et al., 2004).

Table 2 The Haddon Matrix (Peden et al., 2004)

Phase	Human Factors	Vehicle and Equipment Factors	Environmental Factors
Pre-crash	a. Information b. Attitudes c. Impairment d. Police enforcement	a. Roadworthiness b. Lighting c. Braking d. Handling e. Speed management	a. Road design and road layout b. Speed limits c. Pedestrian facilities
Crash	a. Use of restraints b. Impairment	a. Occupant restraints b. Other safety devices c. Crash-protective design	a. Crash-protective roadside objects
Post-crash	a. First-aid skills b. Access to medics	a. Ease of access b. Fire risk	a. Rescue facilities b. Congestion

The three phases of the Haddon Matrix are also known as primary, secondary, and tertiary prevention. Primary prevention is about preventing the crash as a whole, secondary prevention is about minimizing the injuries sustained from the crash while tertiary prevention is about the medical aftercare. This is similar to the three safety dimensions (exposure, crash risk, and consequence) of the safety assessment framework by Kulmala (2010)

From this paragraph can be concluded that traffic safety can be understood in terms of unsafety and safety measures are aimed at reducing this unsafety. It is a complex system with a large variety of factors and potential points to influence this safety.

2.1.2 Measuring traffic safety with indicators

Measuring traffic safety is highly difficult. As discussed above, there is no single definition of traffic safety and there is a large variety of factors that influence traffic safety. No model exist that can fully explain traffic safety, with all the relevant factors and their corresponding importance (Stipdonk, 2013). And even if there was such a model, it would require accurate data on all those factors which is often not available (Stipdonk, 2013).

So, indicators are used to measure traffic safety instead. It is important to note that no single indicator exists that can fully describe traffic safety (Fraade-Blanar et al., 2018). In general, there are two types of indicators for measuring traffic safety at network level: lagging and leading indicators. Lagging indicators measure events like accidents. Thus, the event first needs to happen. Leading indicators are proactive and measure events leading up to accidents, meaning that accidents do not need to happen before traffic safety can be measured (L. T. Aarts, 2018). So, leading indicators can serve as surrogates or proxies for lagging indicators (Fraade-Blanar et al., 2018).
This distinction between the two types of indicators can also be illustrated with the Haddon Matrix (table 2). Leading indicators draw from the first row (pre-crash) while lagging indicators are related to the second and third row (crash and post-crash) (Fraade-Blanar et al., 2018).

2.2 Indicators used for measuring traffic safety from a policy perspective

The Dutch government uses both lagging and leading types of indicators of indicators to measure monitor the development of traffic safety at network level over time. These will be discussed in 2.2.1 and 2.2.2. Evaluating traffic safety can also be done at vehicle level, with ex-ante or ex-post safety evaluation. This is discussed as well, to see if any indicators are used in those process that may be relevant for measuring traffic safety at network level based on vehicle sensor data.

2.2.1 Fatalities and severe injuries

The two most important and widely used lagging indicators are the number of fatalities and number of severe injuries (L. Aarts et al., 2021) . The SWOV, the national scientific institute for road safety research in the Netherlands, publishes the State of Road Safety (*De staat van de verkeersveiligheid*) on a yearly basis which presents these from the year before. In these statistics, a fatality is someone who dies within 30 days of the accident and a severe injury is someone with an AIS score of 2 or higher (SWOV, 2016). AIS stands for Abbreviated Injury Scale and is a way to classify injuries on a scale from 1 (minor) to 6 (maximal/untreatable) with 2 being a moderate injury (SWOV, 2016).
These two lagging indicators have a few strong advantages. They are well established as metric and thus have a clear and uniform definition, resulting in relatively high quality of the data (Blumenthal et al., 2020). Additionally, they are easy to understand for both the general public and policy makers (Blumenthal et al., 2020). Lastly, given their long use it is possible to compare trends over time (Stipdonk, 2013). In other words, fatalities and severe injuries can “tell the final story on if it is safe or not.” (Blumenthal et al., 2020, p12).

However, there are five issues with using crash statistics. The first problem concerns the registration of these statistics. The total number of fatalities come from the official statistics of Statistics Netherlands (CBS) (SWOV, 2020c) which consists of a combination of three sources (Centraal Bureau voor de Statistiek, n.d.). These are the official cause of death forms (*doodsoorzaakformulieren*) as signed by a medical doctor, municipal records and police records (CBS, n.d.). However, this data is only available since 1996 and includes a limited number of characteristics: mode of transportation[1], age, gender, and location (province).
The police also records accidents with fatalities and severe injuries which include up to 40 relevant variables (SWOV, 2016) such as opposing party, road type and weather conditions (Stipdonk, 2013). The problem with the police records, collected in a system called BRON, is that these are not complete. Of the 661 fatalities in the official statistics of the CBS, 586 are registered in BRON, a registration rate of 89% (SWOV, 2020c). In the years from 1996 to 2005 (the first ten years of BRON) the registration rate was 94% on average while the registration rate in 2009 was 83% (Stipdonk, 2013).
The registration rate of injuries is lower and depends on the severity of the injuries where less severe injuries are less likely to be registered (correctly) (SWOV, 2016). Additionally, definitions and databases

[1] Road oriented modes only: car, bicycle, motorcycle, pedestrian, moped, mobility scooter, truck/van

have changed since the first recorded traffic fatalities in 1950, making historical analysis difficult (Stipdonk, 2013).
So, there is underreporting of both fatalities and severe injuries which is not uniformly distributed throughout time and (most likely) over types of crashes. This makes analysis difficult. Next to the problem of data incompleteness, four other problems with crash statistics are identified by researchers.
The second problem with using fatalities and severe injuries as indicators for traffic safety is that crashes are subject to random fluctuations (Chang et al., 2017; ETSC, 2001). So, changes in fatalities throughout the years do not necessarily mean changes in traffic safety. Also, the number of fatalities and severe injuries does not say anything about the processes that have caused these fatalities and severe injuries. In any given situation there will be a probability that a crash might occur, and whether this happens or not is to a certain extent up to chance. So, sometimes a relatively safe situation might result in a crash while a hazardous situation does not. (ETSC, 2001)
The third reason why using fatalities and severe injuries as the only indicators are problematic, is that they do not help understand the processes that cause accidents (ETSC, 2001). According to Chang et al. (2017), crash reports are set up in a way to attribute responsibility instead of searching for the causes of the crash. And this while understanding the causes of crashes are necessary to develop effective safety measures.
A fourth problem is brought forward by many researchers, as is outlined in research by Tarko (2012). As can be seen in figure 8, the riskiness of an event is inversely related to the frequency of an event. This means that accidents are rare occurrences which makes (statistical) analysis more difficult (Chin & Quek, 1997; Tarko, 2012). This is also neatly illustrated in Hydén's Safety Pyramid which can be seen in figure 9.

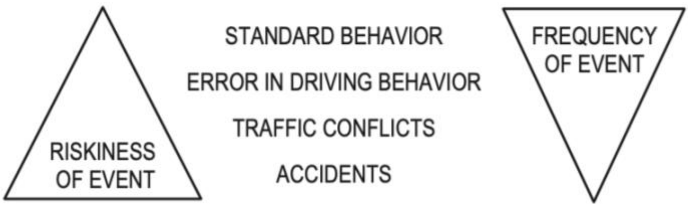


Figure 8 The steps from standard behaviour to actual accident (Klebelsberg, 1964, in Tarko, 2012)

This introduces the fifth and last problem. One of the goals of measuring traffic safety is to monitor and evaluate safety measures but by measuring traffic safety with lagging indicators like injuries and fatalities might take a long time. Using crash statistics is thus a reactive approach in which a sufficiently large number of accidents need to occur, before a problem can be found and addressed. Many researchers (Arun, Haque, Bhaskar, et al., 2021; Chin & Quek, 1997; Mahmud et al., 2017; Tarko, 2012; L. Wang et al., 2020) mention that this raises ethical questions as people first need to crash and perhaps even die before action can be taken to prevent those crashes and fatalities .

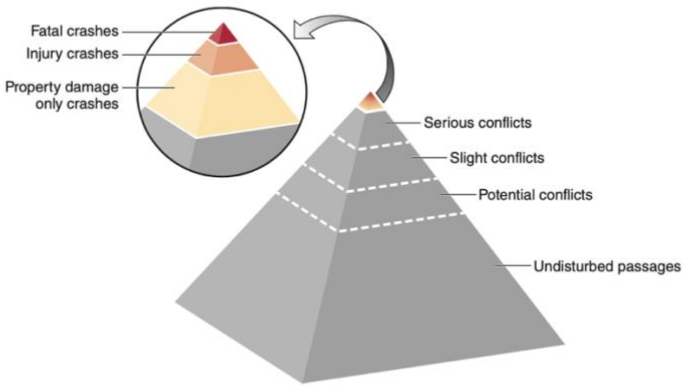


Figure 9 Hydén's Safety Pyramid (Chang et al., 2017)

2.2.2 Dutch SPIs

Leading indicators are proactive indicators. In the context of traffic safety, this means that these indicators can measure traffic safety without accidents needing to happen (L. T. Aarts, 2018).
In the Netherlands, this has resulted in the concept of risk-based policy (*risciogestuurd beleid*). The idea is to map and reduce risks in order to prevent risks (Ministerie van Infrastructuur en Watermanagemant, 2018). These risks are measured in Safety Performance Indicators (SPIs) which the SWOV defines as follows: SPIs are indicators for risk factors or operational conditions of traffic systems that can be used to measure traffic safety and help understand the processes leading to accidents and the accompanying injuries and fatalities (L. T. Aarts, 2018). As an example, the following SPIs are used (Ministerie van Infrastructuur en Watermanagemant, 2018):

- Road quality (share of drivers driving over roads classified as 'sufficiently safe')
- Speed (share of drivers that speeds)
- Vehicles (share of vehicles that meet the norm)
- Participants
    - Usage of protective equipment (share of drivers using protective equipment)
    - Usage of lighting (share of drivers that use the correct lighting)
    - Usage of drugs/alcohol (share of drivers under the influence)
- Handling of accidents (share of traffic victims receiving professional medical help within the set standard)

It is important to note that all the SPIs are defined as the "share of…", and not just the "number of…". The latter does not tell the complete story as exposure is not taken into account (L. T. Aarts, 2018). Stipdonk (2013) illustrates nicely how important it is to take exposure (in any form) into account by showing figure 10, in which both the fatalities and fleet size of mopeds can be seen. In 1974, helmets became mandatory, and the number of fatalities decreased sharply. This, however, does not necessarily mean that helmets are (fully) responsible for this decrease as the fleet size also decreased sharply.

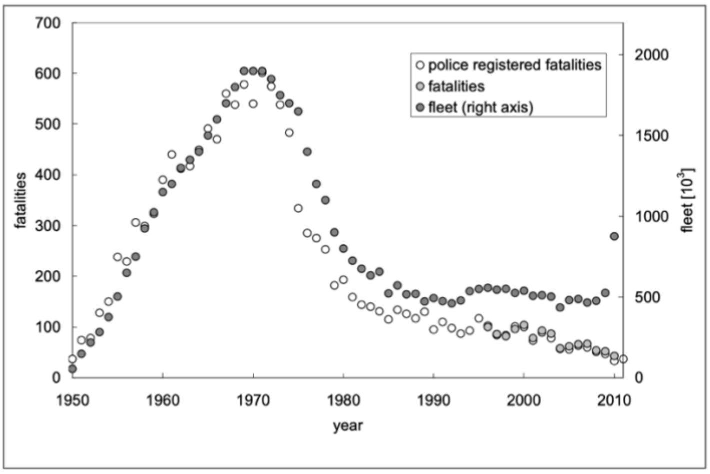


Figure 10 Number of moped fatalities and moped fleet size (Stipdonk, 2013)

Furthermore, the list of SPIs is meant as an example, not a full list with complete definitions[2]. For example, the first SPI about road quality can be specified into motorised traffic and cyclists and protective equipment can refer to seatbelts in cars or to helmets on scooters. Whichever is used depends on the aim of the monitoring. Additionally, this concept of risk-based policy using SPIs is still in the process of being implemented throughout all types of Dutch governments. For the period 2020-2025, the goal is to implement, use and especially learn about the SPIs with an evaluation and professionalisation period between 2025-2030 in which the SPIs can be modified and changed if necessary. (Ministerie van Infrastructuur en Watermanagemant, 2018)

[2] For the full list and definitions: see Aarts (2018)

2.2.3 Ex-ante and ex-post safety evaluation at vehicle level

Measuring traffic safety is a complex task that can be done with several types of indicators. However, measuring the effect of (partial) automation on traffic safety is an even more complex and difficult exercise as “the integration of the driver and vehicle invalidates or changes how traditional measures can be used and gives rise to new needs” (Fraade-Blanar et al., 2018, p14). Looking at safety evaluation at vehicle level may give insight into how to deal with this issue, as vehicles with automation have entered the market, which requires type-approval for those automated systems (Guiting & Striekwold, 2021). Indicators used to measure traffic safety at vehicle levels could potentially be translated to measuring traffic safety at network level based on vehicle sensor data.

Ex-ante safety evaluation

In the EU, all new types of vehicles and individual sub-systems need to be approved by a designated approval authority before it can enter the EU market. This is done to ensure harmonisation among EU member states and to ensure requirements with regards to safety and sustainability are met which is specified in Regulation (EU) 2018/858 (2018) (Guiting & Striekwold, 2021).

Every EU member state has an approval authority which in the Netherlands is the RDW. As can be seen in figure 11, a manufacturer sends its new product to an accredited technical service which tests the product on the aforementioned technical requirements. The approval authority assesses the test report and gives a type-approval to the manufacturer which is then allowed to produce this vehicle or sub-system for the EU market. The approval authority regularly checks the production to confirm it is the same as the tested product. (Guiting & Striekwold, 2021)

For already existing technologies and functionalities, this makes the safety evaluation fairly straightforward. However, the technical requirements as specified in the regulations are functional requirements aimed at mechanical components (Onderzoeksraad voor Veiligheid, 2019). This is problematic for most ADAS as these are new technologies for which such requirements do often not exist (yet). For example, the first Lane Keeping Systems (LKA) entered the market in 2014 but only in 2018 specific requirements were formulated in UNECE[3] regulations and adopted by the EU (Onderzoeksraad voor Veiligheid, 2019).

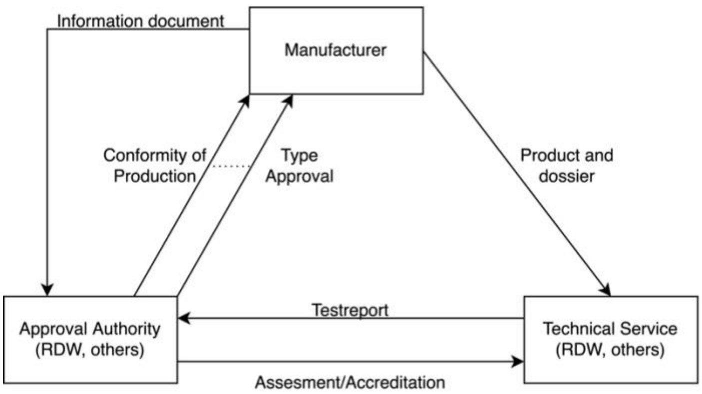


Figure 11 Type approval process within the EU (adapted from Guiting & Striekwold, 2021)

That these systems still got type approval has to do with the fact that these were often admitted through a specific article aimed at stimulating innovation (article 20 in the now obsolete Directive (EU) 2007/46/EG and article 39 in the replacement Regulation (EU) 2018/858) (Onderzoeksraad voor Veiligheid, 2019). In both regulations, it is required that these new technologies are *at least* at the same level of safety as other vehicles (Directive (EU) 2007/46/EG, 2007; Regulation (EU) 2018/858,

---

[3] The UNECE is an organization aimed at pan-European integration in which in specific working groups technical requirements are negotiated. Countries with major vehicle manufacturers like the USA and Japan are also represented in these working groups. (UNECE, 2020)

2018). But because any sub-system admitted through these articles is by definition an exception, no standard safety evaluation method exits.
According to the Dutch Safety Board (2019), it is often stated that it is not possible to demonstrate the effect of this specific ADAS on traffic safety. Since it can also not be proven that this specific ADAS negatively impacts traffic safety it is said to have met the requirement of being at least as safe.
This might be changing though. The first regulatory step for automated driving systems in which the vehicle is in primary control of itself in both the longitudinal and lateral direction has entered into force in January 2021; UN Regulation No 157 [2021/389] (2021). This regulation lays out safety requirements for Automated Lane Keeping Systems (ALKS), as well as requirements regarding the handover from the ALKS to the driver and on Human-Machine Interface (HMI). Examples of specific safety requirements include minimum following distances at different speeds and thresholds for TTC before the vehicle has to avoid a collision with a vehicle cutting into its lane (UN Regulation No 157 [2021/389], 2021).
This makes clear that in ex-ante safety evaluation currently no indicators to measure traffic safety are being used that could potentially translate to measuring traffic safety at network level based on vehicle sensor data.

Ex-post safety evaluation and comparing vehicles with driving automation systems to HDVs
Ideally, it should also be possible for the Dutch government to continually assess the effects of ADAS on traffic safety after they have entered the market. This is however currently not possible because it is not registered which ADAS are present in a vehicle. This is however currently being investigated whether or not this should become a part of the existing vehicle registration system and if this would be technically feasible. (Onderzoeksraad voor Veiligheid, 2019)
And even if this registration would be present, current indicators such as fatalities and severe injuries have their problems associated with accident registrations as discussed in section 2.2. Knowing just whether a vehicle is equipped with a specific ADAS does not mean it is used or if it is used correctly. Additionally, new problems arise when you want to compare vehicles equipped with ADAS to Human Driven Vehicles (HDV).

First of all, vehicles with driving automation systems and HDVs have different capabilities. Disengagement as discussed in the previous section, could possibly be used to say something about the (change) of safety of an vehicles with driving automation systems, but it cannot be compared to HDVs because these cannot disengage. It would even be difficult to compare vehicles with driving automation systems from different manufacturers because they use different systems and report disengagement in a different way (Fraade-Blanar et al., 2018). So, the availability of data on vehicles with automation could be a problem.
The second problem arises because this data is made available by the vehicle manufacturer but without a uniform definition and method of measuring such an indicator, the gameability could become an issue (Blumenthal et al., 2020).
Thirdly, even when there is solid data on the performance of vehicles with automation on specific indicators, this does not mean this data is available for HDVs (Blumenthal et al., 2020). While vehicles with driving automation systems might have the sensors to measure, detect and store/sent vehicle data, older vehicles do not. Additionally, even if a representative sample of conventional vehicles would be retrofitted with measuring equipment, the question could be raised if it is fair to compare these new vehicles with driving automation systems to those older conventional vehicles, given that the average lifespan of Dutch vehicles is about 20 years (MuConsult, 2021).
Fourthly, even if both sufficient data for HDV and vehicles with automation is available, a fair comparison would only be possible if the events and exposure measured could be matched with the ODD of the vehicles with automation (Blumenthal et al., 2020). Automation in vehicles have specific functions and are designed to operate in specific conditions. For example, in tests on public roads, Waymo let their vehicles drive fully autonomous on roadways with speed limits up to and including 45 miles per hour (Schwall et al., 2020). Comparing the recorded accident rates of the Waymo vehicles with driving automation systems to HDV accident rates would be fair if the HDV accident rates on these

types of roads be known. However, at these lower speeds, property-damage-only accidents are frequent these are severely underreported in the USA (Schwall et al., 2020)
Lastly, vehicles with driving automation systems are fundamentally different to conventional vehicles because the former contain large amounts of software that can and are upgraded and updated during their lifespan. This means that vehicles should become safer, which could mean that comparisons even between two vehicles from the same make and model are not 100% fair (Onderzoeksraad voor Veiligheid, 2019).
Ex-post safety evaluation is currently not conducted so no indicators to measure traffic safety are being used that could potentially translate to measuring traffic safety at network level based on vehicle sensor data. Additionally, comparing the safety performance of HDVs and vehicles with automation is currently highly difficult and may in the best case be unfair.

To conclude this section, table 3 gives an overview of indicators used by the Dutch government to measure traffic safety.

Table 3 Overview of indicators used by the Dutch government

Category	Specific examples	Strengths	Weaknesses	Current usage
Lagging indicators	Fatalities, severe injury	1 Easy to understand [a] 2 Uniform definition [a] 3 Historical data available [ab] 4 High validity [a]	1 Reactive approach [cd] 2 Rarity of occurrence [d] 3 Does not give insight into process [ce] 4 Subject to random fluctuations [ce] 5 Data incompleteness [b]	network level [f]
Dutch SPIs	Safe participants, safe speeds	1 Help to give insight [g]	1 Still being implemented [g] 2 Little data available [g]	Still in development, aims for network level at national, regional, and local scale [g]

a Blumenthal et al. (2020)
b Stipdonk (2013)
c Chang et al. (2017)
d Tarko (2012)
e ETSC (2001)
f SWOV (2020c)
g Aarts (2018)

2.3 Indicators used for measuring traffic safety from a scientific perspective

Next to the government, the academic community is involved in measuring traffic safety. And while the previous section shows the government currently does not yet have conclusive way to evaluate traffic safety of vehicles with automation, much research into this is conducted in the academic community, as will be shown in this section.

The effect of vehicles with driving automation systems on safety is determined in different ways by different researchers. These differences in assessment methods are one of the reasons why the found safety effectiveness can vary considerably. Fraade-Blanar et al. (2018) argues that there are four different settings in which testing of vehicles with driving automation systems can take place (table 4). In artificial settings, researchers can explore and test the behaviour of vehicles with driving automation systems without (in simulation) or with minimal (closed course) risk. In both settings, there is no risk to the general public, which is present with public road testing.

Table 4 Classification of safety evaluation methods for vehicles with driving automation systems (adapted form Fraade-Blanar et al., 2018)

Setting		**Safety considerations**
Public roads	With safety driver	Risk to professional driver and to other road users
	Without safety driver	Risk to driver and to other road users
Artificial setting	Closed course	Risk only to professional driver
	Simulation	No risk

Appendix B.1 goes more in depth on what safety evaluation methods exist and are proposed in the research community, but this only describes how something is measured, not what is measured. In those safety evaluation methods, safety is operationalised in indicators. The specific indicators used vary per method and per experiment, depending on the scenarios and goals of the experiment.

Crashes are sometimes used as indicator, for example in Road test data analysis. However, as discussed in the introduction and in section 2.2.1, accidents are rare. Therefore, driving test studies often use traffic conflicts instead, as these are much more common. Gettman et al. (2008) for example found a ratio of traffic conflicts to actual accidents of approximatly 20.000 to 1. Simulation studies also often use surrogate measures of safety (SMoS). For Connected and Automated Vehicles (CAVs), microsimulation using SMoS to estimate traffic safety is even the most common safety evaluation method (C. Wang et al., 2021). SMoS are meant as an alternative/complement to using injuries and fatalities (Johnsson et al., 2021; Laureshyn et al., 2016).

2.3.1 Surrogate Measures of Safety (SMoS)

The basis of SMoS are traffic conflicts, which as a concept was coined already in 1964 by Klebelsberg. Back then, it simply meant a dangerous traffic interaction (Arun, Haque, Bhaskar, et al., 2021). Since then, many more definitions have been proposed yet, similarly to the definition of traffic safety, no single definition exists as the used definition often depends on the goal (Arun, Haque, Bhaskar, et al., 2021). An often-used definition is the definition by Amundsen & Hyden (1977): “traffic conflicts are an observable situation in which two or more road users approach each other in space and time to such an extent that there is a risk of collision if their movements remain unchanged” (in Arun, Haque, Bhaskar, et al., 2021, p4).

Observing SMoS
Originally, traffic conflicts were observed and recorded by trained human observes. Nowadays, this can be done much more accuratly on three different levels according to (Arun, Haque, Bhaskar, et al., 2021): on road user level, on facility level and with microsimulation.
Road user level observations happen from the point of view of the road user. This can be done in naturalistic driving studies in which a vehicle is outfitted with measuring equipment or in a driver simulator as discussed previously. Nowadays however, it could also be done with equipment already present in the vehicle. This method provides the most detailed dataset but is usually also the most costly and time consuming, especially when the time and energy needed to process the raw data is taken into account (Arun, Haque, Bhaskar, et al., 2021).
Facility level observation can be done much quicker because it focusses on specific locations, like (a specific type of) intersections. The obeservation can be done with sensors already in place (e.g. loop detectors or Intelligent Transport Systems (ITS)) or with temporarily placed equipment like cameras or LiDAR (Arun, Haque, Bhaskar, et al., 2021).
Microsimulation can be used for both networkanalyisas and for specifc locations and has as main benefits that the research has full control over the experimental variables and that new, not widely available, technologies can be tested. This method however also has some major disadvantages: the quality of the simulation heaviliy depends on the underlying assumptions and behavourial models, the validity is often questioned and perhaps most imporant: there is often a significant difference in distribution of observed and simulated crashes (Arun, Haque, Bhaskar, et al., 2021).

Thresholds and validity of SMoS
So, SMoS like TTC or PET can by observed by on three levels with different observation methods. In order to draw conclusions about safety, one last element is needed: the threshold. By passing the threshold a normal traffic interaction becomes a traffic conflict or near-miss (Johnsson et al., 2021).
As can be seen in figure 12, a stricter threshold is closer to the actual crash and thus theoretically better if the goal is to estimate the expected total number of crashes. However, events with a stricter threshold are rarer, leading to longer observation periods. This is one of the problems SMoS were developed to overcome in the first place. With too loose thresholds on the other hand, crash risk is not measured, but exposure (Johnsson et al., 2021). Laureshyn et al. (2016) conducted a literature review and found a wide range of thresholds used for both TTC and PET. They found that the most common threshold for TTC is between 1,5s and 3s while for PET it is between 1s and 3s.
Therefore, choosing the right threshold is highly important for the validity of research but is often chosen quite arbitrarily (Arun, Haque, Bhaskar, et al., 2021). The validity of SMoS depends on the relationship between the traffic conflict and the actual crashes. In older studies, no or only a statistically weak relationship was found but with advancing observation and processing technologies, strong correlations have been found (Mahmud et al., 2017). This is confirmed in a literature review by Johnsson et al. (2018) although it is noted that it is difficult to generalise the validity of different indicators because of the variety of methods used. Therefore, Johnsson et al. (2021) suggests using a relative approach to validity in which it is accepted that SMoS may not accurately reflect the number of accidents but that they can still be used to reflect a change in safety (for better or for worse).

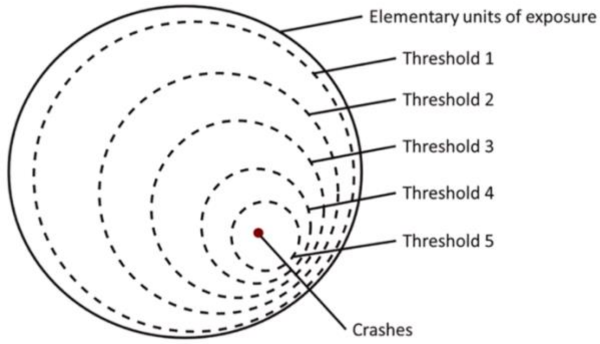


Figure 12 The theoretical relationship between elementary units of exposure, SMoS at different threshold values, and crashes (Johnsson et al., 2021)

Choosing an adequate threshold for an SMoS does pose several challenges. As mentioned before, there is no scientific consensus on thresholds for specific SMoS. A threshold also forms a hard border between safe and unsafe, meaning that for example with a TTC with a threshold of 2s means that a traffic interaction with a TTC of 2,01s is considered safe while 1,99s is not. It can be argued that in reality safety is more a continuum instead (Souman et al., 2021). Additionally, Souman et al. (2021) argue that using a single threshold regardless of the external circumstances like weather and the quality of the road is also problematic.
Lastly, specific SMoS are developed to assess safety in a specific context, like TTC for vehicles moving in the same direction and PET for intersections. This means that in order to assess safety in general, a combination of several indicators must be used. This will result in added complexity.

2.3.1.1 Proximity based SMoS

Because SMoS are much more common than crashes, they have advantage that they can be directly observed in traffic and that it takes considerable less time to assess changes in traffic safety (Johnsson et al., 2021). There is a large variety of SMoS, as is shown by Mahmud et al. (2017) who have identified 38 SMoS in their literature review. However, two SMoS are used significantly more than any other. Both in the literature reviews of Johnsson et al. (2021) and Laureshyn et al (2016), time-to-collision (TTC) is found to be the most used by a margin, followed by post encroachment time (PET).
TTC is the time remaining until an accident occurs if two vehicles maintain their current course and speed (Mahmud et al., 2017). TTC is mainly used for rear-end type of crashes or for hitting pedestrians or objects like a parked vehicle (Mahmud et al., 2017).
For measuring of modelling traffic safety on inttersections, Post Encroachment Time (PET) is more suitable (Mahmud et al., 2017). PET is the time between one vehicle leaving a certain point or area and the arrival of a second vehicle at that point or area (Arun, Haque, Bhaskar, et al., 2021).
Both are of these are called proximity based SMoS as these are based on proximity in time. Other families of SMoS exists, such as those based on spatial proximity, kinematic based SMoS (based on deceleration or acceleration) and those who do not fall in any of the other categories to suit very specific cases (Arun, Haque, Bhaskar, et al., 2021; Mahmud et al., 2017). According to Arun et al. (2021), kinematic based SMoS are often prefered in naturalistic driving studies and are therefore included in this analyisis.

2.3.1.2 Kinematic based SMoS

In urban areas, research has found that the most common evasive action to avoid collision is deceleration (Johnsson et al., 2018). Therefore, (strong) deceleration could indicate a potentially dangerous situation. This could be defined in a relatively simple way in which each event in which the strength of the deceleration exceeds a threshold is counted (Arun, Haque, Bhaskar, et al., 2021). More elaborate indicators have also been developed as described by Mahmud et al. (2017). They describe the Deceleration Rate to Avoid the Crash (DRAC) which also considers vehicle in front.
The main advantage of kinematic based SMoS is that they are easy to understand, objective and physics-based and that they can be used in several different situations (Blumenthal et al., 2020; Mahmud et al., 2017). It is therefore used in naturalistic driving studies (Arun, Haque, Bhaskar, et al., 2021).
Next to deceleration, acceleration could be used in a similar way, as well as evasive action in the form of swerving (Johnsson et al., 2018). Johnsson et al. (2018) does however report that no validation studies have been carried out into deceleration-based indicators.

2.3.2 Disengagement of the automation in the vehicle

Disengagement of the automation in the vehicle happens when the driver retakes full control or when the automation fails (Fraade-Blanar et al., 2018). This could be a proxy for unsafe situations. The use of disengagement has been used as an indicator for safety early on but has since fallen out of use (Blumenthal et al., 2020; Fraade-Blanar et al., 2018). The main reason for this is how sensitive disengagement is to the circumstances like the environment and the driver (Blumenthal et al., 2020). This is also shown in Schwall et al. (2020) in which the situation after disengagement of Waymo vehicles in real world situations were simulated. In more than 99,9%, no simulated contact occurred. Additionally, when it is used in testing by safety drivers to proof how safe a vehicle is (as is done in the past by Tesla in California), it is also highly gameable (Blumenthal et al., 2020).

2.3.3 Engagement of the automation in the vehicle

Similar to how disengagement of automation in the vehicle could be a proxy for unsafe situations, engagement of automation in the vehicle could be used as well. ADAS such as Forward Collision Warning (FCW) and Automatic Emergency Braking (AEB) are safety system that warn the driver or intervene when the system detects that an accident is imminent (Onderzoeksraad voor Veiligheid, 2019). No discussion of using engagement of these systems as a measure for safety was found in academic literature but pilots like the Road Monitor (RoMo) and Kia Insights from vehicle data (see section 2.7) do report using it.

2.3.4 Infractions

For human drivers, there is a statistically significant relationship between infractions and crashes, and it can be assumed that this is similar for vehicles with driving automation systems, at least in this early stage (Blumenthal et al., 2020). Infractions like speeding could be measured by the vehicle itself. However, context matters here as well as sometimes an illegal move could be the safest course of action (Blumenthal et al., 2020). And while there is a statistically significant correlation between infractions and crashes, it is not well-understood or very strong (Blumenthal et al., 2020).

2.3.5 Roadmanship

Roadmanship is a concept that aims to measure whether or not a vehicle drives safely and responds well to the unsafe driving of others. It is a concept still in development and no single best way to define and measure it has emerged (Blumenthal et al., 2020). It can be measured in by for example SMoS like TTC or PET, but no single metric can yet cover all of its aspects (Blumenthal et al., 2020). So often, more holistic measures like the safety envelope concept are proposed. In this concept, of which the Responsibility Sensitive Safety (RSS) model is an example, an envelope is defined around the vehicle which is the minimum safe distance to other vehicles around it, depending on the environment and speed of the vehicle (Fraade-Blanar et al., 2018). Violation of this envelope then indicates an unsafe situation. This method is however very data intensive and not yet clearly defined (Fraade-Blanar et al., 2018). Wishart et al. (2020) have made a more comprehensive list of metrics for this and other similar metrics for the safety evaluation of vehicles equipped with automation.

For these indicators, exposure needs to be into account, as it is done in the SPIs used by the Dutch government. A common way to do that is by measuring, for example, disengagements per vehicle miles travelled (VMT) (Fraade-Blanar et al., 2018). Using disengagement per million VMT is done by for example Matysiak & Razin (2018).

To conclude this first part of the literature review which focussed on measuring traffic safety and indicators for doing so, table 5 shows an overview of indicators used in to evaluate safety in the academic world. Together with table 3 (Overview of indicators used by the Dutch government) this gives an answer to the first research question: How is traffic safety measured currently?

Table 5 Overview of indicators used in the academic world

Category	Specific examples	Strengths	Weaknesses	Current usage
Proximity based SMoS	TTC, PET	1 Directly observable in traffic [h] 2 Objective and physics-based [a]	1 Data intensive 2 Specific for situation [i] 3 Validity of threshold [hij]	(simulation) experiments [hi]
Kinematic SMoS	Deceleration, acceleration, swerving	1 Easy to understand [a] 2 Objective and physics-based [a] 3 Suitable for several situations [i]	1 Validity of threshold [hij]	Naturalistic driving studies [j]
Disengagement of ADAS	ACC, LKA	1 Easy to measure [a]	1 Sensitive to context [ak] 2 Low validity [a] 3 Gameable [a]	Field tests [k]
Engagement of ADAS	BSW, ACC, LKA, FCW	1 Easy to measure [l]	?	Public-private pilots [mn]
Infractions	Speeding	1 Statistically significant relationship with crashes [a] 2 Could be compared to HDV [a]	1 Context dependent [a] 2 Relationship not well understood or strong [a]	?
Holistic Roadmanship measures	Safety envelop violation	1 Objective and physics-based [a]	1 No uniform definition of roadmanship [ao] 2 Data intensive [aop]	Still in development [ao]

a Blumenthal et al. (2020)
h Johnsson et al. (2021)
i Mahmud et al. (2017)
j Arun et al. (2021)
k Schwall et al. (2020)
l Presumed to be similar to disengagement of ADAS
m Interview with Vrijens, ministry of I&W, Appendix C
n Kia Nederland (2021)
o Fraade-Blanar et al. (2018)
p Wishart et al. (2020)

2.4 Vehicle sensor data

The previous sections discussed how traffic safety in general can be measured and what types of indicators are used. The data necessary for measuring these indicators varies. As is highlighted in the introduction, new vehicles are becoming increasingly smarter, connected, and are equipped with more sensors which has led to the idea that vehicles can collect the necessary data for measuring traffic safety at network level. The next several sections are aimed at answering the second research question: What are possible scenarios for using vehicle sensor data to measure traffic safety in practice? It first discusses existing types of vehicle data that are used to various ends to see how these are accessed. It then looks specifically at vehicle sensor data and how these can be shared both in theory and in practice by discussing pilots doing just that.

2.4.1 Regulated types of vehicle data

Vehicles collect a variety of data to various ends. Table 6 shows the availability of three types of vehicle data which regulated by various EU regulations, SRTI, RMI and accident data. These are regulated because they serves a societal goal like improving traffic safety or fair market competition (Ecorys, 2020). All three types of data are based on vehicle data but have a different aim and a different target group. RMI data is not (directly) meant for drivers but for mechanics, either from official dealers or independent (Ecorys, 2020). eCall can warn emergency services after a crash has taken place (Regulation (EU) 2015/758, 2015) while SRTI is aimed at warning drivers at a specific location for eight different dangerous situations (Henkens et al., 2020).
The availability of SRTI messages is regulated in Regulation (EU) No 886/2013 (2013) although there is a discussion on how wide this regulation should be interpreted (Henkens et al., 2020). According to this regulation, SRTI messages should be made available to a National Access Point (NAP) in a specified format (DATEX II).

Each EU member state was required to set up an NAP by Regulation (EU) 2015/962 (2014) which can be used to access, exchange and re-use both static and dynamic road status data and traffic data.

Table 6 Regulated vehicle data accessible (Ecorys, 2020)

Datatype	Explanation	Examples	Access
Repair and Maintenance Information (RMI)	RMI are technical data used for repair and maintenance	a. Diagnostics b. Fault codes	Accessed via OBD portal, supplemented by information from OEM
Accident data (via eCall)	System calls 112 when an accident is detected (automatic or manually)	Minimum Set of Data (MSD): a. Time of activation b. Vehicle Identification Number (VIN) c. Number of Passenger d. Location and driving direction e. Type of vehicle f. Propulsion type	Via public mobile wireless communications network
Safety Related Traffic Information (SRTI)	SRTI are warning messages provided in-car, based on vehicle data	a. Temporary slippery road b. Animal, people, obstacles, debris on the road c. Unprotected accident area d. Short-term road works e. Reduced visibility f. Wrong-way driver g. Unmanaged blockage of a road h. Exceptional weather conditions	From vehicle to National Access Point (NAP), back to vehicles near the relevant location; see figure 14, p33

Next to these three regulated data types are two unregulated data types: the data that is intellectual property of the vehicle manufactures and the 'other' data that has commercial value. The former consists of data necessary to operate the vehicle like firmware while the latter is all the other data collected by the vehicle that could, after processing, have commercial value (Ecorys, 2020).

Table 7 Unregulated vehicle data (Ecorys, 2020)

Datatype	Explanation	Examples
Intellectual property VM	Data necessary to operate the vehicle and its systems	a. Software updates b. Firmware
'Other'	The remainder of the data that may have commercial value	a. Data on car usage b. Data on interface usage c. Data on driving behaviour

It is also possible to retrieve this vehicle sensor data without the help of vehicle manufacturers, although it is y discusses three ways: with a CAN-bus reader, with an OBD dongle or with Stand-alone kits and smart phones. Via the Controller Area Network (CAN)-bus all electronic signals are sent through the system, meaning that (almost) all vehicle sensor data passes through the CAN-bus. This data is encrypted but can (partially) be deciphered with aftermarket devices for access near real-time data. This is typically done by large fleet owners.
The OBD portal is used to acquire RMI information in garages. With an OBD dongle, all information from the OBD portal can also be acquired while driving.
Stand-alone kits or smartphones can also collect data through external sensors. This technically is not vehicle sensor data because the data is not collected by the vehicles themselves. It therefore is also more limited in the information it can provide. (Ecorys, 2020)

All three options are cumbersome and not very suited for large scale data-collection as they require aftermarket devices. This means that any effort for large scale data-collection would require cooperation with OEMs.

2.4.2 Vehicle sensor data

Modern vehicles can be equipped with a wide range of driving automation systems. In 2022, a variety of ADAS will become mandatory in new type of vehicles and in 2024 in all newly produced vehicles, as is discussed in the introduction and in 3.1.3 (Scope). These ADAS include amongst others: Intelligent Speed Assistance (ISA), Advanced Emergency Breaking (AEB), Emergency lane-keeping system (ELKA), Driver drowsiness and attention warning and Advanced driver distraction warning (table 1, p11). These ADAS perform a variety of tasks and need a variety of sensors to achieve these tasks. Figure 13 shows which sensors are generally used to perceive the environment, although the exact usage may differ among different vehicle manufacturers (Ackermann et al., 2019).

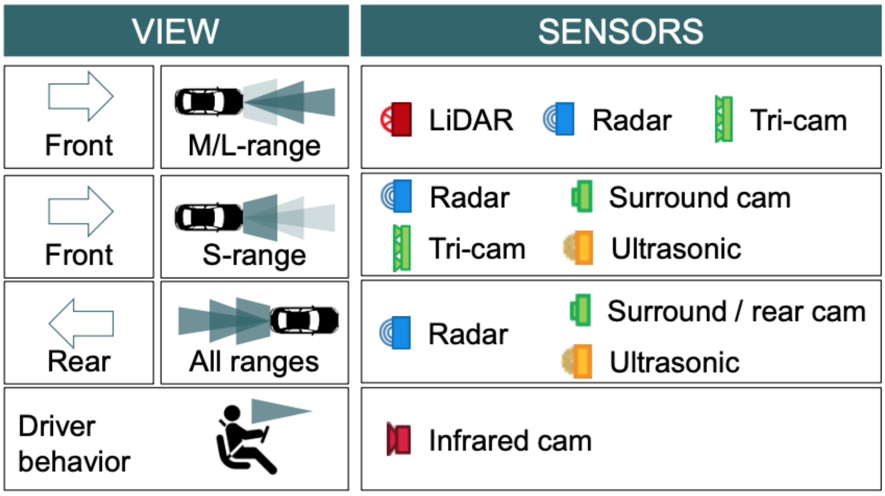


Figure 13 Overview of sensors (adapted from Ackermann et al., 2019).

Data from these sensors and ADAS can have significant commercial value to car manufacturers and related actors. McKinsey (2018) expect services around vehicle sensor data to become increasingly important for vehicle manufacturers while at the same time their conventional markets of selling and maintaining vehicles will be put under pressure. They predict that the market around mobility data could be worth between 450 and 750 billion USD by 2030 through three main value creation models: generating revenue directly, reducing costs, and increasing safety and security. Generating revenue directly could be done by selling additional services to the customers (the drivers), by using the vehicle sensor data to tailor advertising (for both vehicle maintenance and for non-vehicle related promotions), and by selling the vehicle sensor data to third parties for any purpose (McKinsey, 2018). In many cases permission from the customer is needed to share their data. Therefore, McKinsey (2018) sees communicating the added value to the customers as one of the key challenges in this market, although further research by them suggests that a significant portion of customers is willing to share their data related to navigation and mobility.

2.5 Models for sharing vehicle sensor data

In order to use vehicle sensor data, it needs to be collected, processed, and shared. The sharing of the data can be done through the built-in SIM cards that make vehicles connected (Ecorys, 2020). There are however several ways in which such a system of data sharing can be designed.
The Platform for the Deployment of Cooperative Intelligent Transport Systems in the European Union (C-ITS Platform) is a platform in which a large variety of stakeholders, including vehicle manufacturers and member states, worked on identifying success and failure factors for the deployment of C-ITS in

the EU (European Commission, n.d.-c). Working Group 6 (WG6) of this platform proposed three main technical solutions to allow access to in-vehicle data and resources (TRL, 2017):

- In-vehicle interface
- On-Board Application platform
- Data Server Platform

2.5.1 In-vehicle interface

The first solution proposed is already in the market. With an in-vehicle interface, data available through the OBD interface is accessed and transmitted through an external device connected to the OBD interface(TRL, 2017). The current generation OBD-II does have two disadvantages. Firstly, it is not secure enough and secondly, it cannot deliver data at a high enough bandwidth to support the transmitting of large amounts of real-time data. To achieve this, the current OBD-II needs to be replaced with the more advanced OBD+.

2.5.2 On-Board Application platform

With an On-Board Application platform data can be accessed through a dedicated platform integrated in the vehicle. This platform would allow the installation of apps on the HMI (Human Machine Interface) of the vehicle. This would allow third parties on one hand to directly access the CAN-network (Ecorys, 2020)and transmit this data to the party operating the application and on the other hand to send information back to interact with the driver (TRL, 2017). The main issues with this solution are again security and safety. Firstly, a large number of applications running on the vehicles internal systems could degrade the performance of this system as a whole. Secondly, it could be possible to make changes to the software of the vehicle through the applications, introducing concerns about liability and security (TRL, 2017). These risks could however be mediated by putting a system in place for testing and certifying applications (by either the vehicle manufacturer or an independent actor) before allowing installation of said app. This concept is sometimes also referred to as Open Telematics Platform (OTP).

2.5.3 Data Server Platform

The first two solutions process data within the vehicle while the third solution does this outside of the vehicle. With a data server platform solution, data is sent directly to a separate server through the mobile network. Three models exist for this solution:

- Extended vehicle (ExVe)
- Shared server
- B2B marketplace

In the extended vehicle model (ExVe), vehicle sensor data is transmitted encrypted to dedicated servers of the vehicle manufacturer. The vehicle manufacturer can then make (processed) data available to third parties.

In the shared server model, the data is not sent to a server operated by the vehicle manufacturer but to a neutral server operated by a consortium of stakeholders. This model would also allow to have data from multiple vehicle manufacturers on one server. The vehicle manufacturers would then need to deliver their data in a standardized format.

The last model, B2B marketplace, is a combination of both models in which a neutral service provider would be able to access the servers of the vehicle manufacturers to forward data to a neutral server. On this server, third parties could gain access. It is likely that this neutral service provider would be a large big data or IT company like Google (TRL, 2017).

One of the main issues of these models is that the vehicle manufacturer controls which data can be accessed, unlike with the on-board application platform. It also does not allow for communication of third parties with the driver through the HMI resulting in concerns about fair market competition (TRL, 2017).

As a result of this, the Alliance for the Freedom of Car Repair (AFCAR)[4] opposes this solution and is in favour of the on-board application platform while the European Automobile Manufacturers' Association (ACEA) sees the ExVe model as the only solution due to security and safety concerns (Ecorys, 2020). In 2016, the ACEA together with the European Association of Automotive Suppliers (CLEPA) proposed a fourth model to meet the concerns of AFCAR: the ExVe/neutral server model.
In this model, the ExVe model is supplemented with a neutral server allowing the operator of the neutral server to better negotiate with the vehicle manufacturers and at the same time giving third parties the option to choose between going to the vehicle manufacturer directly or indirectly through the neutral server (TRL, 2017). In late 2017, CLEPA decided to no longer support this model due to concerns around fair market competition (Ecorys, 2020).

2.6 Pilots using vehicle sensor data

The collecting and sharing of vehicle sensor data as described in the previous sections is not just theory, it is already being put in practice. This section will show several examples of projects in which this is done to show the development throughout the years and to show different scopes of projects, in amongst others purpose, size, and geographical scope.

Praktijkproef Amsterdam (PPA)

One of the largest and oldest Smart Mobility experiments in the Netherlands in the past several years (2012-2021) has been the Praktijkproef Amsterdam (PPA) (Groenendijk, 2021). The goal of this project was to optimise mobility in urban regions by letting in-car and roadside systems communicate and work together (Groenendijk, 2021). This project was not necessarily aimed at improving traffic safety. However, the various pilots done in the PPA did lay a foundation for further projects and delivered many important lessons on both technical aspects of such systems and, perhaps even more important, on working together in complex public-private partnerships (Groenendijk, 2021). An example of this is that the experiences of the PPA on cooperating between governments, knowledge institutes and companies (car manufacturers, data- and service providers and other suppliers of smart technologies) have been developed into a framework in SOCRATES$^{2.0}$ (Groenendijk, 2021).

SOCRATES$^{2.0}$

SOCRATES$^{2.0}$ was an EU wide project which third and final phase of the PPA was part of. This project ran from 2018 until 2021 with most of the testing being done in 2020 (Groenendijk et al., 2021). SOCRATES$^{2.0}$ demonstrated that it is possible for public and a variety of private parties to work together on the topic of traffic management and data exchange on five different use cases[5] across different EU member states (Groenendijk et al., 2021). Partners in the SOCRATES$^{2.0}$ project have been somewhat surprised by: "the openness in the international setting and the willingness to experiment with working together in a new way (within the boundaries of the project)" (Groenendijk et al., 2021, p31).

Talking Traffic

Talking Traffic is a corporation in the Netherlands between (local) governments and 20 (inter-) national companies aimed at improving traffic flow, traffic safety and sustainability (Bourdeaud'hui et al., 2020). In two use cases - in vehicle signage and speed limit advice, and potentially dangerous situations and road works – vehicle sensor data was used to research the effects of the use cases. Both used individual GPS traces acquired in other parts of the Talking Traffic value chain (Bourdeaud'hui et al., 2020). For one month (November 2019), GPS traces were logged every 4 seconds for about 15% of all vehicles in the Netherlands resulting in 600 million datapoints per day (Bourdeaud'hui et al., 2020).

[4] A large coalition of repair and maintenance companies, insurance companies and consumer organisations
[5] network optimisation, smart destination, environmental zone information, lane information and road works

Proof of Concept Data for Road Safety

While the sharing of vehicle data for road safety applications has been regulated since 2013, as mentioned in section 2.4.1, it was not until 2017 that plans were developed to actually make this happen (Henkens et al., 2020). A Proof on Concept (PoC) for generating and sharing Safety Related Traffic Information (SRTI) then ran from June 2019 until October 2020 in which several EU members states worked together with service providers (SVP) like TomTom Traffic and OEMs such as BMW (Henkens et al., 2020). Figure 14 shows the process of generating and sharing the SRTI messages.

- It starts when a certain trigger condition is met based on input of vehicle sensors like wheel speeds, steering deflection, wipers, and others.
- A data package of a single vehicle is sent to the access/aggregation point by the OEM or SVP. This can be the National Access Point (NAP) that every EU member state should have.
- Then, this data is validated with data from other vehicles and/or other OEMs and SVPs.
- This then forms SRTI message which is sent back to the end user by an SVP.

The sharing of the data between the parties is done through a Data Server Platform, specifically with an extended vehicle (ExVe) concept (Data for Road Safety, 2021) as explained in section 2.5.3.

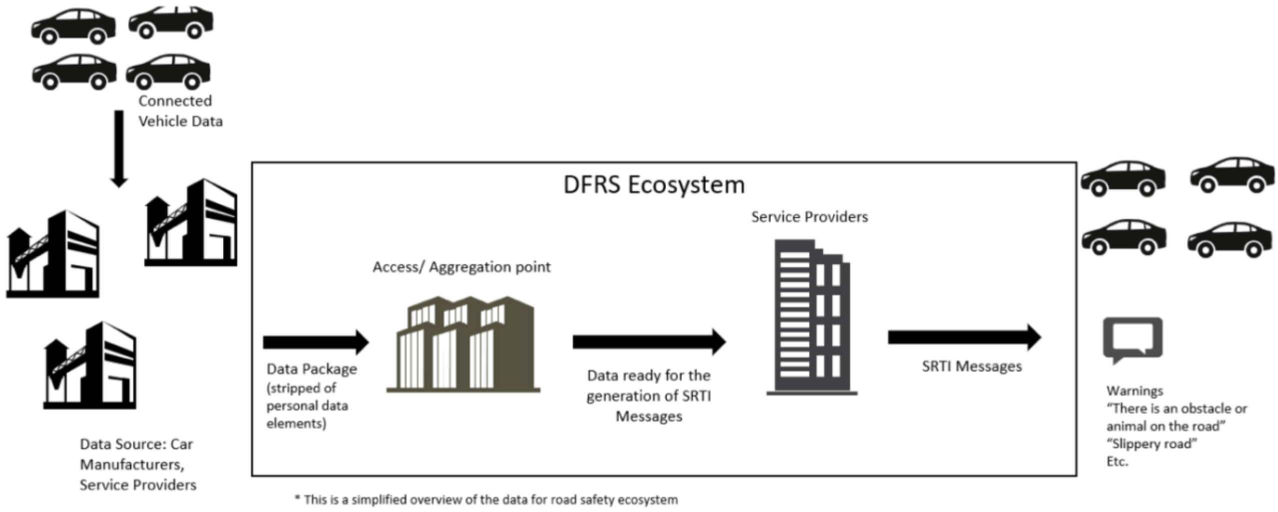


Figure 14 Simplified overview of the process of the PoC Data for Road Safety (Data for Road Safety, 2021)

Road Monitor (RoMo)

The Road Monitor (RoMo) is a new project of the Dutch Ministry of Infrastructure and Water management in which vehicle sensor data is processed into useful information for the road authorities. One of the three focus areas is detecting unsafe situations. The data comes from sensors present in the vehicles. This can be the input of drivers like sudden braking or abrupt steering wheel motions, or it can be the intervention of specific ADAS like Forward Collision Warning (FCW). The collection and processing of the data is done by one OEM and the information is then delivered to the ministry of I&W in the form of an interactive dashboard, which helps to guarantee the privacy of the drivers. This dashboard can then be used to identify dangerous locations and also provide context on why these locations are dangerous. (Interview with Vrijens, ministry of I&W, Appendix C)

Kia Insights from vehicle data

in 2021, Kia started a project with the ministry of I&W to explore the potential added value of vehicle data for various public and/or private purposes (Kia Nederland, 2021). For three months, vehicle sensor data was collected from 4000 Kia vehicles to explore insights in charging behaviour and in safety. To gain insights into traffic safety, the deployment and status together with time and location of four ADAS (Blind spot detection, LKA, ACC, FCW) were tracked throughout these months, resulting in 41 million observations (Kia, 2021). In this project, like in RoMo, no data was shared with the ministry of I&W but only the insights gained from this data to ensure that the government cannot trace the data back to individual users (Kia Nederland, 2021).

2.7 Potential barriers for using vehicle sensor data

The projects and pilots have provided opportunities to explore the potential of using vehicle sensor data to various ends. In doing so, issues have been raised and lessons have been learned. This section will discuss several of the most mentioned lessons and issues.

2.7.1 Technical feasibility

Collecting, sharing, and processing vehicle sensor data is a technically complex operation. However, as the PoC Data for Road Safety and the Kia project show, it is possible to extract data from sensors in the vehicle and to process this data outside of the vehicle. Henkens et al. (2020) conclude in the PoC Data for Road Safety that the supply of data is constant, stable, and timely. Additionally, this project shows that it is technically possible to share data between several parties with the ExVe concept. Furthermore, TRL (2017) finds that the other possible solutions (In-vehicle Interface and On-board Application Platform) are also technically feasible.
It is important to note that this is all done on a pilot level, which are limited in terms of scope. Kia (2021) concludes that collecting and storing data is costly, meaning that it is important to be selective in collecting only the data that is needed to gain the wanted insights. In order to scale up Henkens et al. (2020) conclude that additional actions are necessary to scale up the system of SRTI messages because of large differences in development among partners.

Both the PoC Data for Road Safety and the SOCRATES$^{2.0}$ project stress the importance of using standards for the exchange of data (Groenendijk et al., 2021; Henkens et al., 2020). SENSORIS and DATEX-II are such standards which are widely adopted in Europe although in the SOCRATES$^{2.0}$ project it is concluded that the standard “is not as mature as often thought“ (Groenendijk et al., 2021, p28) and that [about DATEX-II] “there is a lot of own interpretation and therefore ambiguity and miscommunication possible” (Groenendijk et al., 2021, p8).

2.7.2 Legal feasibility

There are two main legal issues with using vehicle data for various purposes as described in the examples: the privacy of the drivers and fair market competition

Privacy

Most legal issues in this case have to do with privacy. Under the GDPR, data that can be traced back to an individual is personal data and as a result needs to adhere to stricter rules (Ecorys, 2020). There is an ongoing debate whether or not vehicle sensor data is personal data in which various stakeholders and even the supervisory authorities on privacy of different EU member states disagree (Ecorys, 2020). In all examples given in 6.4 where vehicle sensor data is collected (PoC Data for Road Safety, Road Monitor, Kia project) the owners of the vehicles have given permission to collect their data, an important aspect in the GDPR. Additionally, in the PoC Data for Road Safety all parties are contractually bound make sure no personally identifiable data is exchanged within the ecosystem, by stripping off all personal data elements and only sharing the bare minimum (Data for Road Safety, 2021). This is not an issue in the Kia project, as no data is shared with the ministry of I&W (only insights), thus avoiding many privacy related issues (Kia Nederland, 2021).
But even with permission to collect data, problems arise. The owner gives permission to collect data, but the vehicle manufacturer decides to share which information with which party (Henkens et al., 2020). Additionally, this permission must also be able to be revoked, which in practice often is difficult (Ecorys, 2020). This becomes even more complex when the vehicle is bought second-hand or when the owner is not the user, for example in the case of a lease car or a shared vehicle (Ecorys, 2020).

Fair market competition

A second set of legal issues may arise around fair market competition, given the dominant position of the vehicle manufacturers. In the TRL research (2017) is concluded that existing laws on this topic should be strong enough to prevent unfair market competition but does explicitly state that the practical application of these laws is highly complex. (Ecorys, 2020) agrees, especially for the coming few years. They also point at the practical aspects because under these fair market competition laws it would need to be proven that a company has a dominant position and also abuses this position. They deem this unlikely, also given the fact that this is a new and quickly developing market.

2.7.3 Cybersecurity

Cybersecurity is one of the most discussed emerging risks in the transition towards more autonomous driving (Ryan, 2020). As vehicles are being equipped with more autonomous functions, both the amount of software[6] as the level of connectivity increases strongly (Onderzoeksraad voor Veiligheid, 2019). The latter is important because an increase in the number of external connections in a vehicle comes with an increase in the potential points of attack (Fraade-Blanar et al., 2018; Onderzoeksraad voor Veiligheid, 2019; Ryan, 2020). This does not just refer to V2V or V2I connections but also to Over-The-Air (OTA) updates.

The risks are also substantial: in recent years researchers and ethical hackers have managed to gain remote access to vehicles which allowed them to take over control of systems like the brakes, engine, and steering wheel (Onderzoeksraad voor Veiligheid, 2019; Ryan, 2020). Besides the direct threat to human life present in such a hacked vehicle, other risks are also present such as data breaches and ransomware attacks (Fraade-Blanar et al., 2018; Ryan, 2020).

The Dutch Safety Board did not find any accidents where cybersecurity may have played a part in but at the same time concludes that this does not mean that no such accidents have taken place. There is no active centralised monitoring of such cybersecurity breaches and often, not enough information is available to establish such a breach after the fact (Onderzoeksraad voor Veiligheid, 2019). Both OEMs and policymakers are taking steps to minimise the risk of cybersecurity attacks, but residual risk will remain (Ryan, 2020). How large this residual risk is, and how to deal with it, is currently not known.

2.7.4 Willingness of stakeholders

Any system that uses vehicle sensor data for any purpose will involve a variety of stakeholders. As discussed in section 2.4.2, a potentially large market exists around use cases involving vehicle sensor data. Who exactly will be involved depends on the scope that is applied but Ecorys (2020), McKinsey (2018), and TRL (2017) all discuss OEMs, suppliers, and service providers.

OEMs are an essential player as these control access to the human machine interface (HMI) and to data collected and produced by the vehicle itself (McKinsey, 2018). Many OEMs see the potential value of using vehicle sensor data for a variety of purposes, as is also shown by the fact that the ACEA (organisation for European OEMs) sees a new business in data and information (Henkens et al., 2020). However, it should not be underestimated how large and complex these OEMs are (McKinsey, 2018). For example, within an OEM conflicting interests may exist between for example the sales department wanting new features and the engineers focusing on quality and complexity reduction for the driver.

Suppliers are another important stakeholder as they deliver parts of the vehicles and may therefore also have control over the access to specific data that can provide them direct economic value or serve as a source of differentiation to competitors (McKinsey, 2018). They can monetize the data through the OEM or directly to their consumer (not necessarily B2C but also B2B), for example with predictive maintenance and failure diagnostic, or through feedback- based R&D optimization (Ecorys, 2020; McKinsey, 2018).

[6] A modern vehicle equipped with several ADAS will have anywhere from 5 to 20 times more lines of code than a f-35 (JSF) fighter jet (± 50-100 million lines of code) (Onderzoeksraad voor Veiligheid, 2019)

Service providers are playing an increasingly larger role in vehicles and the interaction with the drivers. Initially, the services were related to infotainment and navigation but a shift can be seen towards a larger variety of services and further in-vehicle integration (McKinsey, 2018). Examples of this are Apple Carplay and Android Auto which can connect a smartphone to the car's screen and system or HERE Europe and TomTom Traffic who supplied and/or used data, or served as an aggregator in the PoC Data for road safety (Henkens et al., 2020). The service providers thus have an important role in integrating and connecting the vehicle with other technology and the wider ecosystem (McKinsey, 2018).

2.7.5 From pilot to reality

Besides the issues on specific aspects as discussed above, a more general factor is also relevant to consider: the paradox of the pilot. Groenendijk (2021) calls the belief that a successful pilot can simply be scaled up and will lead to the same successful results naïve. Certain factors (like those in table 8 below) that contribute to a pilot's success can equally lead to failure in a scaled-up version.

Table 8 The paradox of the pilot (adapted from Groenendijk, 2021)

Factor	Success for pilot	Failure for scaled-up version
More space	Room to experiment	Distance to regular day-to-day business
Budget	Extra budget for a specific pilot	Not in the regular budget
Enthusiastic employees	Motivated people	Other employees do not feel responsible to use the results from the pilot

This second part of the literature review focussed on vehicle sensor data and collecting it and in doing so aimed to answer RQ2: What are possible scenarios for using vehicle sensor data to measure traffic safety in practice? This review has shown that vehicle manufacturers are currently already obliged and capable of collecting and transmitting data related to safety, accident data via eCall and the SRTI data. Several theoretical models for sharing data are discussed of which the ExVe/neutral server model is supported by the vehicle manufacturers, making this the most viable scenario. This model has been tested in successfully in pilots, although various barriers to using vehicle sensor data have been identified.

3. Methodology and operationalisation of the Delphi study

In the literature review, this research has identified a variety of categories of indicators that can be used to measure traffic safety, together with their respective strengths and weaknesses. Since no consensus has been reached in the scientific literature on which category of indicator would be the most suitable to supplement fatalities and severe injuries as indicator for traffic safety on a network level, this will be the first aim of the Delphi study and in doing so, answering RQ3: *What are feasible and suitable indicators based on vehicle sensor data to measure traffic safety at network level?* Additionally, the literature review discusses several potential barriers of using vehicle sensor data to measure traffic safety in practice as faced in various pilots. Different pilots faced different issues. To gain more insight into these barriers and to what extend these barriers exist, the assessment of experts will be sought as well. This will then be used to answer RQ4: *What barriers exist for the collection and usage of vehicle sensor data to measure traffic safety in practice by the Dutch Ministry of Infrastructure and Water Management?*

In a Delphi study, the opinion of experts is leveraged to improve the understanding of a problem, opportunity or solution when there is incomplete knowledge about this problem (Skulmoski et al., 2007). That is the case here. Several indicators have been identified but the question remains which, if any, could be suitable to supplement fatalities and severe injuries as indicator for traffic safety on a network level.

There are different types of Delphi studies and different ways to execute them, but it usually has the following four characteristics: anonymity, iteration, controlled feedback, and statistical group response (Fritschy & Spinler, 2019; von der Gracht, 2012). Experts are asked questions, separate and anonymous to prevent influencing each other. In subsequent rounds, the same (or similar questions) are asked again but this time with controlled feedback. This feedback can consist of statistics on the group response of the previous round and/or often used arguments. The experts are then given the chance to revaluate their answer, with the goal of reaching consensus or establishing clear dissensus. (Beiderbeck et al., 2021a; Belton et al., 2019; Hsu & Sandford, 2007)

Any Delphi study starts with defining the goal of the study which is the basis for the design and development of the study (Belton et al., 2019). For this study the goal will be twofold:

1. To assess which, if any, indicators based on vehicle sensor data are the most suitable for supplementing fatalities and severe injuries as indicator for assessing traffic safety on a network level.
2. To assess whether or not the large-scale collection and usage of vehicle sensor data for measuring traffic safety will be feasible in practice and what factors act as barriers.

This chapter will focus on the methodological design of the Delphi study and the operationalisation by consecutively discussing how experts are selected (3.1), the operationalisation of the questions in the Delphi survey (3.2) including how questions are asked and how the feedback to experts is provided, dealing with bias (3.3), as well as practical aspects of the Delphi survey (3.4).

3.1 Expert selection

A Delphi survey is fundamentally different from a regular survey because it does not have the goal to generalise results of a representative sample to a larger population, but instead to reach consensus among experts (Okoli & Pawlowski, 2004). Therefore, participants are not selected randomly but rather purposely and based on their expertise (Keeney et al., 2006). And while the Delphi method has proven to produce valid results in the past (Förster & von der Gracht, 2014; Landeta, 2006), the selection of appropriate experts is a highly important part in the process, as the quality of the experts directly relates to the quality of the results (Hsu & Sandford, 2007; Keeney et al., 2006).

However, as Keeney et al. (2006, p209) states: “there is no magic formula to help researchers decide on who are the experts and how many there should be.”

3.1.1 Expert identification methods

Most studies using Delphi study report on how they have selected experts, but fail to report why this selection is made (Devaney & Henchion, 2018; Mauksch et al., 2020). And while there is a general consensus that an expert is “someone who is skilful and well-informed in some special field” (Mauksch et al., 2020, p2) no single best way to define and measure expertise has been developed (Devaney & Henchion, 2018; Hasson et al., 2000; Mauksch et al., 2020). Not clearly defining what an expert constitutes for a specific study and how these experts are selected can cast doubt on the results of such a Delphi study (Donohoe & Needham, 2009).

Therefore, Mauksch et al. (2020) have constructed an overview of expert identification methods which can be seen in table 9 below. Of these eight methods, five are not suited for this study. In past performance experts are selected on their performance in earlier Delphi surveys. It is aimed at repeated studies and needs data on previous forecasting performance of the experts, which is not the case here. Knowledge tests and psychological traits both use tests to assess and subsequently select experts. Knowledge tests focusses on the amount of knowledge a potential expert has on a specific topic while psychological traits tests try to assess expertise based on various cognitive criteria (inner consistency, discrimination ability, the drawing of analogies etc.). Both these methods are highly time-intensive for both the author and for potential experts, and require extensive psychological expertise, making them not suitable for this study.

Social Acclamation where experts nominate peers is advised against by Winkler & Moser (2016) as it is likely to introduce strong bias and create a panel of too like-minded experts. Self-Ratings circumvent the problem of measuring expertise by letting (potential) experts rate themselves. However, empirical research shows little proof that self-rated experts perform better (Mauksch et al., 2020). Self-Rating could be beneficial when it is used in an intra-individual way rather than an inter-individual way, as suggested by Ward et al. (2002). This does however require substantial more work for both the researcher as the experts (increasing the risk of drop-out) and requires a more in-depth knowledge of psychology making it again not suitable for this study.

This leaves Political Influence, Personal Involvement and External Cues as potential expert identification methods for this study. Political influence is an expert identification method that selects experts based on their potential political influence. It is aimed more at Delphi studies being a socio-political learning activity in which the goal is to develop policies, making it a “form of negotiation around far-minded plans for the future” (Mauksch et al., 2020, p7).

Personal involvement is similar to this approach in the sense that it selects experts based on their involvement in the topic at hand but it takes a wider scope than just including policy makers by also including researchers for example (Mauksch et al., 2020). People that are interested in the topic of the Delphi study tend to have higher response rate and are less likely to drop-out (Hasson et al., 2000). Additionally, selecting experts close to the topic in the Delphi study makes it more likely to produce valid results (Donohoe & Needham, 2009). This could however also lead to bias, as the experts could potentially be influenced by the outcome of the study (Hasson et al., 2000).

The last expert identification method discussed here is External Cues. Mauksch et al. (2020) define this as a wider term that includes any criteria potentially available to the researcher. This is an easy to implement method and the most popular as well (Mauksch et al., 2020). Criteria can be based on any factor deemed relevant by the researcher. Because the cues are defined by the researcher, this approach is likely to have (hidden) biases which is important for the researcher to be aware of (Croce et al., 2016)

Table 9 Overview of expert identification methods (adapted from Mauksch et al., 2020)

Method	**Definition/ explanation**	**Pro**	**Contra**
Social Acclamation	Peer Nomination	• Field experts are good assessors of other domain-specific experts • “Democratic” and holistic approach • Co-nominated experts outperformed laymen in some studies • Peer assessments correlate with external cues • Inclusion of unknown experts • Drop-outs are less likely in nomination procedures	• Social desirability bias or popularity effect, i.e., acclaimed expertise correlates with popularity of a person
Political Influence	Selecting experts with potential political impact	• Easily accessible information • Foresight projects may achieve positive social/environmental change	• Politically powerful individuals are not always experts
Personal Involvement	Selection based on personal interest in the subject	• Deliberate practice as an important element of expertise • Higher response rate • More inclusive expert selection	• Requires operationalization/ measurement • Self-selection bias • Materialistic personal interests may be involved
External Cues	Assessment based on externally available criteria, e.g., years on the job, job position, certification, publications etc.	• Easily accessible information • External reflection of skills • Experience often correlates with improved cognitive skills and deeper knowledge	• Risk of hidden biases in the research team's decisions of who counts as an expert • More experience does not always mean more expertise • Professionals move always up, but seldom down; some professionals never become experts • Not applicable to domains lacking institutionalized criteria of expertise • Good theorists are not necessarily good practitioners (domain expertise vs. process expertise) • Neglects creativity and imagination
Self-Ratings	Self-assessment of expertise	• Low-barrier method for fields that lack objective criteria of expertise • Experts are theoretically the best judges of their performance; experts excel strong self-monitoring skills • Self-rated experts outperform self-rated non-experts	• Ambivalent results from decision research • Biases: overoptimism, overconfidence
Past Performance	Selection based on past performance	• Low-barrier, low-cost method if past performance data is available • Well-established “gold standards” and rating systems exist in some domains (e.g., in several scientific disciplines) • Applicable to repetitive tasks with evaluable outcomes	• Depends on measurable outcomes • Negative evidence for the forecasting of extreme or rare events • Biases: overconfidence after success • Fails to acknowledge non-quantifiable, non-observable aspects about experts (e.g. tacit knowledge)
Knowledge Tests	Selection based on verifiable knowledge	• Allows for sub-selection within a group • Answers are verifiable	• Important, but not sufficient: knowledge elicitation is more important than knowledge alone • Neglects experience-related heuristics • Ethical issues/waste of resources: Participants are tested and then excluded • Neglects creativity and imagination
Psychological Traits	Cognitive tests assessing expertise	• Identification of “true” domain experts who meet important cognitive criteria (inner consistency, discrimination ability, the drawing of analogies etc.)	• High demand for preparation • Some fields lack knowledge about cognitive processes in experts • Not applicable if experts are confronted with new tasks • Involves sophisticated and time-intensive procedures • Neglects creativity and imagination

3.1.2 Expert identification in this study

Mauksch et al. (2020) recommend using a combination of expert identification methods to help mitigate the drawbacks and potential biases of each individual method. This reserach will use a combination of External Cues and Personal Involvement using the expert continuum model of Donohoe & Needham (2009) (see figure 15). The personal involvement approach will be first be used to identify groups of stakeholders and specific and relevant organisations. The external cues approach will then be used to assess the expertise of the potential experts.
Political Influence will not be used as an expert identification method for two reasons. Firstly, this method is more focused on developing policies on politically sensitive and contested topics with several opposing interests which is not the case in this study. Secondly, Political influence is quite closely linked to be Personal Involvement. It is expected that there will be significant overlap in experts identified through the Political Influence method and those identified through the Personal Interest method of the mandated closeness class.

Personal involvement

The main strength of Personal Involvement can also be its main weakness: involving those who might be affected by the outcome of the study can lead to self-selection bias (Mauksch et al., 2020). To help find a balance between impartiality and interest in the topic, the expert continuum model of Donohoe & Needham (2009) could be used (Devaney & Henchion, 2018). In this model, the definition of an expert is worked out in three groups by defining experts along a continuum (figure 15): subjective closeness are experts with hands-on experience, experts with mandated closeness have a formal role (policy/legal) while experts with objective closeness are those involved in the topic from an objective standpoint (Devaney & Henchion, 2018; Donohoe & Needham, 2009).

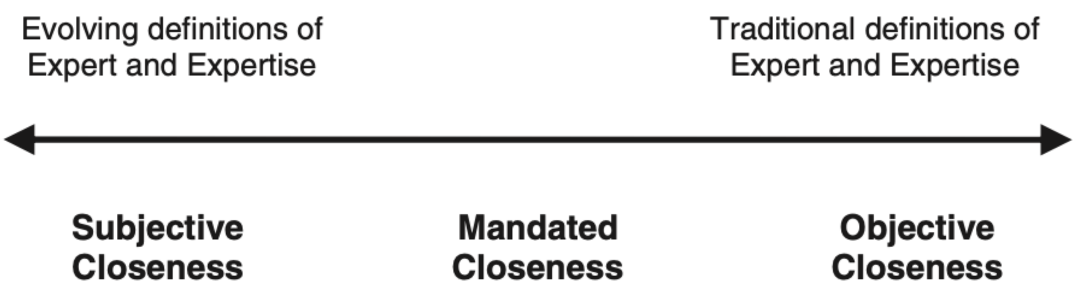


Figure 15 The expert continuum model (adapted from Donohoe & Needham, 2009)

Using this model helps to capture a wide range of experts involved in the topic which in turn could help to balance potential interests and additionally helps to gain a better understanding of the problem and its solutions (Donohoe & Needham, 2009). Table 10 operationalises the model for this research. For each class, relevant organisations are selected based on knowledge gathered in the literature review and discussions with supervisors.

Table 10 The expert continuum model in this research (based on Donohoe & Needham, 2009)

Class of the expert continuum model	Personal interest	Description of organisation	Relevant organisations
Subjective closeness	Hands-on experience	Companies involved in collecting and using vehicle sensor data for safety purposes	OEM BOVAG/RAI
Mandated closeness	A formal role (policy/legal)	Policy makers on the topic of measuring traffic safety and smart mobility	Ministry of Infrastructure & Water Management - Department of Smart Mobility - Department of Traffic Safety RDW Rijkswaterstaat (RWS) CROW
Objective closeness	An objective standpoint	Researchers on the topic of measuring traffic safety or on intelligent vehicles and vehicle sensor data	University researchers - TU Delft - TU/e - UTwente SWOV TNO

External cues

Many researchers report using external cues as a way to measure expertise but often these are not reported in a detailed way or not even reported at all. Table 11 shows several examples of authors that did include criteria for the expert selection for their Delphi study in their paper. Of these, Devaney & Henchion (2018) is the most explicit by also including that any expert must meet three of their five criteria. The criteria in table 11 are also mostly in line with the research of Mauksch et al. (2020) who define three sub-types of external cues that could potentially reflect expertise: experience, certifications and publications, and work positions. Experience is an obvious indicator for expertise and widely accepted as such in the academic literature (Mauksch et al., 2020). One often used way to operationalise this is in years working in the field or in a particular job.

Certifications and publications could serve as a proxy for knowledge. However, in (highly) specific topics no certifications exits, which is why there are experts needed in the first place (Mauksch et al., 2020). Publications could therefore be a more useful criteria, especially in theoretical fields like in this research.

Work positions is the last of the three categories. This is a disputed criteria as a higher job position does not necessarily lead to more expert knowledge (Mauksch et al., 2020). It could serve a role in Delphi studies aimed at for example developments at a strategy level or those concerning policy development, but that is not the case in this study.

Given the goals of the Delphi study (1. to assess the suitability of indicators for traffic safety on network level and 2. to assess the feasibility of collecting and using vehicle sensor data for safety purposes), experts are needed on both measuring traffic safety and on collecting and using vehicle sensor data. It is recognised that it will be difficult for experts to have extensive knowledge on both topics so an expert will be considered as such if the expert has knowledge on either of the two topics.

This, together with the research discussed above and in table 11 below has resulted in the following criteria to assess experts:

1. The expert has at least 5 years of experience with the topic
2. The expert has published a paper or spoken at conference on the topic
3. The expert has a relevant work position the field
4. The expert has a relevant academic background
5. The expert is based in the Netherlands or in a comparable EU country

An expert will be considered having sufficient expertise when they meet at least two of the five criteria. Additionally, the expert needs to have the time and willingness to participate.

Table 11 A non-exhaustive overview of criteria used in other Delphi studies

von der Gracht & Darkow (2010)	Devaney & Henchion (2018)	Roßmann et al. (2018)	Schuckmann et al. (2012)	Warth et al. (2013)
Management level	Contributor to a bioeconomy-related strategy document	Work position	Individual's professional capacity	Knowledge and experience of the issues under investigation
Academic background	Member of a bioeconomy-related advisory, taskforce, industry, or foresight committee	Academic status	Willingness	Capacity and willingness to participate
Job specialisation	Possess at least 5 years of experience in a bioeconomy–related area	Years of experience	Sufficient amount of time capacity to participate.	Sufficient time to participate
Publications	Published in a bioeconomy area or invited to speak at a relevant national event in the last three years	Published works or conference participations		Effective communication skills
Functions inside and outside of the organisation	Represent a media contact point on bioeconomy-related issues	Functional area		
Education		Educational level		
Age		Age		
		Gender		

3.1.3 Number of experts

Next to the question of identifying experts is the question of how many experts should be selected. Similar to the previous question, this question also does not have a clear and unambiguous answer. No consensus exists in the literature on the optimal, or even minimal, number of experts needed for a Delphi study (Donohoe & Needham, 2009; Hsu & Sandford, 2007; Keeney et al., 2006; Okoli & Pawlowski, 2004). The range of number of experts used in past Delphi studies is large, ranging from just three (Norani et al., 2012) to 149 (Devaney & Henchion, 2018) or even several hundreds, as discussed by Keeney et al. (2006). However, typical ranges suggested are between 5 to 20 (Belton et al., 2019), 7 to 15 (Donohoe & Needham, 2009), 10 to 18 (Okoli & Pawlowski, 2004) or 15 to 20 (Hsu & Sandford, 2007).

The number of experts needed depends on the specifics of the study where a sample size should be large enough to have sufficient (and representative) input on the topic (Hsu & Sandford, 2007). At the same time, a larger expert panel also imposes a larger burden on both the researcher as on the experts in terms of data analysis which could result in a lower response rate (Hasson et al., 2000; Hsu. & Sandford, 2007). Keeney et al. (2006, p208) finally conclude that “it [the number of experts] appears to be related to common sense and practical logistics”.

Therefore, in line with the ranges suggested above and the fact that the aim of this Delphi study is quite specific, this study will aim to include between 10 to 15 experts.

To conclude, by employing a combination of two expert identification a balanced expert panel can be created. The personal involvement approach using the expert continuum model helps to introduce heterogeneity in the panel. Heterogeneity can come in many forms but in this case, it means difference in terms of closeness to the topic and the accompanying specific knowledge. Heterogeneous panels have proven to create a wider range of perspectives than an homogeneous panel which helps to prevent bias and framing effects (Förster & von der Gracht, 2014; Winkler & Moser, 2016). This will result in an “inclusive expert population” which will help to mitigate bias in the research (Donohoe & Needham, 2009, p427).

3.1.4 Expert drop-out

A common issue with any Delphi survey is the risk of experts dropping out during the study, especially given the iterative character of the study and the accompanying time commitment (Belton et al., 2019; Hasson et al., 2000; Keeney et al., 2006). Experts dropping out can negatively impact the quality of the study (Belton et al., 2019; Hasson et al., 2000; Keeney et al., 2006). To prevent this, both Belton et al. (2019) and Keeney et al. (2006) point at the importance of making the experts see the goal and relevance of the Delphi study. If the expert selection is done correctly, the experts are interested and involved in the topic, and are perhaps interested or even affected by the outcomes of the study (Belton et al., 2019; Keeney et al., 2006). This can, if the questionnaire is also clear and understandable, help give experts some project ownership and therefore prevent experts dropping out (Keeney et al., 2006). Establishing and communicating a clear timeline and sending remainders via email can also help increase response rates (Turnbull et al., 2018). Experts in the research of Turnbull et al. (2018) do not report these reminders as annoying, making this a viable method of increasing response rates.
These strategies are applied in this research to try and minimise expert drop-out.

3.2 Operationalisation of the Delphi Survey

Developing the questions of a Delphi survey is an essential step that must be carefully deliberated (Belton et al., 2019). Several aspects like the type of questions, the wording of questions, and dealing with bias are discussed in this section. The full questionnaires as sent to the experts are found in Appendix D for round 1 and Appendix E for round 2.

3.2.1 Questions of the Delphi survey

In a Delphi survey, both open-ended and structured questionnaires have been used (Toma & Picioreanu, 2016). While open ended questions can provide a wealth of information and a variety of perspectives, it can also fail to include major subjects known from the literature because experts fail to bring it up (Nowack et al., 2011; Toma & Picioreanu, 2016). Structured questions are much better at this by providing an opportunity to the expert to give their judgement on a specific issue and built upon that (Toma & Picioreanu, 2016). Therefore, Delphi surveys usually consist of a set of questions that require a numerical response, often followed by an opportunity of the expert to comment on the question (Belton et al., 2019).
Next to the type of questions asked, the wording of the questions is also relevant. Research has shown that the wording of questions can influence the quality of the survey data (Markmann et al., 2021). Markmann et al. (2021) have analysed the effect of language on the assessment of respondents on Likert-type scales and have found the following results:

- More abstract language results in more moderate assessment
- Examples lead to more extreme assessments

Additionally, a weak relation was found between lengthy questions and more moderate responses while positive or negative connotations, or the use of modifiers (short words like only, most, often, etc.) were not found to have a clear effect.
Furthermore, a trade-off exists regarding the length of the survey, where a longer survey increases the amount and depth of information gathered but at the same time increases the load on the experts. This could increase the risk of experts dropping out, which negatively impacts the legitimacy of the results (Belton et al., 2019).

So, taking these findings into account, the questionnaire has been developed. The survey will consist of two parts, consistent with the two aims as discussed at the beginning of this chapter. The first aim is to assess the suitability of specific types of indicators to measure traffic safety on a network level.
Seven types of indicators for measuring traffic safety at network level are identified in the literature review. However, from the literature review it has become clear that not every type of indicator identified is as suitable as the others for various reasons. Additionally, if all seven types of indicators are included in the Delphi survey it will become very time-intensive for the experts which increases

the risk of high expert drop-out. Therefore, a selection of four types of indicators is included as will be explained in section 3.2.2 below.

In the Delphi survey, the experts are asked to rate the four types of indicators on four criteria on a scale from 1 (strongly disagree) to 7 (strongly agree). They are then given the opportunity to explain their assessment.
The criteria are based on research by Aarts (2018) for the SWOV where they are used to assess traffic safety indicators. The criteria are made less abstract by using them in a sentence (table 12), which will decrease the room for interpretation by the expert and increase the assessment following from the findings of Markmann et al. (2021). Based on the same research of Markmann et al. (2021), examples of the types of indicators are also included in the questions.

Table 12 Criteria used to evaluate types of indicators in the Delphi survey

Criteria in Aarts (2018)	Criteria operationalised in Delphi survey
Validity	a. This type of indicator reflects traffic safety well
Reliability	b. This type of indicator could be measured in a reliable way
Sensitivity	c. This type of indicator is sensitive to external changes, i.e., it will respond to future traffic safety interventions
Understandability	d. This type of indicator is understandable for different end-users such as researchers and policy makers

Part two of the survey addresses the second aim of the survey, to assess the feasibility of using vehicle sensor data to measure traffic safety. The experts are asked to rate the severity of potential barriers identified in section 2.7 on a 7-point scale (no barrier at all to unsurmountable barrier) and again provide an explanation. All potential barriers discussed in 2.8 are included, and economic feasibility is added as an additional potential barrier.

The Likert-type scale used in this Delphi survey has seven response options. Research suggests that 7-point scales are the most reliable (Beiderbeck et al., 2021a; Toma & Picioreanu, 2016). By having an uneven number of response options, a neutral halfway point is also available. This gives the experts the possibility of explaining a neutral point of view (Toma & Picioreanu, 2016). To make sure this halfway point is truly neutral, an 8th option is also given: I do not know. This will give the experts the option to say 'I do not know" instead of just choosing the middle (neutral) option (Toma & Picioreanu, 2016).
The options are all labelled, as Beiderbeck et al. (2021a) suggest that this will leave little room of interpretation for the experts and gives the highest psychometric quality.

3.2.2 Types of indicators included

This section will discuss the in the literature review identified leading indicators and whether or not these will be included in the Delphi study. This will be decided based on the identified strengths and weaknesses, as well as their suitability to be applied to measuring traffic safety at a network level.

Category	Specific examples	Strengths	Weaknesses	Current usage	Included in Delphi?
Dutch SPIs	Safe participants, safe speeds	1 Help to give insight [g]	1 Still being implemented [g] 2 Little data available [g]	Still in development, aims for network level at national, regional, and local scale [g]	Yes

g Aarts (2018)

The Dutch SPIs will be included in the Delphi study. These are specifically developed to solve some of the issues of using only fatalities and severe injuries as indicators for traffic safety. They can mainly

help gain insight into causes of accidents. They are included because they are still in development and questions about how to measure them remain.
For this study, the two indicators safe participants and safe speeds are the most relevant as these directly concern the driver and its behaviour as opposed to the three other indicators of safe roads, safe vehicles and high-quality trauma care which are out of the scope of this research.

Proximity based SMoS	TTC, PET	1 Directly observable in traffic [h] 2 Objective and physics-based [a]	1 Data intensive 2 Specific for situation [i] 3 Validity of threshold [hij]	(simulation) experiments [hi]	Yes

a Blumenthal et al. (2020)
h Johnsson et al. (2021)
i Mahmud et al. (2017)
j Arun et al. (2021)

Proximity based SMoS will be included in the Delphi study as these are objective metrics which can directly be observed in traffic. For this reason, they are widely used in all sorts of experiments. The question is whether or not these could be useful on a network level, especially given how potentially data intensive this could become. The validity of the threshold could also be a problem. This might be solved by taking a relative approach in which only the change in direction matter.

Kinematic SMoS	Deceleration, acceleration, swerving	1 Easy to understand [a] 2 Objective and physics-based [a] 3 Suitable for several situations [i]	1 Validity of threshold [hij]	Naturalistic driving studies [j]	Yes

a Blumenthal et al. (2020)
h Johnsson et al. (2021)
i Mahmud et al. (2017)
j Arun et al. (2021)

Kinematic SMoS are already being used in naturalistic driving test and form an easy to understand and objective metric. Additionally, they can be used for different types of situations, both when vehicles move in the same direction on one road and when they come from different directions like intersections. This could make them suitable for measuring traffic safety on a network level although the same problem about the validity of threshold values exists as with proximity based SMoS.

Disengagement of ADAS	ACC, LKA	1 Easy to measure [a]	1 Sensitive to context [ak] 2 Low validity [a] 3 Gameable [a]	Field tests [k]	No

a Blumenthal et al. (2020)
k Schwall et al. (2020)

Disengagement of ADAS has been used in the past to measure traffic safety but has since fallen out of favour. This is mainly due to how sensitive the disengagement of the ADAS is to the context which results in questions about its validity. Disengagement of LKA does not need to mean that a situation is unsafe, it could also be due to missing lane markings (the environment) or due to the fact that the driver does not like to use LKA anymore (the driver). Therefore, disengagement of ADAS is deemed to not be a suitable metric to assess traffic safety on a network level and is not included in the Delphi study.

Engagement of ADAS	BSW, ACC, LKA, FCW	1 Easy to measure [l]	?	Public-private pilots[mn]	Yes

i Mahmud et al. (2017)
m Interview with Vrijens, ministry of I&W, Appendix C
n Kia Nederland (2021)

Engagement of ADAS is used in pilot projects discussed in section 2.7 like the Road Monitor and Kia Insights from vehicle data. Similar to disengagement of ADAS, it is easy to measure and perhaps also sensitive to the context in which it takes place. There are however two important differences. Firstly, FCW is a passive ADAS that only intervenes when necessary. This makes that an engagement of FCW is more likely to be a sign of an unsafe situation than the disengagement of an active ADAS like LKA. Secondly, while disengagement of ADAS is discussed in scientific literature and deemed to not be valid, this is not the case for engagement of ADAS. No research into its validity was found, leaving this as an open question. Therefore, this is included in the Delphi study.

Infractions	Speeding	1 Statistically significant relationship with crashes [a] 2 Could be compared to HDV [a]	1 Context dependent [a] 2 Relationship not well understood or strong [a]	?	No

a Blumenthal et al. (2020)

While there is a statistically significant relationship between infractions and accidents, it is weak and not well understood. Additionally, just because there is a statistically significant correlation between the two, it does not necessarily mean that there is a causal relationship between the two. The context of the infraction is important as sometimes, as sometimes an illegal move like speeding is the safest move if it can help avoid an accident. Therefore, using infractions as a metric to assess safety on a network level is not a good idea and not included in the Delphi study.

Holistic Roadmanship measures	Safety envelop violation	1 Objective and physics-based [a]	1 No uniform definition of roadmanship [ao] 2 Data intensive [aop]	Still in development [ao]	No

a Blumenthal et al. (2020)
o Fraade-Blanar et al. (2018)
p Wishart et al. (2020)

Roadmanship is a concept coined by the RAND cooperation that aims to describe the ability of a vehicle to drive safely. It is an interesting development as it goes deeper than other metrics describe here and aims to be more holistic in its assessment. However, it is a new concept without a clear definition that is still in development. No papers were found specifically on this concept in the databases of ScienceDirect or Scopus. Therefore, it is deemed too underdeveloped and unknown to be useful and thus not included in the Delphi study.

So, based on the analysis of the strengths and weaknesses of the types of indicators found in part 1, four types of indicators will be included in the Delphi study:

- Dutch SPIs
- Proximity based SMoS
- Kinematic SMoS
- Engagement of ADAS

3.2.3 Additional questions included in round 2

In addition to the questions where the experts had to evaluate either types of indicators or potential barriers, three other open questions were asked. The experts were invited to suggest types of indicators (question 1.5) and potential barriers (question 2.7) not discussed in the survey. This was done to help make sure all relevant types of indicators and potential barriers are included in the survey. Table 13 below shows the results, and it can be seen that both for the type of indicators and potential barriers one suggestion was made by several experts. 5 experts suggested to use driver distraction as a type of indicator for traffic safety. 5 experts suggested that the willingness of people to let their data be used for traffic safety research could be a barrier as well. Therefore, these two have been added as an additional type of indicator and as an additional potential barrier in the second round of the survey.

Table 13 The experts' answers to questions about additional types of indicators and potential barriers

Question: Are there any other types of indicators based on vehicle sensor data not mentioned above that you think could potentially be a suitable indicator for measuring traffic safety at network level?		**Number of responses (N=16)**
Responses by experts	N/A	5
	Driver distraction/driver behaviour	5
	Lane deviation (due to sleepy driver)	1
	ADAS active?	2
	Environmental/external conditions	2
	Infractions (running a red light)	1
Question: Are there any other factors not mentioned above that you think could potentially be a barrier for implementing a system of using vehicle sensor data to measure traffic safety?		
Responses by experts	N/A	5
	No	1
	Willingness of people to participate	5
	Storage of data	1
	Sufficient number of vehicles to get reliable results	1
	Organisational problems (combination of technical issues, economic viability, and willingness of stakeholders)	3

3.2.4 Number of rounds

Typical for a Delphi study is the iterative character of it by asking question over several rounds (Belton et al., 2019; Fritschy & Spinler, 2019). But similarly to the number or selection of experts, there is no single answer to how many rounds the Delphi survey should have (Schmalz et al., 2021). Early Delphi studies that were highly focussed on reaching consensus used to have four rounds, but this has since then decreased to two or three rounds (Hasson et al., 2000). This is also shown in a review by Nowack et al. (2011) who found that a large majority of the Delphi studies they reviewed conducted two rounds, a minority three and only one did four rounds. The number of rounds usually depends on the set-up of the Delphi survey and on the time available (Hasson et al., 2000; Keeney et al., 2006). From the perspective of the experts are also two important aspects to consider. Too lengthy and complex Delphi surveys can lead to expert fatigue and thus a lower response rates over each subsequent round (Frewer et al., 2011; Fritschy & Spinler, 2019; Keeney et al., 2006). This already happens after two rounds, especially when experts are busy (Keeney et al., 2006). Additionally, most revisions of the expert's own opinions happen after the first round and not afterwards (Fritschy & Spinler, 2019).
So, for these reasons the Delphi survey in this study will consist of two rounds.

3.2.5 Feedback to experts in round 2

Next to anonymity and iteration, controlled feedback and statistical group response are the two other important characteristics of a Delphi study (Belton et al., 2019; Fritschy & Spinler, 2019). Controlled feedback refers to the feedback presented to the experts in the second round and is called "controlled" feedback because the researcher decides on the type of feedback and how it is presented (von der Gracht, 2012). This feedback usually consists of the statistical group response, possible augmented with qualitative feedback (von der Gracht, 2012).

In this research, the controlled feedback that will be provided to the experts in the second round consists of a histogram showing the distribution of scores given by the experts in the first round to a specific question. This gives the experts a visual overview of the scores given per question in the first round. This is augmented with qualitative feedback that consists of arguments brought up by the experts, both in favour and against. Not all arguments made in the first round are included in the survey to keep the workload for the experts reasonable. A selection is made of the most used arguments of which some are edited to remove any references that could reveal the identity of the

experts. Changes to improve the grammar or flow of certain arguments are made as well, although an effort is made to keep the arguments as much in the same wording as possible.
Furthermore, a suggestion by Winkler & Moser (2016) (see 3.3 below) is followed for most questions to sort the arguments and put those contrary to the central tendency on top, to make sure that experts are confronted with opposite opinions and are thus forced to rethink their own opinion.

3.2.6 Distribution of the Delphi survey

The Delphi survey is sent to the experts via email in the form of a Word document, similarly to Fritschy & Spinler (2019). This method is chosen over dedicated survey software as this does not require an active internet connection, it is easy for the experts, it allows them to spent as much time as necessary and finally, because it provides an easy platform to transfer data to other digital platforms for later analysis (Toma & Picioreanu, 2016).
Included in the questionnaire is an introduction including the goal and setup of the survey, and contact details, as recommended by Schmalz et al. (2021). Because the research concerns human subjects, approval of the Human Research Ethics Committee of the TU Delft is sought which includes a Data Management Plan to ensure the privacy of the experts. The experts are notified of this in the introduction where they also give their informed consent to participating in the study.
The questions are presented in a consistent format to let the experts get accustomed to the format, as suggested by Beiderbeck et al. (2021a). It also gives an estimate of the time necessary to complete the survey.

3.3 Dealing with bias

Even in carefully constructed questionnaires, several types of biases may still be present that negatively affect the results. Winkler & Moser (2016) have identified four types of bias and potential solutions as is summarised in table 14. Framing bias and Desirability bias can exist in both rounds of the survey while the Bandwagon effect and Belief-persevereance two are specific to the second round.

Table 14 Types of bias present in Delphi studies bias (Winkler & Moser, 2016)

Bias	**Explanation**	**Solution**
Framing bias	Presentation of questions influences the assessment of the issue	1 Heterogeneity in panel 2 More involved experts are less susceptible 3 Explicit warning against framing (Cheng and Wu, 2010)
Desirability bias	Desirability of an event positively influences a person's likelihood judgment	1 Design of Delphi study (rounds) 2 Heterogeneity in panel
Bandwagon effect	Urge to confirm to the majority opinion	1 Design of Delphi study (anonymity) 2 Heterogeneity in panel 3 Configuration of the provided feedback 4 Consider not using statistical feedback
Belief-perseverance	Experts may overweight their own judgment and underweight other available advice	1 High-quality argumentative feedback 2 Give participants one or two examples of good and poor reasons 3 Make use of warning and counter- arguments

Several of the counters against the biases have to do with nature and design of a Delphi study and with the composition of the expert panel, which have been discussed extensively in section 3.1. The quality of the feedback provided to the experts in the second round is important as well. According to Winkler & Moser (2016), not all arguments of experts from the first round should be included in the second round as feedback. The researcher should filter the arguments to prevent duplicates and low-quality arguments, instead, the feedback should “provide good cues about where the most accurate answers lie” (Winkler & Moser, 2016, p70). By doing this, a significant influence of bias should be prevented.

3.4 Analysis of the Delphi survey

The analysis of the Delphi survey consists of several aspects that are discussed below. A syntax analysis is used to assess the commitment of experts. Non-response bias could influence the results and it is therefore checked if this is present. The level of agreement is used to measure consensus, but different metrics exist. This section will discuss which will be used in this research.

Syntax analysis

A syntax analysis of the explanations given by the experts in round 1[7] is conducted to show their level of engagement (Beiderbeck et al., 2021a). The syntax analysis as is done in this study was originally developed by Förster & von der Gracht (2014) and also applied in Roßmann et al. (2018) and in Beiderbeck et al. (2021b). In a syntax analysis, the experts' explanations are classified as whole sentences, phrases, catchwords, or no explanation. A high percentage of whole sentences indicates a high level of commitment and thus serves as a quality measure (Beiderbeck et al., 2021a).

Non-response bias

Influential research by Armstrong & Overton (1977) suggested that those who do not respond to physically mailed surveys may be substantially different from those who do respond, thus introducing bias. Hudson et al. (2004) showed that surveys distributed via the internet are not significantly different from physically mailed surveys in this aspect. One way of controlling for this bias is by comparing the answers of early respondents to those of late respondents (Díaz de Rada, 2005). This can be done by dividing the group of respondents in two by order of responses (Warth et al., 2013) or by those who responded initially and those who only responded after reminders (Díaz de Rada, 2005). The idea behind both methods is that the latter groups were more reluctant to answer the survey and thus are more similar to those who did not respond at all (Armstrong & Overton, 1977; Díaz de Rada, 2005). If these two groups differ difference in a statistically significant way can be tested with a Mann-Whitney U test (Piecyk & McKinnon, 2010; Warth et al., 2013).

Level of agreement

Measuring the level of agreement is often done in a rather simple way by defining a certain level of agreement such as an (absolute) majority or a standard deviation of ± 1.64 (von der Gracht, 2012). The choice for criteria and thresholds are usually made based on the goal of the study but are still chosen rather arbitrarily (von der Gracht, 2012). This study will use two established and robust metrics to measure the level of agreement to decrease this subjectivity: the IQR and the Coefficient of variation (V).

An often used and seen as rigorous and objective metric of level of agreement is the IQR (von der Gracht, 2012). The IQR is a measurement of dispersion based on the median and is the difference between the 25th and the 75th percentile values. Thus, the range of the IQR depends on the size of the scale. On 5-point scales, an IQR of less than 1 is often used as a level of agreement that signals consensus (Raskin, 1994; Ray & Sahu, 1990) while 2 is used as a threshold on a 9-point scale (von der Gracht & Darkow, 2010) or on a 10-point scale (Linstone & Turoff, 1975; Scheibe et al., 2002). No research using both the IQR, and a 7-point scale was found, so a IQR of less than 1,5 will be used as threshold for having reached a satisfying level of agreement.

Next to the IQR, the coefficient of variation (V) will be used to this end. V allows for comparing the distributions of answers on a scale like IQR but is based on the mean (von der Gracht, 2012). It is calculated as the standard deviation divided by the mean multiplied by 100 (English & Kernan, 1976; von der Gracht, 2012). It is interpreted as is shown in table 15.

[7] This analysis is only done for round 1 because in round 2 the explanation was optional (to decrease the workload for the experts filling in the survey).

Table 15 Coefficient of Variation (V) and consensus (English & Kernan, 1976; von der Gracht, 2012)

Coefficient of variation (V)	Decision rule
$0 < V \leq 0,5$	Good degree of consensus. No need for additional round
$0,5 < V \leq 0,8$	Less than satisfactory degree of consensus. Possible need for additional round.
$V > 0,8$	Poor degree of consensus. Definite need for additional round.

Stability over rounds

In many Delphi studies, specific questions are not included in the subsequent round of the survey when a specific level of agreement is reached (von der Gracht, 2012). However, Dajani et al. (1979) discusses that a specific level of agreement can be meaningless when stability over rounds is not reached or not tested. Stability over rounds in this case means the stability of the group's responses on a specific question, as opposed to individual stability. This is because the focus of a Delphi study is the opinion of the group (von der Gracht, 2012).

Stability over rounds can be measured in a variety of ways but is usually measured based on the change in level of agreement over rounds (von der Gracht, 2012). This study uses two metrics to determine the level of agreement which both are based on the dispersion, namely IQR and the coefficient of variation (V). If these stays at a similar level or decrease, it signals stability or convergence of opinions on that question.

Additionally, inferential statistics can be used to test for stability between rounds in a more robust way (von der Gracht, 2012). These can establish relationships among variables, in this case between the answers on the same questions of round 1 and 2. These answers are dependent as they are given by the same experts to the same questions (Argyrous, 1997). Therefore, a McNemar Chi-square test (in the case of interval scales) or a Wilcoxon matched-pairs signed-ranks test (in the case of ordinal scales) can be used to test if the differences in two sets of answers are statistically significant (von der Gracht, 2012). Given the ordinal scales in this research the Wilcoxon signed rank test is used, where a p-value of lower than 0,05 indicates that there is a statistically significant difference between the two observations (at a 95% confidence level). This was in none of the 25 questions the case, meaning that in none of the questions, the answers from the first and second round differ in a statistically significant way.

4. Results

The results of the Delphi study consist of the numerical answers of the experts to the questions, as well as their written explanations. This chapter will discuss these results and initial analysis. It is split in three parts that each serve a different goal. Descriptive statistics give an overview of the data and provide the basis for the feedback to the experts in the second round. Analysing and quantifying the degree of consensus (or dissent) is one of the main goals of a Delphi study (Diamond et al., 2014). Further analysis can help reveal where dissent comes from.

4.1 Descriptive statistics

This section will discuss descriptive statistics on both the experts and their responses, as well as on the actual quantitative results. The former shows the quality and level of engagement of the experts while the latter gives an overview of the data.

4.1.1 Experts and engagement

27 experts were invited to participate in the Delphi study. 16 experts completed round 1 which is a response rate of 59%. All experts that responded to round 1 met at least three of the five criteria formulated in section 3.1 with most of the experts scoring four or five out of five. On average, the experts had 11 years of experience with traffic safety and/or vehicle sensor data. This confirms a high level of expertise of the invited experts.

Table 16 shows the distribution of experts over the three classes which is highly similar in both the group of invited experts and in the group that responded in both rounds 1 and 2. Academia (university researchers, SWOV and TNO) form a significantly larger group than the other two. Overrepresentation of one group could potentially introduce bias. However, given the nature of the topic - a future development – it makes sense to have a larger share of researchers than those with hands-on experience (industry) or a formal role (government).

Table 16 Distribution of experts per class

	Invited		Response R1		Response R2	
Government	6	22%	3	19%	2	18%
Academia	17	63%	10	63%	7	64%
Industry	4	15%	3	19%	2	18%
Total	27		16		11	

In the Delphi survey the experts are asked to make an assessment on a 7-point scale and then explain why they make this assessment. In 90% of the questions of round 1, a score of 1 to 7 was given. The remaining 10% were mostly a 0, meaning “I do not know” or in a few cases were not filled in. Given the fact that the latter occurred only a few times and are spread out across questions and experts, not filled in scores are presumed to be the same as a 0.

Table 17 below shows the syntax analysis for round 1. It can be seen that in total 74% of the explanations were given in whole sentences. This is a lower than the percentage of whole sentences in Beiderbeck et al. (2021b) (87%) but similar to those in Förster & von der Gracht (2014) (72%) and Roßmann et al. (2018) (78%) and signals a high degree of engagement by the experts. Table 17 also shows no large difference between part 1 (type of indicators) and part 2 (potential barriers). Only the share of phrases is larger in part 2 (13% vs 2% in part 1). This could be due to the differences in the questions. The questions in part 1 consisted of one textual explanation per four scores (for the four criteria). Part 2 asked for with one explanation per score which may require a less elaborate response.

Table 17 Syntax analysis (based on Förster & von der Gracht, 2014)

	Part 1 (type of indicators)		Part 2 (potential barriers		Total	
	Total	Percentage	Total	Percentage	Total	Percentage
Whole sentences	50	78%	104	72%	154	74%
Phrases	1	2%	18	13%	19	9%
Catchwords	0	0%	1	1%	1	0%
No explanation	13	20%	21	15%	34	16%

Splitting the group in half based on response order showed no significant differences between the groups. Splitting the group by those who received a reminder and those who did not showed one question (1.2c: sensitivity of proximity based SMoS) where a statistically significant difference was found at an alpha of 0.05 (p=0.039). Since only one out of twenty-five questions showed a statistically significant difference for one of the approaches, it can be concluded that no strong non-response bias is present.

4.1.2 Quantitative results

Descriptive statistics in a Delphi study often includes the central tendency (in the form of the mean or median) and the dispersion (e.g. standard deviation) (von der Gracht, 2012). The mean is often used in Delphi surveys to show the central tendency, although this is strictly speaking not correct. A Likert scale is an ordinal scale, making the mean an invalid statistic and thus the median the preferred choice for a metric of central tendency (von der Gracht, 2012). Additionally, the median is less sensitive to outliers than the mean (Gordon & Pease, 2006; von der Gracht, 2012).
Consequently, the standard deviation technically not the correct statistic to show the dispersion of the experts' assessments, as it is based on the mean. Instead, Interquartile Range (IQR) will be used as it is a measure of dispersion for the median. The IQR is the distance between the 25th and the 75th percentile values (De Vet et al., 2005). Therefore, a value below 1 means that more than 50% of the scores are within one point on the scale (De Vet et al., 2005). Tables 18 and 19 below show the descriptive statistics for round 1

Table 18 Descriptive statistics Delphi survey part 1

Type of indicator	Criteria	Round 1					Round 2				
		N	Median	Mean	IQR	SD	N	Median	Mean	IQR	SD
1: Dutch SPIs	Validity	16	5	5,3	1	1,2	11	5	5,2	1	1,0
	Reliability	16	6	5,6	0,25	1,4	10	6	6,0	0	0,7
	Sensitivity	15	5	5,3	1	1,3	11	5	5,4	1,5	1,5
	Understandability	16	6	5,9	1	1,5	10	6	6,1	0	0,6
2: Proximity based Surrogate Measures of Safety (SMoS)	Validity	15	6	5,4	1	1,4	11	6	5,5	1	0,8
	Reliability	14	5	4,6	1	1,2	10	5	4,8	0	1,1
	Sensitivity	13	5	5,0	1	1,2	11	5	5,4	1	0,7
	Understandability	15	5	4,5	2	1,7	11	5	4,5	1	1,4
3: Kinematic based Surrogate Measures of Safety (SMoS)	Validity	15	5	5,5	1	0,9	11	6	5,5	1	0,8
	Reliability	15	6	5,5	1	1,2	11	5	5,0	2	1,3
	Sensitivity	13	5	5,2	1	0,8	10	5	5,0	0	0,7
	Understandability	15	5	5,1	2	1,3	11	5	4,9	2	1,3
4: Engagement of ADAS	Validity	15	6	5,6	1,5	1,4	11	6	5,5	1	1,2
	Reliability	15	6	5,3	2,5	1,4	11	5	4,9	2	1,6
	Sensitivity	13	5	5,1	2	1,4	11	5	4,9	1	1,2
	Understandability	15	6	5,5	1	1,0	11	5	5,5	1	0,8
5: Driver distraction	Validity	-	-	-	-	-	8	5	5,4	1	0,9
	Reliability	-	-	-	-	-	8	4	4,0	2	1,3
	Sensitivity	-	-	-	-	-	8	4,5	4,8	1,25	0,9
	Understandability	-	-	-	-	-	9	5	5,3	1	1,0

Table 19 Descriptive statistics Delphi survey part 2

Potential barrier	Round 1					Round 2				
	N	Median	Mean	IQR	SD	N	Median	Mean	IQR	SD
Technical feasibility (collecting data within a single vehicle)	16	4	3,8	2,25	1,6	11	4	3,6	1	1,2
Technical feasibility (extracting and processing data from a fleet)	15	4	3,9	1	1,4	11	4	4,0	1,5	1,0
Legal feasibility	15	5	5,1	1,5	1,2	11	5	5,3	1	1,1
Economic feasibility	14	4,5	4,4	1	1,5	10	4	4,1	1,75	1,4
Cybersecurity	14	4	4,5	2	1,5	11	4	4,5	2	1,4
Willingness of OEMs	15	5	5,0	1,5	1,4	10	5	4,8	0,75	1,2
Willingness of suppliers	8	5	4,5	3,25	1,9	6	4	4,0	2	1,5
Willingness of service providers	12	2	3,0	1,25	1,5	9	2	2,8	1	1,3
Willingness of people	-	-	-	-	-	11	5	4,1	2	1,6
From pilot to reality	14	4,5	4,3	2,5	1,6	10	4,5	4,6	2	1,7

4.2 Analysing consensus

Measuring consensus is an essential part of any Delphi survey, yet defining consensus has proven to be difficult (von der Gracht, 2012). Defining and measuring consensus can, and is, done in a large variety of ways, as is shown and discussed by von der Gracht (2012). An important conclusion of von der Gracht (2012) is that measuring consensus consists of two aspects: the level of agreement and the stability of opinions over rounds.

4.2.1 Level of agreement

Using the IQR and coefficient of variation (V) as metrics for the level of agreement, it can be concluded that in part 1 consensus is reached after two rounds on the majority of the questions (on 16 out of 20 questions). Table 20 shows the level of agreement for part 1 in rounds 1 and 2 while table 21 shows this for part 2. The V is well below the threshold of 0,5 in all questions. The IQR is above the threshold of 1,5 on four criteria, spread out over three types of indicators. In part 2, consensus is reached on five out of ten potential barriers after the second round. The remaining five do not reach consensus of which one is the Willingness of people which was only asked in the second round.

Table 20 Level of agreement in part 1

Type of indicator	Criteria	Round 1			Round 2		
		IQR	V	Consensus?	IQR	V	Consensus?
1: Dutch SPIs	Validity	1	0,23	Yes	1	0,19	Yes
	Reliability	0,25	0,24	Yes	0	0,11	Yes
	Sensitivity	1	0,24	Yes	1,5	0,28	Yes
	Understandability	1	0,26	Yes	0	0,09	Yes
2: Proximity based Surrogate Measures of Safety (SMoS)	Validity	1	0,26	Yes	1	0,15	Yes
	Reliability	1	0,25	Yes	0	0,24	Yes
	Sensitivity	1	0,23	Yes	1	0,13	Yes
	Understandability	**2**	0,38	**No**	1	0,32	Yes
3: Kinematic based Surrogate Measures of Safety (SMoS)	Validity	1	0,17	Yes	1	0,15	Yes
	Reliability	1	0,22	Yes	**2**	0,27	**No**
	Sensitivity	1	0,16	Yes	0	0,13	Yes
	Understandability	**2**	0,25	**No**	**2**	0,26	**No**
4: Engagement of ADAS	Validity	1,5	0,24	Yes	1	0,22	Yes
	Reliability	**2,5**	0,27	**No**	**2**	0,33	**No**
	Sensitivity	**2**	0,28	**No**	1	0,25	Yes
	Understandability	1	0,18	Yes	1	0,15	Yes
5: Driver distraction	Validity	-	-	-	1	0,17	Yes
	Reliability	-	-	-	**2**	0,33	**No**
	Sensitivity	-	-	-	1,25	0,19	Yes
	Understandability	-	-	-	1	0,19	Yes

Table 21 Level of agreement in part 2

Potential barrier	Round 1			Round 2		
	IQR	V	Consensus?	IQR	V	Consensus?
Technical feasibility (collecting data within a single vehicle)	**2,25**	0,43	**No**	1	0,33	Yes
Technical feasibility (extracting and processing data from a fleet)	1	0,35	Yes	1,5	0,25	Yes
Legal feasibility	1,5	0,23	Yes	1	0,21	Yes
Economic feasibility	1	0,33	Yes	**1,75**	0,35	**No**
Cybersecurity	**2**	0,33	**No**	**2**	0,31	**No**
Willingness of OEMs	1,5	0,27	Yes	0,75	0,26	Yes
Willingness of suppliers	**3,25**	0,43	**No**	**2**	0,39	**No**
Willingness of service providers	1,25	**0,51**	**No**	1	0,47	Yes
Willingness of people	-	-	-	**2**	0,37	**No**
From pilot to reality	**2,5**	0,37	**No**	**2**	0,39	**No**

4.2.2 Stability over rounds

In round 2, experts were given the chance to revaluate their own scores and change them if they wished based on the additional information provided. 54 changes were made out of the 258 possible changes (20,9%). The largest number of changes per expert was twelve while three experts did not make any change. On average each expert made approximately five changes while the median number of changes is four. 35% of the changes were in a positive direction, meaning for part 1 a higher score was given on a criteria for a type of indicator while for part 2 a potential barrier was evaluated as lower. Consequently, 65% of the changes were in a negative direction. The proportion of changes were similar for both parts.

Of all the revisions, the large majority (81,5%) were changes of one point on the scale and the remainder of two points in the scale. One change was the exception with a change of -4, which can be attributed to a misunderstanding of the question in round 1 based on the comment made by the expert.

While the share of revisions made is a little lower than those in Fritschy & Spinler (2019) and Roßmann et al. (2018) with 25% and 37% respectively, the proportions of upwards and downwards revisions are similar (61% up/39% down and 56% up/44% down in Fritschy & Spinler, 2019, and Roßmann et al., 2018)

Tables 23 and 24 below show whether or not consensus exists in rounds 1 and 2, based on the IRQ and V. It then shows how this has changed between the two rounds and if there is a converging trend. A change occurred in 6 questions out of 25 (16 in part 1, 9 in part 2). Twice this change was from consensus to dissent and four times from dissent to consensus. In the latter, both the IQR and the V become smaller indicating a decline in the dispersion of answers. This should ideally be the case for all questions as this signals a greater degree of consensus in the second round. In 19 out of 25 questions this convergence does take place. In the six questions where no convergence can be observed, two questions go from consensus in round 1 to dissent in round 2 while two questions have dissent in both rounds. This means that two questions have no converging trend from round 1 to round 2 but still stay below the threshold for consensus in both rounds.

The last column finally shows whether further analysis is done. Further analysis is done when no consensus is reached after the second round to reveal why this is the case. This will be done in the next section for the seven questions where no consensus was reached after two rounds and for two of the additional questions where no consensus was reached in the first and only round.

Table 22 Consensus and stability in part 1

Type of indicator	Criteria	Round 1			Round 2			Comparison					Further research?
		IQR	V	Con-sensus?	IQR	V	Con-sensus?	Change R1 to R2?	IQR change (%)	V change (%)	Wilcoxon signed rank test (p-value)	Conver-ging?	
1: Dutch SPIs	Validity	1	0,23	Yes	1	0,19	Yes	No	0	-16	0,102	Yes	No
	Reliability	0,25	0,24	Yes	0	0,11	Yes	No	-100	-51	0,157	Yes	No
	Sensitivity	1	0,24	Yes	1,5	0,28	Yes	No	50	15	0,564	**No**	No
	Understandability	1	0,26	Yes	0	0,09	Yes	No	-100	-63	0,257	Yes	No
2: Proximity based Surrogate Measures of Safety (SMoS)	Validity	1	0,26	Yes	1	0,15	Yes	No	0	-43	1.000	Yes	No
	Reliability	1	0,25	Yes	0	0,24	Yes	No	-100	-7	0,317	Yes	No
	Sensitivity	1	0,23	Yes	1	0,13	Yes	No	0	-46	0,564	Yes	No
	Understandability	2	0,38	**No**	1	0,32	Yes	**Yes**	-50	-14	1.000	Yes	No
3: Kinematic based Surrogate Measures of Safety (SMoS)	Validity	1	0,17	Yes	1	0,15	Yes	No	0	-10	0,083	Yes	No
	Reliability	1	0,22	Yes	2	0,27	**No**	**Yes**	100	24	0,317	**No**	**Yes**
	Sensitivity	1	0,16	Yes	0	0,13	Yes	No	-100	-14	1.000	Yes	No
	Understandability	2	0,25	**No**	2	0,26	**No**	No	0	5	0,317	**No**	**Yes**
4: Engagement of ADAS	Validity	1,5	0,24	Yes	1	0,22	Yes	No	-33	-8	0,317	Yes	No
	Reliability	2,5	0,27	**No**	2	0,33	**No**	No	-20	23	0,317	**No**	**Yes**
	Sensitivity	2	0,28	**No**	1	0,25	Yes	**Yes**	-50	-12	0,157	Yes	No
	Understandability	1	0,18	Yes	1	0,15	Yes	No	0	-17	0,317	Yes	No
5: Driver distraction	Validity	-	-	-	1	0,17	Yes	-	-	-	-	-	No
	Reliability	-	-	-	2	0,33	**No**	-	-	-	-	-	**Yes**
	Sensitivity	-	-	-	1,25	0,19	Yes	-	-	-	-	-	No
	Understandability	-	-	-	1	0,19	Yes	-	-	-	-	-	No

Table 23 Consensus and stability in part 2

Potential barrier	Round 1			Round 2			Comparison					Further research?
	IQR	V	Con-sensus?	IQR	V	Con-sensus?	Change R1 to R2?	IQR change (%)	V change (%)	Wilcoxon signed rank test (p-value)	Conver-ging?	
Technical feasibility (collecting data within a single vehicle)	2,25	0,43	**No**	1	0,33	Yes	**Yes**	-56	-23	0,317	Yes	No
Technical feasibility (extracting and processing data from a fleet)	1	0,35	Yes	1,5	0,25	Yes	No	50	-29	0,655	**No**	No
Legal feasibility	1,5	0,23	Yes	1	0,21	Yes	No	-33	-9	0,18	Yes	No
Economic feasibility	1	0,33	Yes	1,75	0,35	**No**	**Yes**	75	8	0,564	**No**	**Yes**
Cybersecurity	2	0,33	**No**	2	0,31	**No**	No	0	-8	0,414	Yes	**Yes**
Willingness of OEMs	1,5	0,27	Yes	0,75	0,26	Yes	No	-50	-6	0,317	Yes	No
Willingness of suppliers	3,25	0,43	**No**	2	0,39	**No**	No	-38	-10	0,317	Yes	**Yes**
Willingness of service providers	1,25	0,51	**No**	1	0,47	Yes	**Yes**	-20	-9	0,317	Yes	No
Willingness of people	-	-	-	2	0,37	**No**	-	-	-	-	-	**Yes**
From pilot to reality	2,5	0,37	**No**	2	0,39	**No**	No	-20	4	0,083	Yes	**Yes**

4.3 Further analysis

Common explanations for dissent include the presence of outliers, bipolarity in the responses, and different opinions by different stakeholder groups (Beiderbeck et al., 2021a; Warth et al., 2013).

4.3.1 Outlier analysis

Outliers in a dataset can unrealistically influence the mean and standard deviation (Warth et al., 2013) as well as the IQR (Beiderbeck et al., 2021a). Checking for outliers can be done by converting the scores into standardised z-scores and highlighting those more extreme than the absolute value of 2,58 (the 99% confidence level) (Beiderbeck et al., 2021b; Warth et al., 2013). Fritschy & Spinler (2019) check for outliers by visually inspecting the boxplots in SPSS which automatically flag outliers. Both methods do not show any outliers for the nine questions analysed in this section and thus rule out outliers as a possible reason for the present dissent.

4.3.2 Bipolarity analysis

Bipolarity is a second explanation for dissent because it reveals any possible opposing groups of experts with different opinions (Dajani et al., 1979). If bipolar opinions are present, they will almost always prevent consensus (Beiderbeck et al., 2021a). Checking for bipolarity can be done by visually inspecting the histograms (Beiderbeck et al., 2021b; Warth et al., 2013) and by checking for multiple modes (Beiderbeck et al., 2021b; Scheibe et al., 2002). Out of the nine questions without consensus, four have multiple modes as can be seen in figures 16 and 17. In three of these four cases, the modes are also not directly next to each other[8]. Cybersecurity as a potential barrier is also included in figure 17 even though it does not have two modes, but it does clearly have two different groups of opinions not directly adjacent to each other.

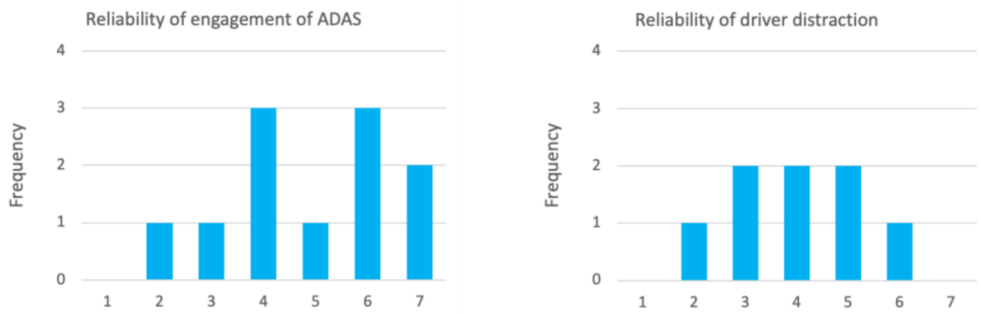


Figure 16 Histograms of the questions that appear to have bipolar opinions in part 1

The bipolarity apparent in reliability of both engagement of ADAS and driver distraction could not be confirmed based on the qualitative comments of the experts. In the questions regarding types of indicators, a single explanation was asked for the four scores given on the four criteria to prevent the Delphi survey from becoming too long and burdening the experts too much. As a consequence, however, not all experts discuss reliability in their qualitative answers, meaning that too little information is available to discuss the apparent bipolarity in these individual criteria in a valid way.
This is different for the questions regarding potential barriers where each score given is accompanied by a comment. The apparent bipolarity in cybersecurity, willingness of suppliers, and willingness of people as potential barriers could be confirmed based on the comments made by the experts, as will be discussed below.

[8] In the 20 other questions that did reach consensus, only four have multiple modes which also lie directly adjacent. This strengthens the case that the bipolarity observed in the five cases here does indeed prevent consensus, and that in the other 20 cases no bipolarity exists.

Figure 17 Histograms of the questions that appear to have bipolar opinions in part 2

In the case of cybersecurity being a potential barrier, all experts agree that there are significant risks, but the disagreement lies in how well these can be contained. On one hand, those who do believe the risks can be minimised refer to “recent advances in cybersecurity and data protection protocols” or to the fact that “the communication of the data is not time critical and vehicle decisions are not based on it”. On the other hand, different experts point data leaks in relation to privacy laws, the consequences for public user acceptance, and one expert made the link between the risk of cybersecurity attacks and the willingness of OEMs to facilitate the sharing of data as no OEM would “jeopardize the IT security for a better software service level, at least not now”.

There are also two groups of opinions on the extent to which the willingness of suppliers a barrier is to the implementation of a system where vehicle sensor data is used to measure traffic safety. One group points out that with a sufficient economic incentive, suppliers will be willing to cooperate. The other group believes that suppliers are unlikely to be willing to participate in any sharing of data as it would threaten their IP and competitive position.

Willingness of people as a potential barrier is divisive as well. Those who think it is a strong barrier believe that people are not willing to share their data because they will not see the benefits, especially when the data can be linked to traffic violations. Other experts see it more as a moderate barrier with one expert expecting “some discussion at the introduction but people will forget it over time” by making a comparison to the public acceptance of smartphones and the data those collect. Other experts propose incentives for the sharing of their data and point to the fact that not all vehicle owners would need to share their data, only a (representative) part.

4.3.3 Stakeholder group analysis

The stakeholder group bias is the third and final analysis that could explain a lack of consensus. Different stakeholder groups could hold different perspectives on the same topic due to different interests (Warth et al., 2013). A Mann-Whitney test can be used to assess if statically significant differences between stakeholders exist (Beiderbeck et al., 2021a; Warth et al., 2013). This research has clearly defined stakeholder groups which were used in the expert selection (academia, government, and industry). Only for the potential barrier economic feasibility statistically significant (p-value <0,05) differences were found between the stakeholder groups academic and industry. The industry rated this as a higher barrier (median score of 6) than the academia (median score of 4).

That no other statistically significant differences were found, does not necessarily mean that these differences do not exist. The limited size of the different stakeholder groups in this research means that large differences would need to exist in the dataset to achieve statistical significance (Norman, 2010). Small differences may exist in a larger population of experts, but these are not found in the current expert population due to the sample being too small. Therefore, Beiderbeck et al. (2021a) suggests that each stakeholder group should consist of at least 15 to 20 participants, which is not the case in this research.

4.4 Overview of results

This section shows and discusses the results of the second round of the Delphi survey for both part 1 and part 2.

4.4.1 Results part 1

Table 24 below shows the five types of indicators and the median and mean scores given per criterion. The scale runs from 1 to 7 where a higher score means that the experts agree more that this type of indicator scores well on that criterium. The median score is more robust than the mean, which is still presented for comparison. In most cases the median and mean are close to each other while in several cases the difference is up to 0,5 points.

Table 24 Median, mean, minimum, and maximum scores (7-point scale) on type of indicators per criteria where the Asterix (*) means no consensus

	Validity				**Reliability**				**Sensitivity**				**Understandability**			
	Median	Mean	Min.	Max.	Median	Mean	Min.	Max.	Median	Mean	Min.	Max.	Median	Mean	Min.	Max.
Dutch SPIs	5,0	5,2	3	6	6,0	6,0	5	7	5,0	5,4	2	7	6,0	6,1	5	7
Proximity based SMoS	6,0	5,5	4	7	5,0	4,8	2	6	5,0	5,4	4	6	5,0	4,5	1	6
Kinematic based SMoS	6,0	5,5	4	6	5,0*	5,0*	3	7	5,0	5,0	4	6	5,0*	4,9*	2	6
Engagement of ADAS	6,0	5,5	3	7	5,0*	4,9*	2	7	5,0	4,9	3	7	5,0	5,5	4	7
Driver distraction	5,0	5,4	4	7	4,0*	4,0*	2	6	4,5	4,8	4	6	5,0	5,3	4	7

As the table shows, the median scores given by the experts are quite closely together. Most median scores are 5,0, followed by 6,0. The median scores of 4,0 and 4,5 are present only once each.

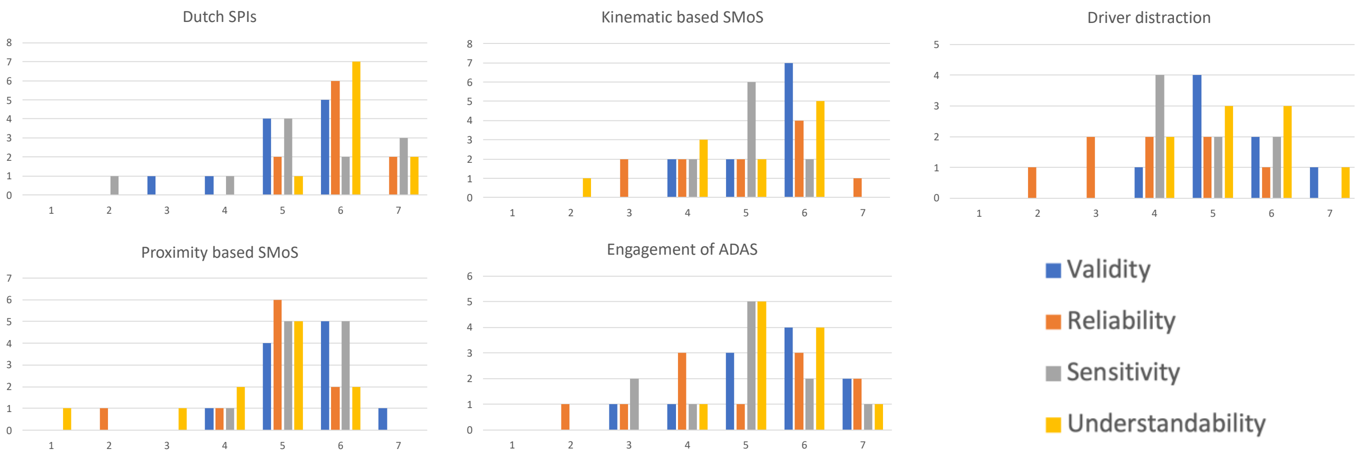


Figure 18 Overview scores per type of indicator and criterium

The Dutch SPIs, like Safe speeds and Safe participants, scores the highest on reliability and understandability of all types of indicators. These scores are also grouped more closely together than in the other types of indicators, as figure 18 shows. Multiple experts mention that this type of indicator represents “basic safety conditions (…) which are pretty easy to collect and analyse”. Especially the example of safe speeds is highlighted by several experts as understandable for policymakers. Additionally, “speed limits have always been and will continue to be important for safety”. It does however “not give a complete picture on safety” as one research expert states: “For example, if one drives at 50kph in a highway, one follows the speed limit, but one also creates a hazard to others.”

Proximity based SMoS like TTC score high on validity which is confirmed by several experts highlighting that it is a “very good precursors of crashes” and “represents critical situations well”. Like the Dutch SPIs, proximity based SMoS are context dependent, but unlike those, it also needs a threshold which comes with additional problems. One academic expert states that “deciding what the threshold should be, has a big impact on what is actually measured” and points out that this might need to change over time as “the introduction of AVs might change what we think of as a critical TTC value.” As can be seen in figure 18, the experts do rate it as quite sensitive, more so than the other type of indicators while it is seen as less understandable for policy makers.

Kinematic based SMoS such as acceleration, deceleration, and swerving did not reach consensus on reliability (IQR = 2, V = 0,27) and on understandability (IQR = 2, V = 0,26). Figure 18 shows that for both of these criteria, the scores given by the experts are spread out over a relatively wide range, 3 to 7 and 2 to 6 respectively. Some experts believe that it can be measured reliably with one expert claiming that: “longitudinal [movement] (acceleration/deceleration) is easier to measure compared to lateral [movement] (swerving)”. Other experts disagree: “Due to the nature of these variables, there are many errors and noises in measuring them via accelerometers, and other devices”.
Interestingly enough, one academic expert claims that “the scientific evidence for the correlation between harsh acceleration and crashes is weak” while a second academic expert says that “there seems to be research that links the behaviours you listed (e.g., hard breakings) with collisions”. This will be further discussed in chapter 5 Discussion.

In the case of engagement of ADAS such as FCW, AEB, and BSW, most experts seem to agree that in essence these engagements signal serious potential conflicts and that they are quite easy to measure. However, the experts also see a variety of practical obstacles, mainly that “a warning from FCW will be different from one OEM to another OEM”. Additionally, some experts have concerns about the reliability (e.g., AEB/FCW false positive) and that the systems change/improve over time, resulting in several experts identifying harmonisation as a key aspect.
Two experts have more fundamental critique on this type of indicator. The activation of these ADAS is based on “some pre-defined thresholds on indicators of safety like TTC or acceleration”. Therefore, they conclude that one might as well just measure those directly.

The suitability of driver distraction as measured by DDAW is only asked about in the second round of the Delphi survey as it was added based on suggestions made by the experts in the first round. Therefore, the stability of consensus could not be measured. Nevertheless, consensus was reached on three out of four criteria with reliability (IQR = 2, V = 0,33) as the exception. This type of indicator has the lowest median scores of all with 4,0 and 4,5 for reliability and sensitivity respectively.
Experts point out the large differences between DDAW systems of different OEMs, much more than those discussed in the previous type of indicator. Additionally, “driver distraction results in specific kind of accidents”. It would make more sense to relate this to specific locations or types of roads than to use this to measure the safety performance of the entire network, as this “gives insights in where drivers are distracted or what circumstances contribute to distraction.

4.4.2 Results part 2

The results of the second part of the Delphi study, on potential barriers for the implementation of a system of measuring traffic safety on network level based on vehicle sensor data, are summarised in figure 19. This figure shows the boxplot for each potential barrier which includes both the minimum and maximum of the answers given, as well as the median and first and third quartile.
Five out of ten potential barriers have a median score of 4 (moderate barrier) and only one is lower, willingness of service providers (2, weak barrier). Willingness of people has a median score of 4,5 while the remaining three have a median score of 5, meaning a somewhat strong barrier. So, most potential barriers identified in the literature review are confirmed by the experts as such, but none of them are

seen as strong or even insurmountable barriers. These barriers are discussed below, with the exception of those who did not reached consensus due to bipolarity and are discussed in the previous section (willingness of suppliers, cybersecurity, willingness of people).

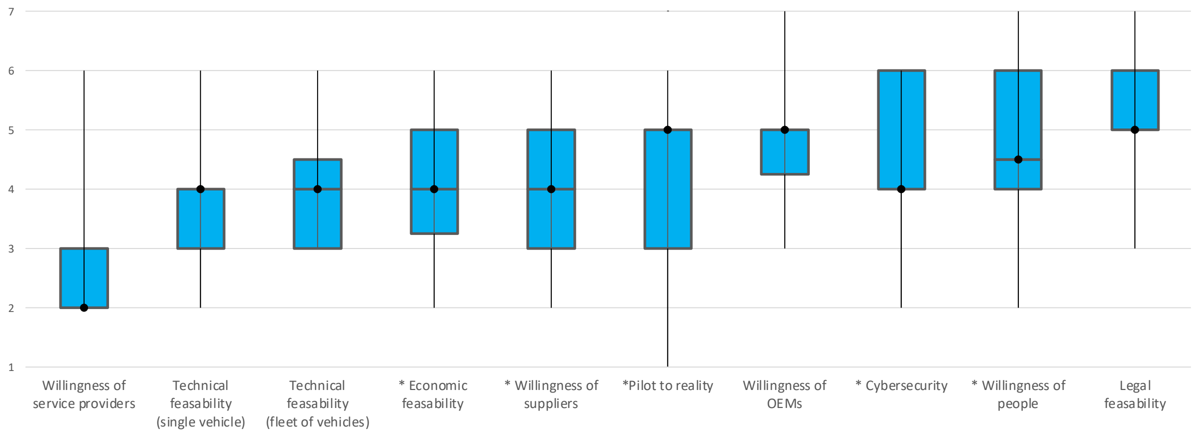


Figure 19 Boxplots potential barrier where the black dot denotes the median score and the Asterix (*) means no consensus. Scale: 1= No barrier at all, 2 = Weak barrier, 3 = Somewhat weak barrier, 4 = Moderate barrier, 5 = Somewhat strong barrier, 6 = Strong barrier, 7 = insurmountable barrier

Willingness of service providers is seen as a weak barrier because “they will see this as new business”. This is supported by an industry expert, claiming that “there are more Service Providers than buyers in the connected car industry, including safety data”. Additional data from service providers may not even be needed according to another experts, disregarding this as a barrier altogether.

The technical feasibility of collecting the data is split in two separate barriers: collecting data within a single vehicle, and extracting, processing, and storing data from a fleet of vehicles. Collecting data within a single vehicle necessary to report any of the in part 1 discussed indicators is seen by the experts as a moderate barrier. Most experts agree that it is technically possible to collect data within a vehicle but that it is “depending on the indicator of interest and the sensor suit present in the vehicle”. Of course, “there are a lot of existing vehicles on the road that are NOT connected in any way, due to their age”. While the experts agree that it should be possible, one government expert refers to experience with a recent pilot using the ExVe/Neutral Server concept and claims it “it works ok but has a lot of flaws and loose ends that have not been solved yet.”
Extracting, processing, and storing the data from a fleet of vehicles from a technical point of view is also seen as a moderate barrier. While several experts point at the required efforts and costs to do this, especially “given the volume and privacy requirements” most experts agree that “it seems very feasible, especially in a country like the Netherlands with great data coverage and a strong network.” The difficulty will lie more in “ensuring the data is comparable between different manufacturers” and in “having all OEM installing the technology (especially for low-cost vehicles)”.
An additional open question on the topic of technical feasibility was asked. The experts were asked if the answers they gave about technical feasibility would be different for any of the specific types of indicators. The experts do not believe that large difference would exist. However, as table 25 shows, four experts point out that combining data from multiple sensors requires more computing power, making it slightly more difficult

Table 25 The experts' answers to a question about differences in their answer on the technical feasibility for any of the discussed specific types of indicators

Question: Would there be a difference in your answers on the technical feasibility for any of the discussed specific types of indicators?	**Number of responses (N=16)**
N/A	4
No	6
I do not know	2
It does not necessarily depend on the type of indicator, but on the sensors that are required to measure that type of indictor. Example: proximity based SMoS (like TTC) needs data on predecessor, meaning that data from different sensors needs to be combined which requires more computing power	4

Economic feasibility is seen as a moderate barrier as well, although no consensus is reached on this potential barrier (IQR = 1,75, V = 0,35) with scores spread quite evenly between 2 and 6. The experts agree that it makes sense for the government to be interested in this data but that "it might be difficult to specify the benefits of network-level safety evaluation to OEMs." Experts do not agree on whether or not a business case for OEMs exists, while this is a crucial factor: "OEMs can stay in a negative business case longer than service providers, but without an eco-system the whole connected industry will struggle". And while the business case improves "if the indicators can be measured with sensors already onboard", an important issue is the role of legislation and the accompanying uncertainty and risks.

Willingness of OEMs is seen as a somewhat strong barrier by the experts. On one hand, several experts state that OEMs are reluctant to share data, as "they are responsible for the safety of the vehicle and its data". On the other hand, it may just be "a point of economic benefit and regulation" as one academic expert states. These economic benefits are disputed by an industry expert: "Making €10M is nice, but that is only a spec for the bigger OEMs". Experts point out that regulation already plays a role with EU Act 2013/886 mandating access to road safety data and the fact that the data is owned by the consumer under the Data Act, meaning that "the OEM can only share this data with consent of the consumer, or on a legal base."

This legal feasibility is seen as a somewhat strong to strong barrier because "legislation can make or break the business". The main relevant legislations discussed by experts are the GDPR and EDPB Guidelines for personal data - in which "lots of vehicle sensor data is considered as personal data and needs thus consent of the owner [to be collected and processed]". Additionally, new legislation is underway with the proposed EU Data Act and future sectoral legislation on vehicle data which is currently open for public consultation.
Several experts point out that while legislation can be changed, this would require strong EU support and a lot of time and effort. And even then, a legal requirement to provide data and insights may need compensation for OEMs to not threaten the market in the long run, as one expert points out.

Going from pilots to reality is seen by some experts as a weak to somewhat weak barrier while most experts rate it as a somewhat strong barrier. The former group argues that pilots are a good way to start and that "anyone doing a pilot (with reason), has the ambition to scale up". The latter sees this differently and sees this as "a big, steep hill to climb". Besides the fact that all barriers discussed needs to be overcome, it would require harmonization of data and/or interfaces of different OEMs which "is difficult and time consuming".

5. Discussion

Vehicle automation has been seen as a promising solution to many problems including traffic safety and is as such a popular topic. However, there is limited insight into the real-world effects of vehicle automation on traffic safety. At the same time, measuring traffic safety is an essential step in monitoring and evaluating traffic safety policies. The current practice uses traffic deaths and severe injuries as indicators for traffic safety at network level which has several shortcomings, mainly that it is a reactive approach.

5.1 Added value of this research

This study explores if a more proactive approach can be viable where data is collected by vehicles equipped with ADAS. In doing so, it contributes to the existing body of knowledge in several ways. Firstly, it looks into what type of proactive indicators would be suitable to apply in the specific context of measuring traffic safety at network level. Given the benefits of leading indicators, many studies exist on this topic but often in an isolated view or focussed on the development of a single leading indicator (Arun, Haque, Bhaskar, et al., 2021). Based on a literature review, this study provides an overview of different types of indicators that could be based on vehicle sensor data for a specific and practical goal: measuring traffic safety at network level.

Secondly, this research discusses to what extent vehicle sensor data can be used to measure traffic safety. Less research exists on this topic. By studying academic literature, policy documents and evaluations of pilots involving vehicle sensor data, this study has collected and discussed several potential barriers for using vehicle sensor data in practice. As far as the author could find, no comprehensive discussion of overview of such potential barriers exists in the current academic literature.

But thirdly and most importantly, this research evaluates both several types of indicators and potential barriers for using vehicle sensor data to measure traffic safety at network level at the same time using a Delphi survey. It therefore allows for a fair comparison between different types of indicators for the application in this context. Additionally, in existing literature potential barriers for collecting and using vehicle sensor data is only discussed in a general and limited fashion. This study provides an overview of multiple potential barriers and then builds on it by letting experts evaluates the potential barriers. Because this is done in the same survey by the same experts, it can give an insight into how the potential barriers relate to each other in terms of size.

5.2 Discussion on results

Limited differences are found in the median scores given by the experts between the types of indicators and the criteria (table 25, p55). Almost all types of indicators score quite high at all criteria with most median scores being five or six on a scale of seven.

This could be because not all types of indicators identified in the literature review were presented to the experts, but only those that were seen as the most promising types of indicators. Additionally, the types of indicators could also be too similar, especially Proximity based SMoS, Kinematic based SMoS or Engagement of ADAS. If this is the case, only limited differences are present between these types of indicators, and all could potentially be used to measure traffic safety at network level.

It could also be the case that only limited differences are found for other reasons. The types of indicators could also appear to be too similar to the experts, either due to unclear questions or to lack of expertise of the experts (to be discussed below). In either case, further research is required which will be discussed under recommendations.

The results regarding potential barriers for collecting and using vehicle sensor data are clearer. Most of the potential barriers initially identified in the literature review are confirmed as such by the experts. However, none of the barriers are rated as insurmountable which provides perspective for

implementing any system that uses vehicle sensor data in practice. So, does that mean that it would be possible to use vehicle sensor data to measure traffic safety?
The experts agree that, while some challenges need to be overcome, from a technical point of view it should be doable to both collect data within a vehicle and to share it through an ExVe/neutral server as discussed in section 2.5. Both service providers and suppliers to the vehicle manufacturing industry should be willing to cooperate based on this research as they see it as new business.
There are however four sizable barriers that need to be overcome of which legal feasibility is the largest. Privacy regulations are discussed by all experts because these dictate if and in what form data can be collected and extracted from vehicles. Several experts also discuss the uncertainty surrounding upcoming regulation. The EC has proposed the Data Act which aims to regulate the access and use of all data generated in the EU across all economic sectors (European Commission, n.d.-b), and a public consultation has just finished on additional legislative measures concerning access to in-vehicle generated data for vehicle-related and mobility services (European Commission, n.d.-a). While the aim of this additional sectoral legislation is to help implement the Data Act and aims at "creating benefits from different economic, social and environmental perspectives" it is also focused on the protection of intellectual property of OEMs and personal data of consumers (European Commission, n.d.-a). At this point, it is not known how the additional sectoral legislation would look like, but it has the potential to have a significant impact on any effort to use vehicle sensor data for measuring traffic safety. This will determine the framework and amount of room available to use vehicle sensor data to measure traffic safety, if any at all.
The second important barrier is the willingness of OEMs to cooperate. On one hand, the experts think that OEMs are reluctant to share any data because of market competition or because of costs consideration. On the other hand, experts point out that costs might be relatively low since sensors are already present in most new vehicles, and OEMs could be (partly) compensated for remaining costs of data extraction and processing. Additionally, regulation could force OEMs to share certain types of data. This is not without precedent as OEMs are already obliged to share SRTI data and data from eCall or the Event Data Recorder (EDR) in the case of an accident (see section 2.4).
Although the experts do not reach consensus on willingness of people to cooperate, it can be established that this is third barrier that needs to be overcome, especially as the new Data Act will allow people more control over their own data and who can access it (European Commission, n.d.-b). Experts believe that people may be unwilling to share their data for several reasons. From a privacy perspective people can be opposed to sharing their data which will especially be the case if the data could be linked to traffic violations. Although it is unlikely that this would even be possible in practice because of privacy regulations, it is the perception of people that matters. The willingness of people to share data could depend on what personal benefits people gain from it, experts point out. This could range from quite abstract benefits as contributing to traffic safety to individual reports or even insurance benefits, as one expert states. Important to note is that in both the pilot Kia Insights from vehicle data and the Road Monitor a large share of the people invited were willing to cooperate.
The fourth barrier that needs to be addressed according to all experts is cybersecurity. And while all agree that it is an important topic, disagreement remains on how far the risks of cyberattacks or data leaks can be contained. Further research efforts should be made to better understand this, especially as these risks could tie into other important factors such as willingness of people and OEMs to participate.

Next to these four barriers that need to be overcome, two points are brought up by most experts throughout the Delphi survey. The first point is that there are no standards for ADAS so an OEM can build and define it in the way they want to. This means that a warning from FCW will be different from one OEM to another OEM. Using an FCW as a proxy for a traffic conflict could therefore mean that a vehicle from one OEM would signal this potentially dangerous situation while another vehicle from a different OEM would not. The same goes for the thresholds for proximity based or kinematic SMoS. Experts therefore point at the need for standardization and harmonization.

The second point experts make is that all vehicle sensor data comes from relatively new (roughly less than four years old as one government expert estimates) or high-end vehicles that have these sensors. This is only a (currently small) portion of all vehicles on the roads in the Netherlands. Does this modern subset of all vehicles represent the entire fleet of vehicles in a fair way? Related to this is the suggestion of a government expert to only select a subset of all potential vehicles any, as monitoring all vehicles may be very time and cost intensive. This would require so assumptions for extrapolation, which in turn could lead to measurement error in data. An important sidenote that must be made here is that the proportion of vehicles equipped with at the sensors needed to measure the types of indicators discussed in the Delphi survey, will increase steadily in the coming years as these are necessary for those ADAS that become compulsory in all new vehicles in 2024.

5.3 Wider context

No single indicator can tell the entire story of safety. It is important to stress that any type of indicator discussed here is aimed to supplement the current practice of measuring traffic at network level with severe injuries and fatalities.

In proposing to use leading indicators based on vehicle sensor data it is important to keep in mind that only a portion of all traffic is able to collect data. Motorised vehicles such as cars and trucks can have sensors and forms of automation that are able to collect data which could eventually be used to measure traffic safety, but other modes of transport like bicycles, pedestrians, or motorcycles to some extent cannot do that. As a result, it will be difficult to directly measure the traffic safety of these road users. This is not unlike the current situation, but it does come with a bigger problem. Evaluating traffic safety policies based on vehicle sensor data will favour those policies that increase the traffic safety of motorised traffic. It therefore risks that this increased traffic safety could come at the expense of others not represented in these indicators and as a result make it less safe for vulnerable road users. If it would be possible from both a legal and a technical perspective to measure any of the proposed indicator per road type (highway, provincial roads, urban streets, etc), this would be less of an issue. These indictors could then be treated more cautiously on those road types where many vulnerable road users are present.

Furthermore, it be noted that the assumption behind using proximity based SMoS, kinematic based SMoS, and engagement of ADAS as indicators is that these signal traffic conflicts which are proxies for accidents. And while traffic conflicts are regarded as suitable proxies for accidents (Arun, Haque, Bhaskar, et al., 2021), they still have their limitations. According to Arun, Haque, Bhaskar, et al. (2021) there are two main theoretical models to explain the relationship between crashes and traffic conflicts: statistical association and causal relationship. Both models would work as the theoretical basis behind the idea of this study, but it is important to note that additional validation is needed to confirm the relation between traffic conflicts and crashes, as is pointed out by several experts and by Arun, Haque, Bhaskar, et al. (2021) and Arun, Haque, Washington, et al. (2021).

The types of indicators evaluated in the Delphi survey may also measure different types of traffic conflicts. Proximity based SMoS based on TTC can only measure traffic conflicts when two vehicles move in the same direction and ignore evasive action, measuring only rear-ended accidents (Arun, Haque, Bhaskar, et al., 2021). Kinematic SMoS do not have this limitation. However, kinematic SMoS will miss any dangerous traffic conflict where no evasive action is taken, either by swerving, braking, or acceleration. Additionally, it may be reasonable to assume that thresholds for kinematic SMoS based on strong accelerations or decelerations should be different for different types of vehicles and perhaps speeds. A certain acceleration could be reasonable when entering a highway while the same acceleration would be dangerous in a city centre. Finally, using engagement of ADAS such as FCW is limited by how the OEM has implemented the FCW which could depend on the speed of the vehicle.

It can be concluded that there is reason to believe that not all types of indicators discussed here are suited for every road type, or that varying thresholds should be used. Taking a relative approach to thresholds may be an interesting direction as well, as discussed in section 2.3. Further research should

be conducted to identify and validate specific thresholds for specific types of indicators for specific road types. Depending on this, it may be a good approach to use multiple types of indicators as they can identify different types of traffic conflicts.
The Dutch SPIs and Driver distraction based on DDAW do not signal traffic conflicts but underlaying risks associated with accidents such as speeding and distracted driving. These give insight into how these risks exist in practice and can be a starting point for developing specific policies to target these specific risks.

Using leading indicators instead of the current lagging indicators is a step forward as it solves some of the problems discussed in section 2.2. It does however not solve them all. The ethical issue of having to wait for sufficient dangerous traffic situations to happen before measures can be justified to improve traffic safety still exists. It may lead to quicker policy responses to dangerous developments in traffic than with lagging indicators, but it still requires those dangerous events to happen.
New issues may arise as well. If OEMs are going to collect more data and share that data with the government or potentially with other parties with the aim to increase traffic safety, special attention should be paid to other public interests such as privacy. Nissenbaum (2004) has developed the notion of privacy as a contextual integrity, which means that any gathering of data should be appropriate to a specific context, and not be used for anything else to prevent risking “driving to the panopticon[9] ” (Reiman, 1995). Zimmer (2005) builds upon this in the specific context of intelligent vehicles and points at the importance of designing technology in such a way that privacy is insured by design. And while the notion of privacy by design has been incorporated in EU regulation (Onderzoeksraad voor Veiligheid, 2019), fear for the large scale gathering of data by the government or companies remain, for example in news media (Hofman, 2021; Modderkolk, 2020; Teeffelen, 2022) and in the Dutch parliament (van Huffelen, 2022; Van Nieuwenhuizen, 2019). It is therefore important that any future discussion or research should not just approach this research from a purely “technical” point of view focussing on how and under what conditions the large-scale collection of data can take place, but also on whether or not this should be desirable and appropriate. Ideally, these discussions should not just be confined to for example academic researchers but be part of a larger political and societal debate, not only specifically on the collection of vehicle sensor data alone, but in a wider context.

5.4 Methodological limitations

Much is written on the validity and reliability of the Delphi method, see for example Förster & von der Gracht (2014), Hasson et al. (2000), Landeta (2006), and von der Gracht (2012). However, as Belton et al. (2019) conclude, Delphi method is not a standardized technique and its form can vary between studies. It is therefore of limited use to discuss the reliability or validity in general terms, but more so to focus on the individual execution (Belton et al., 2019). This study has tried to execute the Delphi method in the best way possible, following advice from academic literature on the Delphi method (see chapter 3).
Belton et al. (2019) discusses that reliability (the ability to produce similar results when the study is repeated) is difficult to establish for a Delphi study, although there have been studies in the past that found similar results across two separate panels. One type of reliability that could be tested is the intra-respondent test-retest reliability where experts are asked the same question twice in different wordings (Belton et al., 2019). This comes however at the drawback of a much lengthier survey and may risk a higher expert dropout, which is the reason why this is not applied in this research.

[9] The panopticon is a specific type of prison design where all cells can be surveilled from one point, used by the French philosopher Michel Foucault as an metaphor for the mechanisms of large-scale social control (Reiman, 1995)

Content validity (the extent to which a survey measures what it is supposed to measure) and what Belton et al. (2019) calls external validity (the extent to which findings can be generalized to a wider population) can be enhanced by making sure that the experts involved form a representative, heterogenous, and large enough sample of relevant experts, that are motivated and have enough opportunity to give their input (Belton et al., 2019; Okoli & Pawlowski, 2004). This study has aimed to reach these goals by using the expert continuum model to help capture a wide and heterogenous range of experts involved in the topic. Additionally, criteria are used to establish a minimum level of expertise. Experts were also given the opportunity to give both quantitative as qualitative input at each question, as well as several open-ended questions to give the experts the chance to give any input they wanted. The syntax analysis (table 16, p46) showed that they used these opportunities, showing a high level of commitment.
Based on ranges suggested in academic literature (5 to 20, 7 to 15, 10 to 18, 15 to 20, see section 3.1) the goal of this study was to include 10 to 15 experts which was reached with eleven experts completing both rounds. It is however important to note that given the wide range the backgrounds of experts, the research would have benefited from more experts, especially from the government and industry. This would either have solidified the current opinions or may have added additional input. Doing this would increase the external validity.
Moreover, the Delphi survey focused on two topics: types of indicators for measuring traffic safety and potential barriers for using vehicle sensor data to do so. It is acknowledged that while a relation is made between these topics in this study, it still requires separate knowledge that not all experts may have. In turn, this could have influenced the results. This is partly countered by adding the option "I do not know" at each question, which experts also used throughout the survey, showing their awareness of their own lack of knowledge. One expert even only answered questions in the second part of the survey, acknowledging the limits of its knowledge. Additionally, the design of a Delphi study helps in this regard, by letting experts reevaluate their own opinions in the second rounds based on additional arguments brought up by other experts. It can be reasoned that experts with less knowledge on a specific question would be more likely to reevaluate their answer from the first round. However, it can be assumed that more experts would have helped to alleviate this issue.

The external validity could also be increased by better assessing the expertise of the involved experts. This could have been done with self-rating in an intra-individual way as proposed by Ward et al. (2002) or by using deep surface variables as proposed by Spickermann et al. (2014). Both were not viable in this study due to a lack of in-depth psychologic knowledge of the author, as well as practical and time constraints. A third way in which this issue could be addressed is by splitting the parts into two separate Delphi surveys with two separate panels. This would have as additional benefit that the survey would be shorter, which could either make it easier for experts to answer it or which could be used to ask more or more in-depth questions. It may also have given room to increase the intra-respondent test-retest reliability in the way as suggested by Belton et al. (2019) and discussed above. However, this would have required substantially more effort to recruit sufficient experts for two panels. It would also have required more time and effort to develop and execute both surveys.
Lastly, the quantitative data that followed from the Delphi survey is of ordinal scale, as the data can be categorised and ranked but it cannot be presumed that the intervals between values are equal (Norman, 2010). Therefore, in the papers of Beiderbeck et al. (2021a) and Belton et al. (2019), which provide a practical guide of developing a Delphi survey, the usage of only non-parametric statistics and tests is recommended. However, an extensive academic discussion exists on this topic with where researchers like Argyrous (1997) and Jamieson (2004) take this point of view while others such as Norman (2010) and Shields et al. (1987) argue that parametric tests can be used as well. Some argue that the data can be treated as being of interval level (e.g. Shields et al., 1987) while others (e.g. Norman, 2010) do not contest that answers on a Likert scale are of ordinal level, but rather argue that parametric tests are robust enough to be used in analysis. In practice, both are used, although not always based on a clear consideration (von der Gracht, 2012). This research followed to strict interpretation of Beiderbeck et al. (2021a) and Belton et al. (2019) to stay on the cautious side. Other

effects may have been found using a less strict approach in for example statistical differences between the two rounds or between stakeholder groups.

5.5 Recommendations

The aim of this research was to explore if data collected by vehicles equipped with ADAS can be used to measure traffic safety at network level in a more proactive approach than the current approach. As demonstrated above, this research has made several contributions towards the academic knowledge of this topic. However, as an explorative research, it also serves as a starting point for further analysis. This discussion highlights several areas that require further efforts, both in (academic) research (1) as in more practical oriented ways by the Dutch ministry of Infrastructure and Water Management (2).

1. The results of the Delphi study seem to indicate that all of the four initially included types of indicators could potentially be used to measure traffic safety at network level. However, more scientific research is required to understand:
    a. Why limited differences are found in the Delphi survey between the different types of indicators.
    b. Which types of indicators signal which types of accidents. This would on one hand allow for a better understanding of their meaning if they would be measured in reality, and on the other hand allow for a more conscious and informed choice on what (combination of) type of indicators should be measured.
    c. How valid the relationship is between traffic conflicts measured by these types of indicators and crashes, on different road types.
    d. Which thresholds would be suitable for any type of indicator per road type.
2. Next to academic research, more applied research should be organised by the ministry of Infrastructure and Water Management on the two topics below. This research can be conducted by various actors, such as the SWOV, TNO, or consultancy firms, possible in combination with academic researchers.
    a. Differences between vehicle sensors and ADAS of different OEMs should be understood if the indicators are based on vehicle sensor data. A way to deal with the differences should be developed or a push for a standard and harmonisation should be made. While it may be difficult or even undesirable to harmonise sensors and ADAS, any indicator should be clearly defined so all OEMs can report the same measure.
    b. Related to this is the need for research on how well the share of vehicles that are equipped with such sensors represent the entire vehicle fleet in the Netherlands. This determines how, or how well, this subset of the vehicle fleet can represent all vehicles.
3. The following actions need to be undertaken by the Dutch ministry of Infrastructure and Water Management according to this research, to address the barriers for using vehicle sensor data to measure traffic safety:
    a. A legal analysis should be developed to identify the legal room for using vehicle sensor data for measuring traffic safety. This should take the proposed Data Act into account, as well as the upcoming additional sectoral legislation on this topic. This would require some time, as these are in development at the moment.
    b. A discussion should be opened with OEMs (either directly or through ACEA) to better understand how willing they are to cooperate in a system where they will be required to collect and share vehicle sensor data, and under what conditions. This will determine how such a system would be able to exist (economic benefit for the OEMs, because of the OEMs reputations, or if they would need to be forced by legislation, etc.).
    c. Included in this discussion should be the issue of cybersecurity and in how far the risks related to cybersecurity can be contained. This may require not just the input of the OEMs but also the internal knowledge of the ministry of Infrastructure and Water

Management on this topic, as well as external input from consultancy firms of academic researchers.

d. In any implementation of such as system, it is important to take the consumer and its privacy seriously. Allow them to give informed consent of sharing their data and clearly communicate to what end it will be used. Make clear who has access to what data, so it is also clear what can or cannot be done with the data. This will help to increase the willingness of people to cooperate.

6. Conclusion

This research has explored the idea of using data collected by vehicles to measure traffic safety at network level. Much research is carried out into vehicle automation and its effect on traffic safety, although the size of the safety benefits of a combination of ADAS in practice remains unclear. At the same time, the current practice of using fatalities and severe injuries as indicators for measuring traffic safety at network level has several problems, mainly that it is a reactive approach that only measures severe crashes which are rare. Being able to monitor the effects of any policy intervention, and of vehicle automation specifically, on traffic safety is important to evaluate current policies and develop new policies to address existing issues. This has given rise to the idea of using data collected by the vehicles themselves to measure traffic safety at network level. This vehicle sensor data could allow for using proactive indicators that could measure traffic conflicts, traffic interaction that are dangerous but do not necessarily result in an accident. As these traffic conflicts are much more common, traffic safety could be measured in a more proactive way, solving some of the problems of the current practice.

This has led to the following research question: (How) can data from sensors of vehicles equipped with ADAS be used to measure traffic safety at network level? There are two aspects to this question. First comes the theoretical aspect of how traffic safety can be measured with indicators based on vehicle sensor data. What should be measured? The second aspect is more practically focussed and discusses whether or not it is feasible to collect, process and use vehicle sensor data in practice.

To identify potential feasible and suitable types of indicators based on vehicle sensor data for measuring traffic safety at network level, a literature review is carried out. As this is a relatively novel idea, limited academic research exists on this topic. Therefore, existing indicators for measuring traffic safety are discussed that could be adapted to use vehicle sensor data as a source. Expert judgement in the form of a Delphi study is used to assess how suitable those indicators would be for measuring traffic safety at network level. The four most promising types of indicator identified in the literature review were evaluated both in a quantitative and qualitative way on four criteria over two rounds by a panel of experts, plus one additional type of indicator suggested by several of the experts. The panel consisted of eleven experts from with various backgrounds, ranging from government to academia to industry to ensure sufficient heterogeneity and varying perspectives.
The experts rated four of the five types of indicators in a favourable way, with limited differences in median scores given. This could mean that all these types of indicators would be suitable and feasible to measure traffic safety at network level. That limited differences are found could however also be due to the study design or the expert panel. Further research is required to assess whichever is the case.
The same combination of literature research and Delphi study is applied to identify barriers for collecting and using vehicle sensor data in practice. From this can be concluded that there are no insurmountable barriers that prevent using vehicle sensor data in practice. However, there are four sizable barriers that need to be addressed: legal feasibility, willingness of OEMs to cooperate as well as people's willingness to cooperate and cybersecurity. Any usage of vehicle sensor data must happen within the regulatory framework set by privacy regulations. The new Data Act proposed by the EC will give consumers more control over their own data, increasing the importance of the willingness of people to cooperate. Upcoming additional sectoral legislation tied to this Data Act may have significant impact on the conditions under which the collection of vehicle sensor data may take place although at this point it is unknown how this legislation will look like. Additionally, the questions of how well the subset of vehicles equipped with the sensors necessary to collect the vehicle sensor data represents the entire vehicle fleet needs to be answered, as well as how to deal with differences between systems of different OEMs.

So, the answer to the main research question - (how) can data from sensors of vehicles equipped with ADAS be used to measure traffic safety at network level- is that it depends. If certain conditions are met, it should be possible to collect and use vehicle sensor data to measure traffic safety. The main condition is that there should be legal room available under the proposed Data Act and on the upcoming additional sectoral legislation. It is therefore recommended to conduct a legal analysis when the regulations are published.

It is important to note that the aim of this research was not to give a conclusive answer on how vehicle sensor data can be used to measure traffic safety, but to explore this novel idea and see if it would be possible at all. If the legal analysis of the upcoming regulations gives room for collecting vehicle sensor data, this research has proven that there are sufficient possibilities to warrant further research and discussions into using vehicle sensor data for measuring traffic safety.

Bibliography

Aarts, L. T. (2018). Prestatie-indicatoren voor verkeersveiligheid (SPI's); Overzicht van beschikbare kennis over SPI's als basis voor risicogestuurd beleid (p. 52). SWOV.

Aarts, L., Wijlhuizen, G. J., Gebhard, S., Goldenbeld, Ch., Decae, R., Bos, N., Bijleveld, F., Mons, C., & Hoekstra, T. (2021). De Staat van de Verkeersveiligheid 2021; Doelstellingen voor 2020 definitief niet gehaald – hoe nu verder? (p. 28). SWOV.

Ackermann, J., Reinhold, M., Christiaens, S., & Luxen, M. (2019). Connected and autonomous vehicles interface analysis. FEV Consulting. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjkpKzKqejzAhVHDewKHXr5BXUQFnoECAYQAQ&url=https%3A%2F%2Fwww.at-north.nl%2Fapp%2Fuploads%2F2019%2F05%2F2019-05-08_CAV_Interface_Study_F_shared.pdf&usg=AOvVaw09hgbPKubxyZFxYsmM_8FN

ADAS Alliantie. (2019). ADAS Convenant: Verhogen veilig gebruik ADAS door het bevorderen van doorontwikkeling, bekendheid en aanschaf. https://www.adasalliantie.nl/

Amundsen, F. H., & Hyden, C. (1977). Proceedings of first workshop on traffic conflicts. Oslo, TTI, Oslo, Norway and LTH Lund, Sweden, 78.

Argyrous, G. (1997). Statistics for social research. Macmillan International Higher Education.

Armstrong, J., & Overton, T. (1977). Estimating Nonresponse Bias in Mail Survey. Journal of Marketing Research, 14, 396–402. https://doi.org/10.1177/002224377701400320

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accident Analysis & Prevention, 153, 106016. https://doi.org/10.1016/j.aap.2021.106016

Arun, A., Haque, Md. M., Washington, S., Sayed, T., & Mannering, F. (2021). A systematic review of traffic conflict-based safety measures with a focus on application context. Analytic Methods in Accident Research, 32, 100185. https://doi.org/10.1016/j.amar.2021.100185

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021a). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. MethodsX, 8, 101401. https://doi.org/10.1016/j.mex.2021.101401

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021b). The impact of COVID-19 on the European football ecosystem – A Delphi-based scenario analysis. Technological Forecasting and Social Change, 165, 120577. https://doi.org/10.1016/j.techfore.2021.120577

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. Technological Forecasting and Social Change, 147, 72–82. https://doi.org/10.1016/j.techfore.2019.07.002

Bhavsar, P., Das, P., Paugh, M., Dey, K., & Chowdhury, M. (2017). Risk Analysis of Autonomous Vehicles in Mixed Traffic Streams. Transportation Research Record, 2625(1), 51–61. https://doi.org/10.3141/2625-06

Blumenthal, M. S., Fraade-Blanar, L., Best, R., & Irwin, J. L. (2020). Safe Enough: Approaches to Assessing Acceptable Safety for Automated Vehicles. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA569-1.html

Boelhouwer, A., van den Beukel, A. P., van der Voort, M. C., Hottentot, C., de Wit, R. Q., & Martens, M. H. (2020). How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands. Transportation Research Interdisciplinary Perspectives, 4, 100103. https://doi.org/10.1016/j.trip.2020.100103

Boggs, A. M., Wali, B., & Khattak, A. J. (2020). Exploratory analysis of automated vehicle crashes in California: A text analytics & hierarchical Bayesian heterogeneity-based approach. Accident Analysis & Prevention, 135, 105354. https://doi.org/10.1016/j.aap.2019.105354

Bourdeaud'hui, J., Devriese, J., Michielsen, K., & Vandenbossche, J. (2020). Evaluatierapport Talking Traffic. Be-Mobile NV.

CBS. (n.d.). Verkeersdoden [Webpagina]. Centraal Bureau voor de Statistiek. Retrieved 19 August 2021, from https://www.cbs.nl/nl-nl/nieuws/2021/15/610-verkeersdoden-in-2020/verkeersdoden

Chang, A., Saunier, N., & Laureshyn, A. (2017). Proactive methods for road safety analysis. SAE Technical Paper.

Chin, H.-C., & Quek, S.-T. (1997). Measurement of traffic conflicts. Safety Science, 26(3), 169–185. https://doi.org/10.1016/S0925-7535(97)00041-6

Croce, V., Wöber, K., & Kester, J. (2016). Expert identification and calibration for collective forecasting tasks. Tourism Economics, 22(5), 979–994. https://doi.org/10.5367/te.2015.0472

Dajani, J. S., Sincoff, M. Z., & Talley, W. K. (1979). Stability and agreement criteria for the termination of Delphi studies. Technological Forecasting and Social Change, 13(1), 83–90. https://doi.org/10.1016/0040-1625(79)90007-6

Data for Road Safety. (2021). Privacy Statement- Data for Road Safety. https://www.dataforroadsafety.eu/images/Documenten/20210706_Privacy_Statement._DFRS_ecosystem.pdf

de Loë, R. C., Melnychuk, N., Murray, D., & Plummer, R. (2016). Advancing the State of Policy Delphi Practice: A Systematic Review Evaluating Methodological Evolution, Innovation, and Opportunities. Technological Forecasting and Social Change, 104, 78–88. https://doi.org/10.1016/j.techfore.2015.12.009

De Vet, E., Brug, J., Nooijer, J., Dijkstra, A., & Vries, N. (2005). Determinants of forward stage transitions: A Delphi study. Health Education Research, 20, 195–205. https://doi.org/10.1093/her/cyg111

Devaney, L., & Henchion, M. (2018). Who is a Delphi 'expert'? Reflections on a bioeconomy expert selection procedure from Ireland. Futures, 99, 45–55. https://doi.org/10.1016/j.futures.2018.03.017

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. Journal of Clinical Epidemiology, 67(4), 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Díaz de Rada, V. (2005). Measure and control of non-response in a mail survey. European Journal of Marketing, 39(1/2), 16–32. https://doi.org/10.1108/03090560510571990

Directive (EU) 2007/46/EG, (2007). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32007L0046

Donohoe, H. M., & Needham, R. D. (2009). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. International Journal of Tourism Research, 11(5), 415–437. https://doi.org/10.1002/jtr.709

Ecorys. (2020). Onderzoek naar het delen van voertuigdata en interfaces [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2020/03/23/rapport-delen-voertuigdata-en-interfaces-ecorys

English, J. M., & Kernan, G. L. (1976). The prediction of air travel and aircraft technology to the year 2000 using the Delphi method. Transportation Research, 10(1), 1–8. https://doi.org/10.1016/0041-1647(76)90094-0

ERTRAC. (2019). Connected Automated Driving Roadmap. ERTRAC. https://www.ertrac.org/index.php?page=ertrac-roadmap

ETSC. (2001). Transport Safety Performance Indicators. European Transport Safety Council. https://etsc.eu/transport-safety-performance-indicators/

European Commission. (n.d.-a). Access to vehicle data, functions and resources. Have Your Say. Retrieved 21 August 2022, from https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13180-Access-to-vehicle-data-functions-and-resources/F_en

European Commission. (n.d.-b). Data Act: Measures for a fair and innovative data economy [Text]. European Commission - European Commission. Retrieved 21 September 2022, from https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113

European Commission. (n.d.-c). Register of Commission expert groups and other similar entities. Retrieved 2 November 2021, from https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?do=groupDetail.groupDetail&groupID=3188

European Commission. (2011). White Paper: Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52011DC0144

European Commission. (2019). EU Road Safety Policy Framework 2021-2030—Next steps towards 'Vision Zero'. https://op.europa.eu/nl/publication-detail/-/publication/d7ee4b58-4bc5-11ea-8aa5-01aa75ed71a1

Regulation EU 2019/2144, (2019). https://eur-lex.europa.eu/eli/reg/2019/2144/oj

Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. PLOS ONE, 12(9), e0184952. https://doi.org/10.1371/journal.pone.0184952

FESTA. (2018). FESTA Handbook (Version 7). https://connectedautomateddriving.eu/wp-content/uploads/2019/01/FEST-Handbook-Version-7.pdf\

Förster, B., & von der Gracht, H. (2014). Assessing Delphi panel composition for strategic foresight—A comparison of panels based on company-internal and external participants. Technological Forecasting and Social Change, 84, 215–229. https://doi.org/10.1016/j.techfore.2013.07.012

Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). Measuring Automated Vehicle Safety: Forging a Framework. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2662.html

Frewer, L. J., Fischer, A. R. H., Wentholt, M. T. A., Marvin, H. J. P., Ooms, B. W., Coles, D., & Rowe, G. (2011). The use of Delphi methodology in agrifood policy development: Some lessons learned. Technological Forecasting and Social Change, 78(9), 1514–1525. https://doi.org/10.1016/j.techfore.2011.05.005

Fridman, L., Brown, D., Glazer, M., Angell, W., Dodd, S., Jenik, B., Terwilliger, J., Patsekin, A., Kindelsberger, J., Ding, L., Seaman, S., Mehler, A., Sipperley, A., Pettinato, A., Seppelt, B., Angell, L., Mehler, B., & Reimer, B. (2019). MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation. IEEE Access, PP, 1–1. https://doi.org/10.1109/ACCESS.2019.2926040

Fritschy, C., & Spinler, S. (2019). The impact of autonomous trucks on business models in the automotive and logistics industry–a Delphi-based scenario study. Technological Forecasting and Social Change, 148, 119736. https://doi.org/10.1016/j.techfore.2019.119736

Gettman, D., Pu, L., Sayed, T., & Shelby, S. G. (2008). Surrogate Safety Assessment Model and Validation: Final Report (dot:39210). FHWA-HRT-08-051. https://rosap.ntl.bts.gov/view/dot/39210

Gordon, T., & Pease, A. (2006). RT Delphi: An efficient, "round-less" almost real time Delphi method. Technological Forecasting and Social Change, 73(4), 321–333. https://doi.org/10.1016/j.techfore.2005.09.005

Grimberg, E., Botzer, A., & Musicant, O. (2020). Smartphones vs. in-vehicle data acquisition systems as tools for naturalistic driving studies: A comparative review. Safety Science, 131, 104917. https://doi.org/10.1016/j.ssci.2020.104917

Groenendijk, J. (2021). Verantwoording Praktijkproef Amsterdam. TwynstraGudde. https://praktijkproefamsterdam.nl/database/documenten/download/595

Groenendijk, J., de Kievit, M., van Min, E., & van Oosten, M. (2021). SOCRATES2.0 final evaluation. TwynstraGudde, Goudappel, Arup. https://socrates2.org/library

Guiting, T., & Striekwold, P. (2021, September 23). Typegoedkeuring, een toelichting.

Harms, I. M., Bingen, L., & Steffens, J. (2020). Addressing the awareness gap: A combined survey and vehicle registration analysis to assess car owners' usage of ADAS in fleets. Transportation Research Part A: Policy and Practice, 134, 65–77. https://doi.org/10.1016/j.tra.2020.01.018

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. Journal of Advanced Nursing, 32(4), 1008–1015. Scopus. https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x

Henkens, N., Sloot, M., & van Rij, M. (2020). Evaluation of the PoC 'Data for Road Safety' (No. SWNL0266365). Sweco. https://www.dataforroadsafety.eu/images/Documenten/Microsoft_Word_-_PoC_DTF_-_monitor__evaluation_report_Sweco.pdf

Hofman, E. (2021, June 9). Een minister van Digitale Zaken moet de datahonger van de overheid in de gaten houden. De Groene Amsterdammer. https://www.groene.nl/artikel/wie-stopt-de-datahonger

Hsu, C.-C., & Sandford, B. (2007). The Delphi Technique: Making Sense Of Consensus. Practical Assessment, Research and Evaluation, 12.

Hudson, darren, Seah, L.-H., Hite, D., & Haab, T. (2004). Telephone Presurveys, Self-Selection, and Non-Response Bias to Mail and Internet Surveys in Economic Research. Applied Economics Letters, 11, 237–240. https://doi.org/10.1080/13504850410001674876

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education, 38(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jang, J., Ko, J., Park, J., Oh, C., & Kim, S. (2020). Identification of safety benefits by inter-vehicle crash risk analysis using connected vehicle systems data on Korean freeways. Accident Analysis & Prevention, 144, 105675. https://doi.org/10.1016/j.aap.2020.105675

Johnsson, C., Laureshyn, A., & Dágostino, C. (2021). A relative approach to the validation of surrogate measures of safety. Accident Analysis & Prevention, 161, 106350. https://doi.org/10.1016/j.aap.2021.106350

Johnsson, C., Laureshyn, A., & De Ceunynck, T. (2018). In search of surrogate safety indicators for vulnerable road users: A review of surrogate safety indicators. Transport Reviews, 38(6), 765–785. https://doi.org/10.1080/01441647.2018.1442888

Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice, 94, 182–193. https://doi.org/10.1016/j.tra.2016.09.010

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. Journal of Advanced Nursing, 53(2), 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x

Kia. (2021). CONCEPT REPORT: Insigts from vehicle data. Kia Nederland, Ministry of Infrastructure and Watermanagement.

Kia Nederland. (2021, November 25). Kia verkent inzichten in voertuigdata met Ministerie van IenW. Kia verkent inzichten in voertuigdata met Ministerie van IenW. https://nieuws.kia.nl/kia-verkent-inzichten-in-voertuigdata-met-ministerie-van-ienw/

Kulmala, R. (2010). Ex-ante assessment of the safety effects of intelligent transport systems. Accident Analysis & Prevention, 42(4), 1359–1369. https://doi.org/10.1016/j.aap.2010.03.001

Landeta, J. (2006). Current validity of the Delphi method in social sciences. Technological Forecasting and Social Change, 73(5), 467–482. https://doi.org/10.1016/j.techfore.2005.09.002

Laureshyn, A., Johnsson, C., De Ceunynck, T., Svensson, Å., de Goede, M., Saunier, N., Włodarek, P., van der Horst, R., & Daniels, S. (2016). Review of current study methods for VRU safety. Appendix 6–Scoping review: Surrogate measures of safety in site-based road traffic observations: Deliverable 2.1–part 4.

Linstone, H., & Turoff, M. (1975). The Delphi Method: Techniques and Applications. In Technometrics (Vol. 18). https://doi.org/10.2307/3150755

Mahmud, S. M. S., Ferreira, L., Hoque, Md. S., & Tavassoli, A. (2017). Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS Research, 41(4), 153–163. https://doi.org/10.1016/j.iatssr.2017.02.001

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research, 1, 1–22. https://doi.org/10.1016/j.amar.2013.09.001

Markmann, C., Spickermann, A., von der Gracht, H. A., & Brem, A. (2021). Improving the question formulation in Delphi-like surveys: Analysis of the effects of abstract language and amount of information on response behavior. FUTURES & FORESIGHT SCIENCE, 3(1), e56. https://doi.org/10.1002/ffo2.56

Matysiak, A., & Razin, P. (2018). Analysis of the advancements in real-life performance of highly automated vehicles' with regard to the road traffic safety. MATEC Web of Conferences, 231, 05003.

Mauksch, S., von der Gracht, H. A., & Gordon, T. J. (2020). Who is an expert for foresight? A review of identification methods. Technological Forecasting and Social Change, 154, 119982. https://doi.org/10.1016/j.techfore.2020.119982

McKinsey. (2018). From buzz to bucks—Automotive players on the highway to car data monetization. McKinsey.

Melander, L. (2018). Scenario development in transport studies: Methodological considerations and reflections on delphi studies. Futures, 96, 68–78. https://doi.org/10.1016/j.futures.2017.11.007

Ministerie van Infrastructuur en Watermanagemant. (2018). Veilig van deur tot deur. Het Strategisch Plan Verkeersveiligheid 2030: Een gezamenlijke visie op aanpak verkeersveiligheidsbeleid. https://www.rijksoverheid.nl/documenten/rapporten/2018/12/05/bijlage-1-het-strategisch-plan-verkeersveiligheid-2030-veilig-van-deur-tot-deur

Ministerie van Verkeer en Waterstaat. (2009). Strategisch Plan Verkeersveiligheid 2008-2020: Van, voor en door iedereen. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiI7cG8gvnvAhVwgf0HHd0QDTAQFjAAegQIAhAD&url=https%3A%2F%2Fwww.rijksoverheid.nl%2Fbinaries%2Frijksoverheid%2Fdocumenten%2Fbeleidsnota-s%2F2009%2F06%2F15%2Fstrategisch-plan-verkeersveiligheid-2008-2020%2Fstrategischplan-n.pdf&usg=AOvVaw33jBUV-oxGY8wjkQztHbAa

Minsterie van Infrastructuur en Waterstaat. (2021, May 10). Inzet van techniek in auto's: Onderhoud wegen efficiënter. Nieuws IenW. https://www.nieuwsienw.nl/1955269.aspx?t=Inzet-van-techniek-in-auto%E2%80%99s-onderhoud-wegen-efficienter--

Modderkolk, H. (2020, November 16). De overheid lijdt aan datahonger en dat kost de burger privacy. de Volkskrant. https://www.volkskrant.nl/gs-b81be6f4

MuConsult. (2021). Bandbreedte ontwikkeling penetratiegraad ADAS. Ministerie van Infrastructuur en Waterstaat.

Nissenbaum, H. (2004). Privacy As Contextual Integrity. Washington Law Review, 79.

Norani, N., Deros, B., Abd Wahab, D., & Ab Rahman, M. (2012). Validation of Lean Manufacturing Implementation Framework Using Delphi Technique. Jurnal Teknologi (Sciences and Engineering), 59, 1–6. https://doi.org/10.11113/jt.v59.1596

Norman, G. (2010). Likert scales, levels of measurement and the 'laws' of statistics. Advances in Health Sciences Education: Theory and Practice, 15(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Nowack, M., Endrikat, J., & Guenther, E. (2011). Review of Delphi-based scenario studies: Quality and design considerations. Technological Forecasting and Social Change, 78(9), 1603–1615. https://doi.org/10.1016/j.techfore.2011.03.006

Noy, I. Y., Shinar, D., & Horrey, W. J. (2018). Automated driving: Safety blind spots. Safety Science, 102, 68–78. https://doi.org/10.1016/j.ssci.2017.07.018

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. Information & Management, 42(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002

Onderzoeksraad voor Veiligheid. (2019). Wie stuurt? Verkeersveiligheid en automatisering in het wegverkeer. http://www.onderzoeksraad.nl/nl/page/4729/wie-stuurt-verkeersveiligheid-en-automatisering-in-het-wegverkeer

Orlovska, J., Wickman, C., & Soderberg, R. (2020). THE USE OF VEHICLE DATA IN ADAS DEVELOPMENT, VERIFICATION AND FOLLOW-UP ON THE SYSTEM. Proceedings of the Design Society: DESIGN Conference, 1, 2551–2560. https://doi.org/10.1017/dsd.2020.322

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A., Jarawan, E., & Mathers, C. (2004). World Report on Road Traffic Injury Prevention. World Health Organisation. https://www.who.int/publications-detail-redirect/world-report-on-road-traffic-injury-prevention

Piecyk, M. I., & McKinnon, A. C. (2010). Forecasting the carbon footprint of road freight transport in 2020. International Journal of Production Economics, 128(1), 31–42. https://doi.org/10.1016/j.ijpe.2009.08.027

Raskin, M. S. (1994). The Delphi Study in Field Instruction Revisited: Expert Consensus on Issues and Research Priorities. Journal of Social Work Education, 30(1), 75–89. https://doi.org/10.1080/10437797.1994.10672215

Ray, P. K., & Sahu, S. (1990). Productivity Management in India: A Delphi Study. International Journal of Operations & Production Management, 10(5), 25–51. https://doi.org/10.1108/01443579010005245

Regulation (EU) 2015/758, (2015). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32015R0758

Regulation (EU) 2015/962, (2014). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32015R0962

Regulation (EU) 2018/858, Pub. L. No. 32018R0858, 151 OJ L (2018). http://data.europa.eu/eli/reg/2018/858/oj/eng

Regulation (EU) No 886/2013, (2013). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32013R0886

Reiman, J. H. (1995). Driving to the Panopticon: A Philosophical Exploration of the Risks to Privacy Posed by the Highway Technology of the Future. In Privacy. Routledge.

Roßmann, B., Canzaniello, A., von der Gracht, H., & Hartmann, E. (2018). The future and social impact of Big Data Analytics in Supply Chain Management: Results from a Delphi study. Technological Forecasting and Social Change, 130, 135–149. https://doi.org/10.1016/j.techfore.2017.10.005

Ryan, C. (2020). Emerging autonomous vehicle risks: The role of telematics and machine learning based risk assessment.

Schagen, I. N. L. G. van, & Aarts, L. T. (2018). DV3 – Huidige situatie, maatschappelijke trends en wensbeelden; Kader voor de visie Duurzaam Veilig Wegverkeer 2018-2030 (p. 74). SWOV.

Scheibe, M., Skutsch, M., & Schofer, J. (2002). IV. C. Experiments in Delphi methodology. The Delphi Method: Techniques and Applications, 257–281.

Schmalz, U., Spinler, S., & Ringbeck, J. (2021). Lessons Learned from a Two-Round Delphi-based Scenario Study. MethodsX, 8, 101179. https://doi.org/10.1016/j.mex.2020.101179

Schuckmann, S. W., Gnatzy, T., Darkow, I.-L., & von der Gracht, H. A. (2012). Analysis of factors influencing the development of transport infrastructure until the year 2030—A Delphi based scenario study. Technological Forecasting and Social Change, 79(8), 1373–1387. https://doi.org/10.1016/j.techfore.2012.05.008

Schwall, M., Daniel, T., Victor, T., Favarò, F., & Hohnhold, H. (2020). Waymo Public Road Safety Performance Data. ArXiv, abs/2011.00038.

Sener, I. N., Zmud, J., & Williams, T. (2019). Measures of baseline intent to use automated vehicles: A case study of Texas cities. Transportation Research Part F: Traffic Psychology and Behaviour, 62, 66–77. https://doi.org/10.1016/j.trf.2018.12.014

Shields, T. J., Silcock, G. W. H., Donegan, H. A., & Bell, Y. A. (1987). Methodological problems associated with the Use of the Delphi technique. Fire Technology, 23(3), 175–185. https://doi.org/10.1007/BF01036934

Singh, S. (2015). Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. Traffic Safety Facts - Crash Stats, Article DOT HS 812 115. https://trid.trb.org/view.aspx?id=1346216&source=post_page

Skulmoski, G., Hartman, F., & Krahn, J. (2007). The Delphi Method for Graduate Research. JITE, 6, 1–21. https://doi.org/10.28945/199

Sohrabi, S., Khodadadi, A., Mousavi, S. M., Dadashova, B., & Lord, D. (2021). Quantifying the automated vehicle safety performance: A scoping review of the literature, evaluation of methods, and directions for future research. Accident Analysis & Prevention, 152, 106003. https://doi.org/10.1016/j.aap.2021.106003

Souman, J., Adjenughwure, K., van Dam, E., van Weperen, M., & Tejada, A. (2021). Quantification of Safe Driving (TNO 2021 R12632 0.1). TNO. https://repository.tno.nl//islandora/object/uuid:df4404be-252b-4187-832e-28dfd1609c4b

Spickermann, A., Zimmermann, M., & von der Gracht, H. A. (2014). Surface- and deep-level diversity in panel selection—Exploring diversity effects on response behaviour in foresight. Technological Forecasting and Social Change, 85, 105–120. https://doi.org/10.1016/j.techfore.2013.04.009

Stipdonk, H. L. (2013). Road safety in bits and pieces; for a better understanding of the development of the number of road fatalities. https://repository.tudelft.nl/islandora/object/uuid%3A435ca957-1b52-446c-a325-f12c433a99d9

SWOV. (2005). State of the art Report on Road Safety Performance Indicators; Deliverable D3.1 of the EU FP6 project SafetyNet (p. 177). European Commission.

SWOV. (2012). Naturalistic Driving: Observing everyday driving behaviour [SWOV-Factsheet].

SWOV. (2016). Gegevensbronnen; Uitgebreid overzicht. SWOV. http://www.swov.nl/sites/default/files/bestanden/wegwijzer/gegevensbronnen.pdf

SWOV. (2020a). Kosten van verkeersongevallen. SWOV-Factsheet, Maart 2020.

SWOV. (2020b). Verkeersdoden in Nederland [SWOV-Factsheet, april 2020.]. SWOV. https://www.swov.nl/feiten-cijfers/fact/verkeersdoden-hoe-heeft-het-aantal-verkeersdoden-nederland-zich-sinds-1950

SWOV. (2020c). De Staat van de Verkeersveiligheid 2020: Doelstellingen 2020 worden niet gehaald. SWOV. https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2020/12/14/bijlage-1-de-staat-van-de-verkeersveiligheid-2020-swov/bijlage-1-de-staat-van-de-verkeersveiligheid-2020-swov.pdf

Talal, M., Ramli, K. N., Zaidan, A. A., Zaidan, B. B., & Jumaa, F. (2020). Review on car-following sensor based and data-generation mapping for safety and traffic management and road map toward ITS. Vehicular Communications, 25, 100280. https://doi.org/10.1016/j.vehcom.2020.100280

Tarko, A. P. (2012). Use of crash surrogates and exceedance statistics to estimate road safety. Accident Analysis & Prevention, 45, 230–240. https://doi.org/10.1016/j.aap.2011.07.008

Teeffelen, K. van. (2022, April 20). Privacywaakhond ziet te grote datahonger bij de overheid. Trouw. https://www.trouw.nl/gs-b050316e

Thulin, H., & Nilsson, G. (1994). Road traffic, exposure, injury risks and injury consequences for different travel modes and age groups (VTI rapport 390A). Statens Väg-och Transportforskningsinstitut.

Toma, C., & Picioreanu, I. (2016). The Delphi Technique: Methodological Considerations and the Need for Reporting Guidelines in Medical Journals. International Journal of Public Health Research, 4, 47–59.

TRL. (2017). Access to In-vehicle Data and Resources; Final Report (No. CPR2419). https://ec.europa.eu/transport/themes/its/studies/its_en. transport.its.menu

UN Regulation No 157 [2021/389], (2021). https://eur-lex.europa.eu/eli/reg/2021/389/oj

UNECE. (2020, June 24). UN Regulation on Automated Lane Keeping Systems is milestone for safe introduction of automated vehicles in traffic | UNECE. https://unece.org/transport/press/un-regulation-automated-lane-keeping-systems-milestone-safe-introduction-automated

van Huffelen, A. (2022, May 18). Antwoorden op Kamervragen over datahonger van overheid—Kamerstuk—Rijksoverheid.nl. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/kamerstukken/2022/05/18/antwoorden-op-kamervragen-over-de-datahonger-van-de-overheid

Van Nieuwenhuizen, C. (2019). Antwoord op vragen van de leden Schonis en Verhoeven over het artikel 'Wie temt het datamonster in de auto-industrie?' https://www.tweedemonitor.nl/kamervraag/2019D21928

Victor, T. W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel. Human Factors, 60(8), 1095–1116. https://doi.org/10.1177/0018720818788164

Virdi, N., Grzybowska, H., Waller, S. T., & Dixit, V. (2019). A safety assessment of mixed fleets with Connected and Autonomous Vehicles using the Surrogate Safety Assessment Module. Accident Analysis & Prevention, 131, 95–111. https://doi.org/10.1016/j.aap.2019.06.001

Vlakveld, W. P. (2019). Veiligheidseffecten van rijtaakondersteunende systemen. SWOV.

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. Technological Forecasting and Social Change, 79(8), 1525–1536. https://doi.org/10.1016/j.techfore.2012.04.013

von der Gracht, H. A., & Darkow, I.-L. (2010). Scenarios for the logistics services industry: A Delphi-based analysis for 2025. International Journal of Production Economics, 127(1), 46–59. https://doi.org/10.1016/j.ijpe.2010.04.013

Wang, C., Xie, Y., Huang, H., & Liu, P. (2021). A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. Accident Analysis & Prevention, 157, 106157. https://doi.org/10.1016/j.aap.2021.106157

Wang, L., Zhong, H., Ma, W., Abdel-Aty, M., & Park, J. (2020). How many crashes can connected vehicle and automated vehicle technologies prevent: A meta-analysis. Accident Analysis & Prevention, 136, 105299. https://doi.org/10.1016/j.aap.2019.105299

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. Advances in Health Sciences Education, 7(1), 63–80. Scopus. https://doi.org/10.1023/A:1014585522084

Warth, J., von der Gracht, H. A., & Darkow, I.-L. (2013). A dissent-based approach for multi-stakeholder scenario development—The future of electric drive vehicles. Technological Forecasting and Social Change, 80(4), 566–583. https://doi.org/10.1016/j.techfore.2012.04.005

Winkler, J., & Moser, R. (2016). Biases in future-oriented Delphi studies: A cognitive perspective. Technological Forecasting and Social Change, 105, 63–76. https://doi.org/10.1016/j.techfore.2016.01.021

Wishart, J., Como, S., Elli, M., Russo, B., Weast, J., Altekar, N., & James, E. (2020, April 14). Driving Safety Performance Assessment Metrics for ADS-Equipped Vehicles. https://doi.org/10.4271/2020-01-1206

Xie, K., Yang, D., Ozbay, K., & Yang, H. (2019). Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure. Accident Analysis & Prevention, 125, 311–319. https://doi.org/10.1016/j.aap.2018.07.002

Young, W., Sobhani, A., Lenné, M. G., & Sarvi, M. (2014). Simulation of safety: A review of the state of the art in road safety simulation modelling. Accident Analysis & Prevention, 66, 89–103. https://doi.org/10.1016/j.aap.2014.01.008

Yue, L., Abdel-Aty, M., Wu, Y., & Wang, L. (2018). Assessment of the safety benefits of vehicles' advanced driver assistance, connectivity and low level automation systems. Accident Analysis & Prevention, 117, 55–64. https://doi.org/10.1016/j.aap.2018.04.002

Zimmer, M. (2005). Surveillance, Privacy and the Ethics of Vehicle Safety Communication Technologies. Ethics and Information Technology, 7(4), 201–210. https://doi.org/10.1007/s10676-006-0016-0

Appendix A Scientific paper

Starting on the next page is a scientific paper summarising the research.

Opportunities and barriers for measuring traffic safety indicators based on vehicle sensor data: a Delphi study

S. Jansen

Keywords
Traffic safety
Vehicle sensor data
Delphi study
Indicators
Consensus

Abstract
Current indicators for measuring traffic safety such as fatalities and severe injuries are reactive and lagging indicators. With the rise of automation in vehicles, the opportunity may arise to measure proactive and leading indicators. This research uses the knowledge of experts on traffic safety and/or vehicle sensor data to evaluate the suitability of five types of indicators to measure traffic safety based on vehicle sensor data. It also evaluates ten potential barriers for using vehicle sensor data in practice. While limited differences are found between the types of indicators, none of the potential barriers is rated as unsurmountable by the experts. In doing so, this research proofs that it would be possible to measure traffic safety with vehicle sensor data and provides a starting point for further research

1. Introduction

In the Netherlands, traffic safety policies have been in place for decades and currently have the ultimate goal to reach zero traffic fatalities by 2050 (Ministerie van Infrastructuur en Water Managemant, 2018). Traffic safety policies aim to influence the operational conditions of road traffic, resulting in less accidents, fatalities, and injuries which in turn reduces the total social costs associated with accidents (SWOV, 2005). At least, if the traffic safety measure is designed and implemented correctly. In order to assess whether or not an implemented safety measure is effective – and to see if the costs are proportionate to the benefits – it is important to monitor its effects on traffic safety (L. T. Aarts, 2018; SWOV, 2005). Measuring traffic safety is not only important for monitoring the effect of safety measures, but also for estimating current (specific) aspects of traffic safety and for comparison with other countries (ETSC, 2001).

No model exist that can fully explain traffic safety, with all the relevant factors and their corresponding importance (Stipdonk, 2013). So, indicators are used to measure traffic safety instead. And while it is important to note that no single indicator exists that can fully describe traffic safety (Fraade-Blanar et al., 2018), the two most important and widely used indicators are the number of fatalities and number of severe injuries (L. Aarts et al., 2021). These indicators have a clear and uniform definition, resulting in relatively high quality of the data and are easy to understand for both the general public and policy makers (Blumenthal et al., 2020). Given their long use it also is possible to compare trends over time (Stipdonk, 2013). In other words, fatalities and severe injuries can “tell the final story on if it is safe or not.” (Blumenthal et al., 2020, p12).

These indicators are however not without their problems. There often are problems with underreporting of both fatalities and severe injuries which are not uniformly distributed throughout time and (most likely) over types of crashes (SWOV, 2016, 2020a). These crashes are also subject to random fluctuations (Chang et al., 2017; ETSC, 2001). But most importantly, crashes are rare occurrences (see figure 1) which makes (statistical) analysis more difficult (Chin & Quek, 1997; Tarko, 2012).

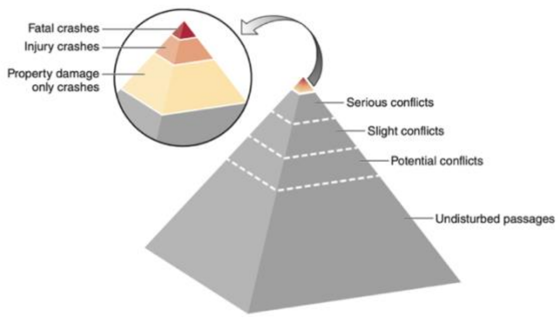


Figure 1 Hydén’s Safety Pyramid (Chang et al., 2017)

Using severe injuries fatalities as indicators to monitor and evaluate traffic safety measures is a reactive approach in which a sufficiently large number of accidents need to occur, before a problem can be found and addressed. Many researchers (Arun et al., 2021; Chin & Quek, 1997; Mahmud et al., 2017; Tarko, 2012; Wang et al., 2020) mention that this raises ethical questions as people first need to crash before action can be taken to prevent those crashes and fatalities.

These issues can be alleviated by using leading indicators. Leading indicators are proactive indicators that measure events leading up to accidents, meaning that accidents do not need to happen before traffic safety can be measured (L. T. Aarts, 2018). So, they measure the types of traffic conflicts lower in the Safety Pyramid of Hydén which can serve as surrogates or proxies for lagging indicators (Fraade-Blanar et al., 2018). An often-used definition for a traffic conflict is the definition by Amundsen & Hyden (1977): “traffic conflicts are an observable situation in which two or more road users approach each other in space and time to such an extent that there is a risk of collision if their movements remain unchanged" (in Arun et al., 2021, p4).

Measuring these traffic conflicts requires data that can be gathered by observational data with fixed point sensors as discussed by for example Talal et al. (2020) and Young et al. (2014) or with on-board data collection as is often done in a Naturalistic Driving Study (Grimberg et al., 2020; Vlakveld, 2019). Both methods have their issues where the former is limited by its fixed location and resulting less variety (Talal et al., 2020; Young et al., 2014) while the latter is expensive because of the need to retrofit vehicles with a range of sensors and a way to store and transmit data, resulting in relatively short data collection methods with a limited number of vehicles (Grimberg et al., 2020; Talal et al., 2020; Vlakveld, 2019).

The rise of the share of vehicles equipped with ADAS may provide an opportunity to overcome this problem. Vehicles equipped with ADAS already have a range of sensors aboard to facilitate the ADAS and are often connected with built-in SIM cards (Ecorys, 2020). The question can be raised if these can be used to measure leading indicators for traffic safety. After all, if in an ideal world all vehicles would be able to collect and transmit data on driving behaviour, it would be possible to conduct what is essentially a Naturalistic Driving Study on an unprecedented scale.

This research aims to explore if and how the emergence of data produced by vehicles equipped with ADAS can be leveraged to measure traffic safety in practice. The research question central to this study is the following: (how) can data from sensors of vehicles equipped with ADAS be used to measure traffic safety at network level?

The research will focus on two aspects of this question: what type of indicators could be used to measure traffic safety based on vehicle sensor data, and is it feasible to use vehicle sensor data in practice?

To develop types of indictors based on vehicle sensor data, a literature review is carried out into existing types of indicators for measuring traffic safety. It will look at existing indicators used by policymakers to measure traffic safety at network level, as well as into indicators for evaluating traffic safety at vehicle level as used in scientific literature to assess the performance of vehicles equipped with ADAS. These indicators are then adapted to be based on vehicle sensor data. Experts on traffic safety and/or vehicle sensor data evaluated these types of indicators in the form of a Delphi study. Similarly, potential barriers for using vehicle sensor data in practice are identified in scientific literature and policy documents and evaluated by experts.

This research is scientifically relevant because it helps contribute to the body of knowledge of traffic safety evaluation. This research helps to explore if data collected by vehicles could help overcome some of the currently existing issues and ultimately improve the way in which traffic safety is measured currently.

This also has a clear societal relevance because it can help to better evaluate traffic safety policies, decrease the number of accidents, reach the goal of Vision Zero and ultimately save lives.

2. Literature review

The first part of the literature review focusses on currently existing and used types of indicators for measuring traffic safety and an overview is given in table 1.

2.1 Types of indicators for measuring traffic safety

Dutch SPIs

In the Netherlands, the concept of risk-based policy (*risciogestuurd beleid*) is develop where the idea is to map and reduce risks in traffic (Ministerie van Infrastructuur en Watermanagemant, 2018). These risks are measured in various Safety Performance Indicators (SPIs) focussing on various areas related to traffic safety such as roads, drivers, or speeds(L. T. Aarts, 2018; Ministerie van Infrastructuur en Watermanagemant, 2018). Stipdonk (2013) emphasises the importance of taking exposure (in any form) into account, which is done in the SPIs by defining all the SPIs as the "share of…", and not just the "number of…" (L. T. Aarts, 2018).

Proximity based SMoS

Surrogate Measures of Safety (SMoS) are indicators aimed at measuring traffic conflicts (Johnsson et al., 2021; Laureshyn et al., 2016) There is a large variety of SMoS, as is shown by Mahmud et al. (2017) who have identified 38 SMoS in their literature review. However, two SMoS are used significantly more than any other, namely time-to-collision (TTC), and post encroachment time (PET) (Johnsson et al., 2021; Laureshyn et al., 2016).

TTC is the time remaining until an accident occurs if two vehicles maintain their current course and speed (Mahmud et al., 2017). TTC is mainly used for rear-end type of crashes or for hitting pedestrians or objects like a parked vehicle (Mahmud et al., 2017). For measuring of modelling traffic safety on inttersections, PostEncroachment Time (PET) is more suitable (Mahmud et al., 2017). PET is the time between one vehicle leaving a certain point or area and the arrival of a second vehicle at that point or area (Arun et al., 2021). Both are called proximity based SMoS as these are based on proximity in time (Mahmud et al., 2017).

Kinematic based SMoS

Kinematic based SMoS are based on deceleration or acceleration (Arun et al., 2021; Mahmud et al., 2017) and are often prefered in naturalistic driving studies (Arun et al., 2021). The reasoning is that in urban areas the most common evasive action to avoid collision is deceleration (Johnsson et al., 2018). Therefore, (strong) deceleration could indicate a potentially dangerous situation. This could be defined in a relatively simple way in which each event in which the strength of the deceleration exceeds a threshold is counted (Arun et al., 2021). More elaborate indicators have also been developed as described by Mahmud et al. (2017). They describe the Deceleration Rate to Avoid the Crash (DRAC) which also considers vehicle in front.

The main advantage of kinematic based SMoS is

that they are easy to understand, objective and physics-based and that they can be used in several different situations (Blumenthal et al., 2020; Mahmud et al., 2017). It is therefore used in naturalistic driving studies (Arun et al., 2021). Next to deceleration, acceleration could be used in a similar way, as well as evasive action in the form of swerving (Johnsson et al., 2018). Johnsson et al. (2018) does however report that no validation studies have been carried out into deceleration-based indicators.

Engagement of automation

In the past, disengagement of the automation in a vehicle is used as proxy for unsafe situations but has since fallen out of use (Blumenthal et al., 2020; Fraade-Blanar et al., 2018). The main reason for this is how sensitive disengagement is to the circumstances like the environment and the driver (Blumenthal et al., 2020).

In recent pilots, engagement of specific types of ADAS in the vehicle such as Forward Collision Warning (FCW) or Automatic Emergency Braking (AEB) is used as a proxy for an unsafe situations. After all, these are safety systems that warn the driver or intervene when the system detects that an accident is imminent (Onderzoeksraad voor Veiligheid, 2019). No discussion of using engagement of these systems as a measure for safety was found in academic literature.

2.2 Potential barriers for using vehicle sensor data in practice

Scenario for using vehicle sensor data

The vehicle sensor data collected by the vehicle is owned by the OEM and while it is possible to collect data with aftermarket devices, this is not very suited for large scale data-collection (Ecorys, 2020). Therefore, any effort for large scale collection would require cooperation with OEMs. A potentially large market exists around mobility data which has the interest of OEMs which could make them willing to cooperate and supply vehicle sensor data to the government for measuring traffic safety (McKinsey, 2018). Forcing OEMs to share specific data concerning safety is also not without precedent, as current Regulation (EU) No 886/2013 (2013) already obliges OEMs to collect and share data about specific cases, the Safety Related Traffic Information (SRTI).

The most likely model to share large amounts of vehicle sensor data between OEMs, the government and potentially third parties is the extended vehicle (ExVe)/neutral server model as proposed by European Automobile Manufacturers' Association (ACEA) which see it as the only solution that can guarantee sufficient security and safety of the data (Ecorys, 2020).

Table 1 Overview of indicators used to measure traffic safety

Category	Specific examples	Strengths	Weaknesses	Current usage
Lagging indicators	Fatalities, severe injurie	1 Easy to understand [a] 2 Uniform definition [a] 3 Historical data available [ab] 4 High validity [a]	1 Reactive approach [cd] 2 Rarity of occurrence [d] 3 Does not give insight into process [ce] 4 Subject to random fluctuations [ce] 5 Data incompleteness [b]	network level [f]
Dutch SPIs	Safe participants, safe speeds	1 Help to give insight [g]	1 Still being implemented [g] 2 Little data available [g]	Still in development, aims for network level at national, regional, and local scale [g]
Proximity based SMoS	TTC, PET	1 Directly observable in traffic [h] 2 Objective and physics-based [a]	1 Data intensive 2 Specific for situation [i] 3 Validity of threshold [hij]	(simulation) experiments [hi]
Kinematic SMoS	Deceleration, acceleration, swerving	1 Easy to understand [a] 2 Objective and physics-based [a] 3 Suitable for several situations [i]	1 Validity of threshold [hij]	Naturalistic driving studies [j]
Engagement of ADAS	BSW, ACC, LKA, FCW	1 Easy to measure [ak]	?	Public-private pilots [lm]

a Blumenthal et al. (2020)
b Stipdonk (2013)
c Chang et al. (2017)
d Tarko (2012)
e ETSC (2001)
f SWOV (2020c)
g Aarts (2018)
h Johnsson et al. (2021)
i Mahmud et al. (2017)
j Arun et al. (2021)
k Presumed to be similar to disengagement of ADAS
l Interview with Vrijens, ministry of I&W, Appendix B
m Kia Nederland (2021)

In this model, vehicle sensor data is transmitted encrypted to dedicated servers of the OEM which can then make (processed) data available to third parties (TRL, 2017). This can be done directly or indirectly through the neutral server to allow third parties an option to choose (TRL, 2017). This model was successfully applied in a Proof on Concept (PoC) for generating and sharing SRTI messages (Henkens et al., 2020). It is important to note that this is all done on a pilot level, which are limited in terms of scope. This pilot, and others, have provided opportunities to explore the potential of using vehicle sensor data. The following potential barriers to doing just that on a large scale in practice are identified and discussed below.

Legal feasibility

Most legal issues in this case have to do with privacy. Under the GDPR, data that can be traced back to an individual is personal data and as a result needs to adhere to stricter rules (Ecorys, 2020). There is an ongoing debate whether or not vehicle data is personal data in which various stakeholders and even the supervisory authorities on privacy of different EU member states disagree (Ecorys, 2020). The owner gives permission to collect data, but the vehicle manufacturer decides to share which information with which party (Henkens et al., 2020). A second set of legal issues may arise around fair market competition, given the dominant position of the vehicle manufacturers (TRL, 2017). This should be less of a concern as existing laws on this topic should be strong enough to prevent unfair market competition although the practical application of these laws is highly complex (Ecorys, 2020; TRL, 2017).

Cybersecurity

Cybersecurity is one of the most discussed emerging risks in the transition towards more autonomous driving (Ryan, 2020). An increase in the number of external connections in a vehicle comes with an increase in the potential points of attack (Fraade-Blanar et al., 2018; Onderzoeksraad voor Veiligheid, 2019; Ryan, 2020). The risks are also substantial: in recent years researchers and ethical hackers have managed to gain remote access to vehicles which allowed them to take over control of systems like the brakes, engine, and steering wheel (Onderzoeksraad voor Veiligheid, 2019; Ryan, 2020). Besides the direct threat to human life present in such a hacked vehicle, other risks are also present such as data breaches and ransomware attacks (Fraade-Blanar et al., 2018; Ryan, 2020).

Willingness of stakeholders

Any system that uses vehicle data for any purpose will involve a variety of stakeholders. Who exactly will be involved depends on the scope that is applied but Ecorys (2020), McKinsey (2018), and TRL (2017) all discuss OEMs, suppliers, and service providers.

Economic feasibility

Closely related to willingness of stakeholder is the economic feasibility. Is collecting and sharing vehicle sensor data on a large scale feasible against reasonable cost to OEMs or the government?

Technical feasibility

Collecting, sharing, and processing vehicle data is a technically complex operation. However, as the PoC Data for Road Safety and the Kia project show, it is possible to extract data from sensors in the vehicle and to process this data outside of the vehicle with the ExVe/neutral server model (Henkens et al., 2020; Kia, 2021).

From pilot to reality

Besides the issues on specific aspects as discussed above, a more general factor is also relevant to consider: the paradox of the pilot. Groenendijk (2021) calls the belief that a successful pilot can simply be scaled up and will lead to the same successful results naïve. For example, the extra budget and room to experiment in a specific pilot may not necessarily be available in the regular day-to-day business of an OEM or government (Groenendijk, 2021)

3. Methodology

To assess if the types of indicators identified in the literature review would be suitable to be based on vehicle sensor data to measure traffic safety, and to assess the size of the potential barriers to use vehicle sensor data in practice, a Delphi study is conducted. There are multiple variations of the Delphi study but it always has the following four characteristics: anonymity, iteration, controlled feedback, and statistical group response (Fritschy & Spinler, 2019; von der Gracht, 2012). In this research, experts are asked questions, separate and anonymous to prevent influencing each other. In subsequent rounds, the same are asked again with additional information, the controlled feedback. This feedback consists of statistics on the group response of the previous round and often used arguments. The experts are then given the chance to revaluate their answer, with the goal of reaching consensus or establishing clear dissensus. (Beiderbeck et al., 2021a; Belton et al., 2019; Hsu & Sandford, 2007)

3.1 Expert selection

A Delphi survey is fundamentally different from a regular survey because it does not have the goal to generalise results of a representative sample to a larger population, but instead to reach consensus among experts (Okoli & Pawlowski, 2004).Therefore, participants are not selected randomly but rather purposely and based on their expertise (Keeney et al., 2006). And while the Delphi method has proven to produce valid results in the past (Förster & von der Gracht, 2014; Landeta, 2006), the selection of appropriate experts is a highly important part in the process, as the quality of the experts directly relates to the quality of the results (Hsu & Sandford, 2007; Keeney et al., 2006). No single best way to define and measure expertise has been developed (Devaney & Henchion, 2018; Hasson et al., 2000; Mauksch et al., 2020). This research follows the recommendations of Mauksch et al. (2020) to use a combination of expert identification methods which helps to mitigate the drawbacks and potential biases of each individual method. Personal Involvement using the expert continuum model of Donohoe & Needham (2009) is used to identify a wide range of experts involved in the topic. Using this model (figure 2), organisations are selected that have either hands-on experience (subjective closeness), a formal role (policy/legal) (mandated closeness), or an objective standpoint (objective closeness).

Figure 2 The expert continuum model (adapted from Donohoe & Needham, 2009)

This is followed by a selection of experts within those organisations using external cues to assess their expertise (Mauksch et al., 2020) where experts need to have at least three of the following five criteria (based on Devaney & Henchion, 2018; Roßmann et al., 2018; Schuckmann et al., 2012; von der Gracht & Darkow, 2010; Warth et al., 2013):

1. The expert has at least 5 years of experience with the topic
2. The expert has published a paper or spoken at conference on the topic
3. The expert has a relevant work position in the field
4. The expert has a relevant academic background
5. The expert is based in the Netherlands or in a comparable EU country

27 experts were invited, 16 completed round 1 and 11 round 2. This is in line with typical ranges suggested between 5 to 20 (Belton et al., 2019), 7 to 15 (Donohoe & Needham, 2009), 10 to 18 (Okoli & Pawlowski, 2004) or 15 to 20 (Hsu & Sandford, 2007). In this way, a heterogeneous panels is selected which have proven to create a wider range of perspectives than an homogeneous panel and helps to prevent bias and framing effects (Förster & von der Gracht, 2014; Winkler & Moser, 2016).

3.2 Operationalisation

The Delphi survey consist of questions that require a numerical response, followed by an opportunity for the expert to explain or comment on the question (Belton et al., 2019) where in the wording of questions advice by Markmann et al. (2021) is followed to prevent influencing the quality of the survey data. The survey consists of two parts where in part one, experts are asked to rate the four types of indicators on four criteria (table 2) on a scale from 1 (strongly disagree) to 7 (strongly agree). They are then given the opportunity to explain their assessment.

Table 2 Criteria used to evaluate types of indicators

Criteria in Aarts (2018)	Criteria operationalised in Delphi survey
Validity	a. This type of indicator reflects traffic safety well
Reliability	b. This type of indicator could be measured in a reliable way
Sensitivity	c. This type of indicator is sensitive to external changes, i.e., it will respond to future traffic safety interventions
Understandability	d. This type of indicator is understandable for different end-users such as researchers and policy makers

Part two of the survey is aimed at assessing the feasibility of using vehicle sensor data to measure traffic safety. The experts are asked to rate the severity of potential on a 7-point-scale (no barrier at all to unsurmountable barrier) and again provide an explanation.

A 7-point Likert-type scale is used in this Delphi survey as this is seen as the most reliable (Beiderbeck et al., 2021a; Toma & Picioreanu, 2016). Following Toma & Picioreanu (2016), an 8th option, “I do not know ”, is included to ensure that the middle option is truly neutral. All options are labelled to leave little room of interpretation by the experts and reach the highest psychometric quality (Beiderbeck et al., 2021a).

Delphi surveys have an iterative character where question are asked over multiple rounds (Belton et al., 2019; Fritschy & Spinler, 2019). This Delphi survey consists of two rounds, as research shows that expert fatigue results in lower response rates over each subsequent round (Frewer et al., 2011; Fritschy & Spinler, 2019; Keeney et al., 2006). Additionally, most revisions of the expert’s own opinions happen after the first round and not afterwards (Fritschy & Spinler, 2019).

The controlled feedback provided to the experts in the second round consists of a histogram showing the distribution of scores given by the experts in the first round per each question. This is augmented with qualitative feedback that consists of arguments brought up by the experts, both in favour and against. Not all arguments made in the first round are included in the survey to keep the workload for the experts reasonable. A selection is made of the most used arguments of which some are edited to remove any references that could reveal the identity of the experts. Furthermore, a suggestion by Winkler & Moser (2016) (see 3.3 below) is followed for most questions to sort the arguments and put those contrary to the central tendency on top, to make sure that experts are confronted with opposite opinions and are thus forced to rethink their own opinion.

3.3 Analysis

Syntax analysis and non-response bias

A syntax analysis of the explanations given by the experts in round will be conducted to show their level of engagement (Beiderbeck et al., 2021a). The syntax analysis as is conducted in this study was originally developed by Förster & von der Gracht (2014) and also applied in Roßmann et al. (2018) and in Beiderbeck et al. (2021b). In a syntax analysis, the experts' explanations are classified as whole sentences, phrases, catchwords, or no explanation. A high percentage of whole sentences indicates a high level of commitment and thus serves as a quality measure (Beiderbeck et al., 2021a).

Additionally, there will be tested for the existence of non-response bias by comparing the answers of early respondents to those of late respondents (Díaz de Rada, 2005). This is done by dividing the group of respondents in two by order of responses (Warth et al., 2013) and by those who responded initially and those who only responded after reminders (Díaz de Rada, 2005) and conducting a Mann-Whitney U test to check for differences (Piecyk & McKinnon, 2010; Warth et al., 2013).

Descriptive statistics and level of agreement

Descriptive statistics in a Delphi study often includes the central tendency and the dispersion (von der Gracht, 2012). Given that the scores from a Likert scale are of ordinal level, the median the preferred choice for a metric of central tendency (von der Gracht, 2012). Additionally, the median is less sensitive to outliers than the mean (Gordon & Pease, 2006; von der Gracht, 2012). Consequently, the Interquartile Range (IQR) will be used as it is a measure of dispersion for the median and to assess the level of agreement.

The IQR is a rigorous and objective metric of level of agreement (von der Gracht, 2012) that is the distance between the 25th and the 75th percentile values (De Vet et al., 2005). Therefore, a value below 1 means that more than 50% of the scores are within one point on the scale (De Vet et al., 2005). On 5-point scales, an IQR of less than 1 is often used as a level of agreement that signals consensus (Raskin, 1994; Ray & Sahu, 1990) while 2 is used as a threshold on a 9-point scale (von der Gracht & Darkow, 2010) or on a 10-point scale (Linstone & Turoff, 1975; Scheibe et al., 2002). No research using both the IQR, and a 7-point scale was found, so a IQR of less than 1,5 will be used as threshold for having reached a satisfying level of agreement. In addition to the IQR, the coefficient of variation (V) will be used to this end. V allows for comparing the distributions of answers on a scale like IQR but is based on the mean (von der Gracht, 2012). It is calculated as the standard deviation divided by the mean multiplied by 100 (English & Kernan, 1976; von der Gracht, 2012) and interpreted as in table 3.

Table 3 Coefficient of Variation (V) and consensus (English & Kernan, 1976; von der Gracht, 2012)

Coefficient of variation (V)	Decision rule
$0 < V \leq 0,5$	Good degree of consensus. No need for additional round
$0,5 < V \leq 0,8$	Less than satisfactory degree of consensus. Possible need for additional round.
$V > 0,8$	Poor degree of consensus. Definite need for additional round.

Stability over rounds

Dajani et al. (1979) discusses that a specific level of agreement can be meaningless when stability over rounds is not reached or not tested. Stability over rounds in this case means the stability of the group's responses on a specific question (von der Gracht, 2012). It will be measured based on the change in level of agreement over rounds (von der Gracht, 2012).

Additionally, inferential statistics are used to test for stability between rounds in a more robust way based on von der Gracht (2012). Given that the same questions are asked to the same experts, the answers are depended (Argyrous, 1997) and of ordinal scale, a Wilcoxon matched-pairs signed-ranks test is used to test if the differences in two sets of answers are statistically significant (von der Gracht, 2012).

Further analysis of dissent

Common explanations for dissent include the presence of outliers, bipolarity in the responses, and different opinions by different stakeholder groups (Beiderbeck et al., 2021a; Warth et al., 2013).

Checking for outliers can be done by converting the scores into standardised z-scores and highlighting those more extreme than the absolute value of 2,58 (the 99% confidence level) (Beiderbeck et al., 2021b; Warth et al., 2013). Checking for bipolarity can be done by visually inspecting the histograms (Beiderbeck et al., 2021b; Warth et al., 2013) and by checking for multiple modes (Beiderbeck et al., 2021b; Scheibe et al., 2002). Different stakeholder groups could hold different perspectives on the same topic due to different interests (Warth et al., 2013). A Mann-Whitney test can be used to assess if statically significant differences between stakeholders exist (Beiderbeck et al., 2021a; Warth et al., 2013).

4. Results

Syntax analysis and non-response bias

The syntax analysis (table 4) shows a high degree of engagement by the experts with a total of 74% of the explanations given in whole sentences. This is a lower than the percentage of whole sentences in Beiderbeck et al. (2021b) (87%) but similar to those in Förster & von der Gracht (2014) (72%) and Roßmann et al. (2018) (78%) and. No large difference exists between part 1 (type of indicators) and part 2 (potential barriers).

Table 4 Syntax analysis (based on Förster & von der Gracht, 2014)

	Total	Percentage
Whole sentences	154	74%
Phrases	19	9%
Catchwords	1	0%
No explanation	34	16%

No strong non-response bias is present as no statistically significant differences are found using a Mann-Whitney U test between the groups when splitting the group in half based on response order. Splitting the group by those who received a reminder and those who did resulted in only one question where a statistically significant difference was found at an alpha of 0.05 (p=0.039).

Descriptive statistics and level of agreement

Tables 5 and 6 show descriptive statistics and the level of agreement of both parts in rounds 1 and 2. In part 1, consensus is reached after two rounds on the majority of the questions (on 16 out of 20 questions). The coefficient of variation is well below the threshold of 0,5 in all questions. The IQR is above the threshold of 1,5 on four criteria, spread out over three types of indicators. In part 2, consensus on five out of ten potential barriers is reached after the second rounds. The remaining five do not reach consensus based on a threshold of 1,5 IQR of which one is the Willingness of people which was only asked in the second round.

Tables 5 and 6 Descriptive statistics and level of agreement

Type of indicator	Criteria	Round 1					Round 2				
		N	Median	IQR	V	Consensus?	N	Median	IQR	V	Consensus?
1: Dutch SPIs	Validity	16	5	1	0,23	Yes	11	5	1	0,19	Yes
	Reliability	16	6	0,25	0,24	Yes	10	6	0	0,11	Yes
	Sensitivity	15	5	1	0,24	Yes	11	5	1,5	0,28	Yes
	Understandability	16	6	1	0,26	Yes	10	6	0	0,09	Yes
2: Proximity based Surrogate Measures of Safety (SMoS)	Validity	15	6	1	0,26	Yes	11	6	1	0,15	Yes
	Reliability	14	5	1	0,25	Yes	10	5	0	0,24	Yes
	Sensitivity	13	5	1	0,23	Yes	11	5	1	0,13	Yes
	Understandability	15	5	**2**	0,38	**No**	11	5	1	0,32	Yes
3: Kinematic based Surrogate Measures of Safety (SMoS)	Validity	15	5	1	0,17	Yes	11	6	1	0,15	Yes
	Reliability	15	6	1	0,22	Yes	11	5	**2**	0,27	**No**
	Sensitivity	13	5	1	0,16	Yes	10	5	0	0,13	Yes
	Understandability	15	5	**2**	0,25	**No**	11	5	**2**	0,26	**No**
4: Engagement of ADAS	Validity	15	6	1,5	0,24	Yes	11	6	1	0,22	Yes
	Reliability	15	6	**2,5**	0,27	**No**	11	5	**2**	0,33	**No**
	Sensitivity	13	5	**2**	0,28	**No**	11	5	1	0,25	Yes
	Understandability	15	6	1	0,18	Yes	11	5	1	0,15	Yes
5: Driver distraction	Validity	-	-	-	-	-	8	5	1	0,17	Yes
	Reliability	-	-	-	-	-	8	4	**2**	0,33	**No**
	Sensitivity	-	-	-	-	-	8	4,5	1,25	0,19	Yes
	Understandability	-	-	-	-	-	9	5	1	0,19	Yes

Potential barrier	Round 1					Round 2				
	N	Median	IQR	V	Consensus?	N	Median	IQR	V	Consensus?
Technical feasibility (collecting data within a single vehicle)	16	4	**2,25**	0,43	**No**	11	4	1	0,33	Yes
Technical feasibility (extracting and processing data from a fleet)	15	4	1	0,35	Yes	11	4	1,5	0,25	Yes
Legal feasibility	15	5	1,5	0,23	Yes	11	5	1	0,21	Yes
Economic feasibility	14	4,5	1	0,33	Yes	10	4	**1,75**	0,35	**No**
Cybersecurity	14	4	**2**	0,33	**No**	11	4	**2**	0,31	**No**
Willingness of OEMs	15	5	1,5	0,27	Yes	10	5	0,75	0,26	Yes
Willingness of suppliers	8	5	**3,25**	0,43	**No**	6	4	**2**	0,39	**No**
Willingness of service providers	12	2	1,25	**0,51**	**No**	9	2	1	0,47	Yes
Willingness of people	-	-	-	-	-	11	5	**2**	0,37	**No**
From pilot to reality	14	4,5	**2,5**	0,37	**No**	10	4,5	**2**	0,39	**No**

From round 1 to 2, the experts made 54 changes out of the 258 possible changes (20,9%). About 35% of the changes were in a positive direction, meaning for part 1 a higher score was given on a criteria for a type of indicator while for part 2 a potential barrier was evaluated as lower. Consequently, 65% of the changes were in a negative direction. Of all the revisions, the large majority (81,5%) were changes of one point on the scale and the remainder of two points in the scale. While the share of revisions made is a little lower than those in Fritschy & Spinler (2019) and Roßmann et al. (2018) with 25% and 37% respectively, the proportions of upwards and downwards revisions are similar (61% up/39% down and 56% up/44% down in Fritschy & Spinler, 2019, and Roßmann et al., 2018).

Stability over rounds

A change occurred in 6 questions out of 25 (16 in part 1, 9 in part 2). Twice this change was from consensus to dissent and four times from dissent to consensus which should ideally be the case for all questions as this signals a greater degree of consensus in the second round.

Using a Wilcoxon signed rank test it is established that in none of the questions, the answers from the first and second round differ in a statistically significant way (alpha = 0,05).

Further analysis of dissent

Nine questions are without consensus. The further analysis found that none of those are due to outliers. The limited size of the different stakeholder groups in this research means that large differences would need to exist in the dataset to achieve statistical significance (Norman, 2010) and none are indeed found.

Multiple modes are found in five of the questions and using the qualitative comments, bipolarity could be confirmed in three of those: cybersecurity, willingness of suppliers, and willingness of people as potential barriers.

4.1 Overview of results part 1

Table 7 shows the median scores given by experts per type of indicator per criteria (on a 7-point scale).

The Dutch SPIs, like Safe speeds and Safe participants, scores the highest on reliability and understandability of all types of indicators. Multiple experts mention that this type of indicator represents "basic safety conditions (…) which are pretty easy to collect and analyse". Especially the example of safe speeds is highlighted by several experts as understandable for policymakers although it does "not give a complete picture on safety" as one expert states: "For example, if one drives at 50kph in a highway, one follows the speed limit, but one also creates a hazard to others."

Proximity based SMoS like TTC score high on validity which is confirmed by several experts highlighting that it is a "very good precursors of crashes" and "represents critical situations well". Like the Dutch SPIs, proximity based SMoS are context dependent, but unlike those, it also needs a threshold which comes with additional problems. One expert states that "deciding what the threshold should be, has a big impact on what is actually measured" and points out that this might need to change over time as "the introduction of AVs might change what we think of as a critical TTC value."

Kinematic based SMoS such as acceleration, deceleration, and swerving did not reach consensus on reliability (IQR = 2, V = 0,27) and on understandability (IQR = 2, V = 0,26). Some experts believe that this can be measured reliable with one expert claiming that: "longitudinal [movement] (acceleration/deceleration) is easier to measure compared to lateral [movement] (swerving)". Other experts disagree: "Due to the nature of these variables, there are many errors and noises in measuring them via accelerometers, and other devices".

Interestingly enough, one academic expert claims that "the scientific evidence for the correlation between harsh acceleration and crashes is weak" while a second academic expert says that "there seems to be research that links the behaviours you listed (e.g., hard breakings) with collisions".

In the case of engagement of ADAS such as FCW, AEB, and BSW, most experts seem to agree that in essence these engagements signal serious potential conflicts and that they are quite easy to measure. However, the experts also see a variety of practical obstacles, mainly that "a warning from FCW will be different from one OEM to another OEM". Additionally, some experts have concerns about the reliability (e.g., AEB/FCW false positive) and that the systems change/improve over time, resulting in several experts identifying harmonisation as a key aspect. Two experts have more fundamental critique on this type of indicator. The activation of these ADAS is based on "some pre-defined threshold

Table 7 Overview of the median scores given by experts per type of indicator per criteria (on a 7-point scale) where an Asterix (*) means no consensus is reached

	Validity	Reliability	Sensitivity	Understandability
Dutch SPIs	5,0	6,0	5,0	6,0
Proximity based SMoS	6,0	5,0	5,0	5,0
Kinematic based SMoS	6,0	5,0*	5,0	5,0*
Engagement of ADAS	6,0	5,0*	5,0	5,0
Driver distraction	5,0	4,0*	4,5	5,0

on indicators of safety like TTC or acceleration". Therefore, they conclude that one might as well just measure those directly.

The suitability of driver distraction as measured by DDAW is only asked about in the second round of the Delphi survey as it was added based on suggestions made by the experts in the first round. Nevertheless, consensus was reached on three out of four criteria with reliability (IQR = 2, V = 0,33) as the exception. This type of indicator has the lowest median scores of all with 4,0 and 4,5 for reliability and sensitivity respectively. Experts point out the large differences between DDAW systems of different OEMs, much more than those discussed in the previous type of indicator. Additionally, "driver distraction results in specific kind of accidents". It would make more sense to relate this to specific locations or types of roads than to use this to measure the safety performance of the entire network, as this "gives insights in where drivers are distracted or what circumstances contribute to distraction.

4.2 Overview of results part 2

Five out of ten potential barriers have a median score of 4 (moderate barrier) and only one is lower, willingness of service providers (2, weak barrier). Willingness of people has a median score of 4,5 while the remaining three have a median score of 5, meaning a somewhat strong barrier. So, most potential barriers identified in the literature review are confirmed by the experts as such, but none of them are seen as strong or even insurmountable barriers

Willingness of OEMs is seen as a somewhat strong barrier by the experts. On one hand, several experts state that OEMs are reluctant to share data, as "they are responsible for the safety of the vehicle and its data". On the other hand, it may just be "a point of economic benefit and regulation" as one expert states. These economic benefits are disputed by one industry expert: "Making €10M is nice, but that is only a spec for the bigger OEMs". Experts point out that regulation already plays a role with EU Act 2013/886 mandating access to road safety data and the fact that the data is owned by the consumer under the Data Act, meaning that "the OEM can only share this data with consent of the consumer, or on a legal base."

This legal feasibility is seen as a somewhat strong to strong barrier because "legislation can make or break the business". The main relevant legislations discussed by experts are the GDPR and EDPB Guidelines for personal data - in which "lots of vehicle data is considered as personal data and needs thus consent of the owner [to be collected and processed]". Additionally, new legislation is underway with the proposed EU Data Act and future sectoral legislation on vehicle data which is currently open for public consultation. Several experts point out that while legislation can be changed, this would require strong EU support and a lot of time and effort. And even then, a legal requirement to provide data and insights may need compensation for OEMs to not threaten the market in the long run, as one expert points out.

In the case of cybersecurity being a potential barrier, all experts agree that there are significant risks, but no consensus is reached due to bipolar opinions. Disagreement lies in how well risks can be contained. On one hand, those who do believe the risks can be minimised refer to "recent advances in cybersecurity and data protection protocols" or to the fact that "the communication of the data is not time critical and vehicle decisions are not based on it". On the other hand, different experts point at data leaks in relation to privacy laws, the consequences for public user acceptance. One expert made the link between the risk of cybersecurity attacks and the willingness of OEMs to facilitate the sharing of data as no OEM would "jeopardize the IT security for a better software service level, at least not now".

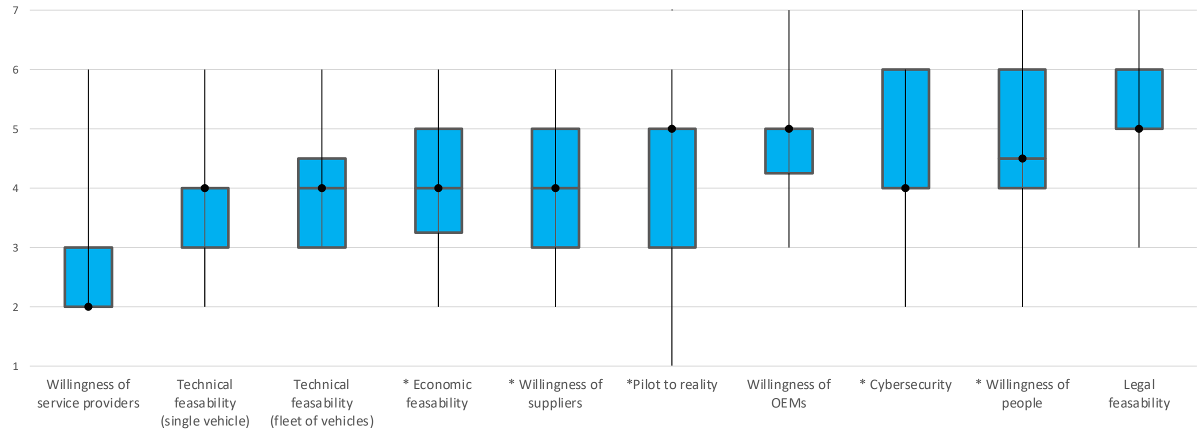


Figure 3 Boxplots potential barriers, where the black dot denotes the median score and the Asterix (*) means no consensus. Scale: 1=No barrier at all, 2=Weak barrier, 3=Somewhat weak barrier, 4=Moderate barrier, 5=Somewhat strong barrier, 6=Strong barrier, 7=insurmountable barrier

Willingness of people as a potential barrier is divisive as well. Those experts who think it is a strong barrier believe that people are not willing to share their data because they will not see the benefits, especially when the data can be linked to traffic violations. Other experts see it more as a moderate barrier with one expert expecting "some discussion at the introduction but people will forget it over time" by making a comparison to the public acceptance of smartphones and the data those collect. Other experts propose incentives for the sharing of their data and point to the fact that not all vehicle owners would need to share their data, only a (representative) part.

Going from pilots to reality is seen by some experts as a weak to somewhat weak barrier while most experts rate it as a somewhat strong barrier. The former group argues that pilots are a good way to start and that "anyone doing a pilot (with reason), has the ambition to scale up". The latter sees this differently and sees this as "a big, steep hill to climb". Besides the fact that all barriers discussed needs to be overcome, it would require harmonization of data and/or interfaces of different OEMs which "is difficult and time consuming".

Two groups of opinions exist on the extent to which the willingness of suppliers a barrier is to the implementation of a system where vehicle data is used to measure traffic safety. One group points out that with a sufficient economic incentive, suppliers will be willing to cooperate. The other group believes that suppliers are unlikely to be willing to participate in any sharing of data as it would threaten their IP and competitive position.

Economic feasibility is seen as a moderate barrier, although no consensus is reached on this potential barrier (IQR = 1,75, V = 0,35) with scores spread quite evenly between 2 and 6. The experts agree that it makes sense for the government to be interested in this data but that "it might be difficult to specify the benefits of network-level safety evaluation to OEMs." Experts do not agree on whether or not a business case for OEMs exists, while this is a crucial factor: "OEMs can stay in a negative business case longer than service providers, but without an eco-system the whole connected industry will struggle". And while the business case improves "if the indicators can be measured with sensors already onboard", an important issue is the role of legislation and the accompanying uncertainty and risks.

The technical feasibility of collecting the data within a single vehicle necessary to report any of the indicators is seen by the experts as a moderate barrier. Most experts agree that it is technically possible to collect data within a vehicle but that it is "depending on the indicator of interest and the sensor suit present in the vehicle". Of course, "there are a lot of existing vehicles on the road that are NOT connected in any way, due to their age". While the experts agree that it should be possible, one expert refers to experience with a recent pilot using the ExVe/Neutral Server concept and claims it "it works ok but has a lot of flaws and loose ends that have not been solved yet." Extracting, processing, and storing the data from a fleet of vehicles from a technical point of view is also seen as a moderate barrier. While several experts point at the required efforts and costs to do this, especially "given the volume and privacy requirements" most experts agree that "it seems very feasible, especially in a country like the Netherlands with great data coverage and a strong network." The difficulty will lie more in "ensuring the data is comparable between different manufacturers" and in "having all OEM installing the technology (especially for low-cost vehicles)".

Willingness of service providers is seen as a weak barrier because "they will see this as new business". This is supported by an industry expert, claiming that "there are more Service Providers than buyers in the connected car industry, including safety data". Additional data from service providers may not even be needed according to another experts, disregarding this as a barrier altogether.

5. Discussion and conclusion

This study explores if a more proactive approach can be viable where data is collected by vehicles equipped with ADAS. In doing so, it contributes to the existing body of knowledge in several ways. Firstly, it looks into what type of leading indicators would be suitable to apply in the specific context of measuring traffic safety at network level and provides an overview of that.

Secondly, this research discusses opportunities for using vehicle sensor data to measure traffic safety and provides an overview of potential barriers for doing so in practice.

But thirdly and most importantly, this research evaluates both several types of indicators and potential barriers for using vehicle sensor data to measure traffic safety at network level at the same time using a Delphi survey. It therefore allows for a fair comparison between different types of indicators for the application in this context. Additionally, in existing literature potential barriers for collecting and using vehicle sensor data is only discussed in a general and limited fashion. This study provides an overview of multiple potential barriers and then builds on it by letting experts evaluates the potential barriers. Because this is done in the same survey by the same experts, it can give an insight into how the potential barriers relate to each other in terms of size.

5.1 Discussion on results

Limited differences are found in the median scores given by the experts between the types of indicators and the criteria. This could that all of the types of indicators evaluated in the Delphi survey could be suitable. It may also be the case that the results are found due to methodological limitations regarding the questions and experts involved. Further research should be undertaken to verify the results found here.

The results regarding potential barriers for collecting and using vehicle sensor data are clearer. Most of the potential barriers initially identified in the literature review are confirmed as such by the experts. However, none of the barriers are rated as insurmountable which provides perspective for implementing any system that uses vehicle sensor data in practice.

Four sizable barriers will need to be overcome of which legal feasibility is the largest. Privacy regulations are discussed by all experts because these dictate if and in what form data can be collected and extracted from vehicles. Uncertainty exists surrounding upcoming regulation as the EC has proposed the Data Act which aims to regulate the access and use of all data generated in the EU across all economic sectors (European Commission, n.d.-b), and a public consultation has just finished on additional legislative measures concerning access to in-vehicle generated data for vehicle-related and mobility services (European Commission, n.d.-a). At this stage, it is not known how the additional sectoral legislation would look like, but it has the potential to have a significant impact on any effort to use vehicle sensor data for measuring traffic safety. This will determine the framework and amount of room available to use vehicle sensor data to measure traffic safety, if any at all.

The other three important barrier that need to be addressed are cybersecurity, the willingness of OEMs to cooperate, as well as that of consumers.

5.2 Wider context

It is important to stress that any type of indicator discussed here is aimed to supplement the current practice of measuring traffic at network level with severe injuries and fatalities.

In proposing to use leading indicators based on vehicle sensor data it is important to keep in mind that only a portion of all traffic is able to collect data. Motorised vehicles such as cars and trucks can have sensors and forms of automation that are able to collect data which could eventually be used to measure traffic safety, but other modes of transport like bicycles, pedestrians, or motorcycles to some extent cannot do that. As a result, it will be difficult to directly measure the traffic safety of these road users. This is not unlike the current situation, but it does come with a bigger problem. Evaluating traffic safety policies based on vehicle sensor data will favour those policies that increase the traffic safety of motorised traffic. It therefore risks that this increased traffic safety could come at the expense of others not represented in these indicators and as a result make it less safe for vulnerable road users. If it would be possible from both a legal and a technical perspective to measure any of the proposed indicator per road type (highway, provincial roads, urban streets, etc), this would be less of an issue. These indictors could then be treated more cautiously on those road types where many vulnerable road users are present.

Furthermore, it be noted that the assumption behind using proximity based SMoS, kinematic based SMoS, and engagement of ADAS as indicators is that these signal traffic conflicts which are proxies for accidents. And while traffic conflicts are regarded as suitable proxies for accidents (Arun et al., 2021), they still have their limitations. The types of indicators evaluated in the Delphi survey may also measure different types of traffic conflicts

5.3 Methodological limitations

Eleven experts completed both rounds of the Delphi study, a number in line with typical ranges suggested in literature (see section 3.1). It is however important to note that given the wide range the backgrounds of experts, the research would have benefited from more experts, especially from the government and industry. This would either have solidified the current opinions or may have added additional input. It would then also be possible to conduct more in-depth stakeholder group analysis.
Assessing the expertise of potential experts in a more sophisticated way would increase the external validity This could be done with self-rating in an intra-individual way as proposed by Ward et al. (2002) or by using deep surface variables as proposed by Spickermann et al. (2014).

5.4 Recommendations for future research

This research is explorative in nature and serves as a starting point for further analysis. This discussion highlights several areas that require further efforts, both in (academic) research as in more practical oriented ways by government.
The results of the Delphi study seem to indicate that all of the four initially included types of indicators could potentially be used to measure traffic safety at network level. However, more scientific research is required to understand:

- Why limited differences are found in the Delphi survey between the different types of indicators.
- Which types of indicators signal which types of accidents. This would on one hand allow for a better understanding of their meaning if they would be measured in reality, and on the other hand allow for a more conscious and informed choice on what (combination of) type of indicators should be measured.
- How valid the relationship is between traffic conflicts measured by these types of indicators and crashes, on different road types.
- Which thresholds would be suitable for any type of indicator per road type.

This research also provides several practical steps that could be undertaken by governments.

- A legal analysis should be developed to identify the legal room for using vehicle sensor data for measuring traffic safety. This should take the proposed Data Act into account, as well as the upcoming additional sectoral legislation on this topic. This would require some time, as these are in development at the moment.
- A discussion should be opened with OEMs (either directly or through ACEA) to better understand how willing they are to cooperate in a system where they will be required to collect and share vehicle sensor data, and under what conditions. This will determine how such a system would be able to exist (economic benefit for the OEMs, because of the OEMs reputations, or if they would need to be forced by legislation, etc.).
- Included in this discussion should be the issue of cybersecurity and in how far the risks related to cybersecurity can be contained. This may require not just the input of the OEMs but also the internal knowledge of the ministry of Infrastructure and Water Management on this topic, as well as external input

from consultancy firms of academic researchers.

- In any implementation of such as system, it is important to take the consumer and its privacy seriously. Allow them to give informed consent of sharing their data and clearly communicate to what end it will be used. Make clear who has access to what data, so it is also clear what can or cannot be done with the data. This will help to increase the willingness of people to cooperate.

5.5 Conclusion

So, the answer to the main research question - (how) can data from sensors of vehicles equipped with ADAS be used to measure traffic safety at network level- is that it depends. If certain conditions are met, it should be possible to collect and use vehicle sensor data to measure traffic safety. The main condition is that there should be legal room available under the proposed Data Act and on the upcoming additional sectoral legislation.

Bibliography

Aarts, L. T. (2018). Prestatie-indicatoren voor verkeersveiligheid (SPI's); Overzicht van beschikbare kennis over SPI's als basis voor risicogestuurd beleid (p. 52). SWOV.

Aarts, L., Wijlhuizen, G. J., Gebhard, S., Goldenbeld, Ch., Decae, R., Bos, N., Bijleveld, F., Mons, C., & Hoekstra, T. (2021). De Staat van de Verkeersveiligheid 2021; Doelstellingen voor 2020 definitief niet gehaald – hoe nu verder? (p. 28). SWOV.

Amundsen, F. H., & Hyden, C. (1977). Proceedings of first workshop on traffic conflicts. Oslo, TTI, Oslo, Norway and LTH Lund, Sweden, 78.

Argyrous, G. (1997). Statistics for social research. Macmillan International Higher Education.

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accident Analysis & Prevention, 153, 106016. https://doi.org/10.1016/j.aap.2021.106016

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021a). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. MethodsX, 8, 101401.https://doi.org/10.1016/j.mex.2021.101401

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021b). The impact of COVID-19 on the European football ecosystem – A Delphi-based scenario analysis. Technological Forecasting and Social Change, 165, 120577. https://doi.org/10.1016/j.techfore.2021.120577

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. Technological Forecasting and Social Change, 147, 72–82. https://doi.org/10.1016/j.techfore.2019.07.002

Blumenthal, M. S., Fraade-Blanar, L., Best, R., & Irwin, J. L. (2020). Safe Enough: Approaches to Assessing Acceptable Safety for Automated Vehicles. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA569-1.html

Chang, A., Saunier, N., & Laureshyn, A. (2017). Proactive methods for road safety analysis. SAE Technical Paper.

Chin, H.-C., & Quek, S.-T. (1997). Measurement of traffic conflicts. Safety Science, 26(3), 169–185. https://doi.org/10.1016/S0925-7535(97)00041-6

Dajani, J. S., Sincoff, M. Z., & Talley, W. K. (1979). Stability and agreement criteria for the termination of Delphi studies. Technological Forecasting and Social Change, 13(1), 83–90.

https://doi.org/10.1016/0040-1625(79)90007-6
De Vet, E., Brug, J., Nooijer, J., Dijkstra, A., & Vries, N. (2005). Determinants of forward stage transitions: A Delphi study. Health Education Research, 20, 195–205. https://doi.org/10.1093/her/cyg111
Devaney, L., & Henchion, M. (2018). Who is a Delphi 'expert'? Reflections on a bioeconomy expert selection procedure from Ireland. Futures, 99, 45–55. https://doi.org/10.1016/j.futures.2018.03.017
Díaz de Rada, V. (2005). Measure and control of non-response in a mail survey. European Journal of Marketing, 39(1/2), 16–32. https://doi.org/10.1108/03090560510571990
Donohoe, H. M., & Needham, R. D. (2009). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. International Journal of Tourism Research, 11(5), 415–437. https://doi.org/10.1002/jtr.709
Ecorys. (2020). Onderzoek naar het delen van voertuigdata en interfaces [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2020/03/23/rapport-delen-voertuigdata-en-interfaces-ecorys
English, J. M., & Kernan, G. L. (1976). The prediction of air travel and aircraft technology to the year 2000 using the Delphi method. Transportation Research, 10(1), 1–8. https://doi.org/10.1016/0041-1647(76)90094-0
ETSC. (2001). Transport Safety Performance Indicators. European Transport Safety Council. https://etsc.eu/transport-safety-performance-indicators/
European Commission. (n.d.-a). Access to vehicle data, functions and resources. Have Your Say. Retrieved 21 August 2022, from https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13180-Access-to-vehicle-data-functions-and-resources/F_en
European Commission. (n.d.-b). Data Act: Measures for a fair and innovative data economy [Text]. European Commission - European Commission. Retrieved 21 September 2022, from https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113
Förster, B., & von der Gracht, H. (2014). Assessing Delphi panel composition for strategic foresight—A comparison of panels based on company-internal and external participants. Technological Forecasting and Social Change, 84, 215–229. https://doi.org/10.1016/j.techfore.2013.07.012
Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). Measuring Automated Vehicle Safety: Forging a Framework. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2662.html
Frewer, L. J., Fischer, A. R. H., Wentholt, M. T. A., Marvin, H. J. P., Ooms, B. W., Coles, D., & Rowe, G. (2011). The use of Delphi methodology in agrifood policy development: Some lessons learned. Technological Forecasting and Social Change, 78(9), 1514–1525. https://doi.org/10.1016/j.techfore.2011.05.005
Fritschy, C., & Spinler, S. (2019). The impact of autonomous trucks on business models in the automotive and logistics industry–a Delphi-based scenario study. Technological Forecasting and Social Change, 148, 119736. https://doi.org/10.1016/j.techfore.2019.119736
Gordon, T., & Pease, A. (2006). RT Delphi: An efficient, "round-less" almost real time Delphi method. Technological Forecasting and Social Change, 73(4), 321–333. https://doi.org/10.1016/j.techfore.2005.09.005
Grimberg, E., Botzer, A., & Musicant, O. (2020). Smartphones vs. in-vehicle data acquisition systems as tools for

naturalistic driving studies: A comparative review. Safety Science, 131, 104917. https://doi.org/10.1016/j.ssci.2020.104917

Groenendijk, J. (2021). Verantwoording Praktijkproef Amsterdam. TwynstraGudde. https://praktijkproefamsterdam.nl/database/documenten/download/595

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. Journal of Advanced Nursing, 32(4), 1008–1015. Scopus. https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x

Henkens, N., Sloot, M., & van Rij, M. (2020). Evaluation of the PoC ‘Data for Road Safety’ (No. SWNL0266365). Sweco. https://www.dataforroadsafety.eu/images/Documenten/Microsoft_Word_-_PoC_DTF_-_monitor__evaluation_report_Sweco.pdf

Hsu, C.-C., & Sandford, B. (2007). The Delphi Technique: Making Sense Of Consensus. Practical Assessment, Research and Evaluation, 12.

Johnsson, C., Laureshyn, A., & Dágostino, C. (2021). A relative approach to the validation of surrogate measures of safety. Accident Analysis & Prevention, 161, 106350. https://doi.org/10.1016/j.aap.2021.106350

Johnsson, C., Laureshyn, A., & De Ceunynck, T. (2018). In search of surrogate safety indicators for vulnerable road users: A review of surrogate safety indicators. Transport Reviews, 38(6), 765–785. https://doi.org/10.1080/01441647.2018.1442888

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. Journal of Advanced Nursing, 53(2), 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x

Kia. (2021). CONCEPT REPORT: Insigts from vehicle data. Kia Nederland, Ministry of Infrastructure and Watermanagement.

Kia Nederland. (2021, November 25). Kia verkent inzichten in voertuigdata met Ministerie van IenW. Kia verkent inzichten in voertuigdata met Ministerie van IenW. https://nieuws.kia.nl/kia-verkent-inzichten-in-voertuigdata-met-ministerie-van-ienw/

Landeta, J. (2006). Current validity of the Delphi method in social sciences. Technological Forecasting and Social Change, 73(5), 467–482. https://doi.org/10.1016/j.techfore.2005.09.002

Laureshyn, A., Johnsson, C., De Ceunynck, T., Svensson, Å., de Goede, M., Saunier, N., Włodarek, P., van der Horst, R., & Daniels, S. (2016). Review of current study methods for VRU safety. Appendix 6–Scoping review: Surrogate measures of safety in site-based road traffic observations: Deliverable 2.1–part 4.

Linstone, H., & Turoff, M. (1975). The Delphi Method: Techniques and Applications. In Technometrics (Vol. 18). https://doi.org/10.2307/3150755

Mahmud, S. M. S., Ferreira, L., Hoque, Md. S., & Tavassoli, A. (2017). Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS Research, 41(4), 153–163. https://doi.org/10.1016/j.iatssr.2017.02.001

Markmann, C., Spickermann, A., von der Gracht, H. A., & Brem, A. (2021). Improving the question formulation in Delphi-like surveys: Analysis of the effects of abstract language and amount of information on response behavior. FUTURES & FORESIGHT SCIENCE, 3(1), e56. https://doi.org/10.1002/ffo2.56

Mauksch, S., von der Gracht, H. A., & Gordon, T. J. (2020). Who is an expert for foresight? A review of identification methods. Technological Forecasting and Social Change, 154, 119982.

https://doi.org/10.1016/j.techfore.2020.119982
McKinsey. (2018). From buzz to bucks—Automotive players on the highway to car data monetization. McKinsey.
Ministerie van Infrastructuur en Watermanagemant. (2018). Veilig van deur tot deur. Het Strategisch Plan Verkeersveiligheid 2030: Een gezamenlijke visie op aanpak verkeersveiligheidsbeleid. https://www.rijksoverheid.nl/documenten/rapporten/2018/12/05/bijlage-1-het-strategisch-plan-verkeersveiligheid-2030-veilig-van-deur-tot-deur
Norman, G. (2010). Likert scales, levels of measurement and the 'laws' of statistics. Advances in Health Sciences Education: Theory and Practice, 15(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y
Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. Information & Management, 42(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002
Onderzoeksraad voor Veiligheid. (2019). Wie stuurt? Verkeersveiligheid en automatisering in het wegverkeer. http://www.onderzoeksraad.nl/nl/page/4729/wie-stuurt-verkeersveiligheid-en-automatisering-in-het-wegverkeer
Piecyk, M. I., & McKinnon, A. C. (2010). Forecasting the carbon footprint of road freight transport in 2020. International Journal of Production Economics, 128(1), 31–42. https://doi.org/10.1016/j.ijpe.2009.08.027
Raskin, M. S. (1994). The Delphi Study in Field Instruction Revisited: Expert Consensus on Issues and Research Priorities. Journal of Social Work Education, 30(1), 75–89. https://doi.org/10.1080/10437797.1994.10672215
Ray, P. K., & Sahu, S. (1990). Productivity Management in India: A Delphi Study. International Journal of Operations & Production Management, 10(5), 25–51.https://doi.org/10.1108/01443579010005245
Regulation (EU) No 886/2013, (2013). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32013R0886
Roßmann, B., Canzaniello, A., von der Gracht, H., & Hartmann, E. (2018). The future and social impact of Big Data Analytics in Supply Chain Management: Results from a Delphi study. Technological Forecasting and Social Change, 130, 135–149. https://doi.org/10.1016/j.techfore.2017.10.005
Ryan, C. (2020). Emerging autonomous vehicle risks: The role of telematics and machine learning based risk assessment.
Scheibe, M., Skutsch, M., & Schofer, J. (2002). IV. C. Experiments in Delphi methodology. The Delphi Method: Techniques and Applications, 257–281.
Schuckmann, S. W., Gnatzy, T., Darkow, I.-L., & von der Gracht, H. A. (2012). Analysis of factors influencing the development of transport infrastructure until the year 2030—A Delphi based scenario study. Technological Forecasting and Social Change, 79(8), 1373–1387. https://doi.org/10.1016/j.techfore.2012.05.008
Spickermann, A., Zimmermann, M., & von der Gracht, H. A. (2014). Surface- and deep-level diversity in panel selection—Exploring diversity effects on response behaviour in foresight. Technological Forecasting and Social Change,105–120. https://doi.org/10.1016/j.tecfore.2013.04.009
Stipdonk, H. L. (2013). Road safety in bits and pieces; for a better understanding of the development of the number of road fatalities. https://repository.tudelft.nl/islandora/object/uuid%3A435ca957-1b52-446c-a325-f12c433a99d9
SWOV. (2005). State of the art Report on Road Safety Performance Indicators;

Deliverable D3.1 of the EU FP6 project SafetyNet (p. 177). European Commission.
SWOV. (2016). Gegevensbronnen; Uitgebreid overzicht. SWOV. http://www.swov.nl/sites/default/files/bestanden/wegwijzer/gegevensbronnen.pdf
SWOV. (2020a). Verkeersdoden in Nederland [SWOV-Factsheet, april 2020.]. SWOV. https://www.swov.nl/feiten-cijfers/fact/verkeersdoden-hoe-heeft-het-aantal-verkeersdoden-nederland-zich-sinds-1950
SWOV. (2020b). De Staat van de Verkeersveiligheid 2020: Doelstellingen 2020 worden niet gehaald. SWOV. https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2020/12/14/bijlage-1-de-staat-van-de-verkeersveiligheid-2020-swov. pdf
Talal, M., Ramli, K. N., Zaidan, A. A., Zaidan, B. B., & Jumaa, F. (2020). Review on car-following sensor based and data-generation mapping for safety and traffic management and road map toward ITS. Vehicular Communications, 25, 100280. https://doi.org/10.1016/j.vehcom.2020.100280
Tarko, A. P. (2012). Use of crash surrogates and exceedance statistics to estimate road safety. Accident Analysis & Prevention, 45, 230–240. https://doi.org/10.1016/j.aap.2011.07.008
Toma, C., & Picioreanu, I. (2016). The Delphi Technique: Methodological Considerations and the Need for Reporting Guidelines in Medical Journals. International Journal of Public Health Research, 4, 47–59.
TRL. (2017). Access to In-vehicle Data and Resources; Final Report (No. CPR2419). https://ec.europa.eu/transport/themes/its/studies/its_en. transport.its.menu
Vlakveld, W. P. (2019). Veiligheidseffecten van rijtaakondersteunende systemen. SWOV.
von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. Technological Forecasting and Social Change, 79(8), 1525–1536.https://doi.org/10.1016/j.techfore.2012.04.013
von der Gracht, H. A., & Darkow, I.-L. (2010). Scenarios for the logistics services industry: A Delphi-based analysis for 2025. International Journal of Production Economics, 127(1), 46–59. https://doi.org/10.1016/j.ijpe.2010.04.013
Wang, L., Zhong, H., Ma, W., Abdel-Aty, M., & Park, J. (2020). How many crashes can connected vehicle and automated vehicle technologies prevent: A meta-analysis. Accident Analysis & Prevention, 136, 105299. https://doi.org/10.1016/j.aap.2019.105299
Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. Advances in Health Sciences Education, 7(1), 63–80. Scopus.https://doi.org/10.1023/A:1014585522084
Warth, J., von der Gracht, H. A., & Darkow, I.-L. (2013). A dissent-based approach for multi-stakeholder scenario development — The future of electric drive vehicles. Technological Forecasting and Social Change, 80(4), 566–583.https://doi.org/10.1016/j.tecfore.2012.04.005
Winkler, J., & Moser, R. (2016). Biases in future-oriented Delphi studies: A cognitive perspective. Technological Forecasting and Social Change, 105, 63–76. https://doi.org/10.1016/j.techfore.2016.01.021
Young, W., Sobhani, A., Lenné, M. G., & Sarvi, M. (2014). Simulation of safety: A review of the state of the art in road safety simulation modelling. Accident Analysis & Prevention, 66, 89–103. https://doi.org/10.1016/j.aap.2014.01.0

Appendix B Safety evaluation methods for vehicles with automation

This appendix gives an overview of safety evaluation methods for vehicles with automation in section B.1 and of data collection methods used in those methods in section C.2.

B.1 Safety evaluation methods for vehicles with automation

That there is no standard way to measure safety effects is illustrated by the fact that studies into the assessment of safety effects of ADAS find different types of assessment methods. For example, Sohrabi et al. (2021) identifies six different safety assessment methods while Vlakveld (2019) and Yue et al. (2018) find five. Table B1 shows these assessment methods which will be discussed in more detail below.

Table B1 Safety evaluation methods for vehicles equipped with automation as discussed in three review studies where the percentage show the share of studies in the review that used that method

		Vlakveld (2019)	Yue et al. (2018)	Sohrabi et al. (2021)	Setting
Road test data analysis		x	x (28%)	x (30%)	Public roads
Driving Test Study (DTS)	Field Operational Test (FOT)	x	x (28% for all DTS combined)		Public roads
	Naturalistic Driving Study (NDs)		x		Public roads
	Driving simulator	x	x	x (21%)	Artificial
Test Track tests		x			Artificial
Traffic simulation				x (27%)	Artificial
Target crash population		x	x (44%)	x (18%)	Artificial
Safety effectiveness				x (2%)	Artificial
System failure risk assessment				x (2%)	Artificial

Road test data analysis

Road test data analysis is one of the three safety assessment methods all three studies discuss. Road test data analysis are studies that analyse reports and statistics on crashes or disengagements of vehicles with and without ADAS driving on roads (Sohrabi et al., 2021; Vlakveld, 2019; Yue et al., 2018). By comparing these crash rates, it can be estimated if the vehicles with ADAS are safer than those without. All three studies see this as the most reliable way of estimating safety effects of vehicles with ADAS. However, they also agree that there are limitations in the data available. Sohrabi et al. (2021) names the small size of datasets and the under reporting of crashes of human driven vehicles (HDVs). Vlakveld (2019) points at the fact that it is usually not known how many kilometres both types of vehicles (with and without ADAS) drive, making for an unfair comparison. This also has to do with the potential self-selection bias in which different type of drivers by different types of cars (Vlakveld, 2019). Finally, Yue et al. (2018) mention that this methodology is only possible for ADAS already present in the market in sufficient numbers.

Within this type of assessment methods, Sohrabi et al. (2021) distinguish two more types: research into the characteristics crashes of vehicles with driving automation systems and research into the safety reliability of such vehicles. Examples of the former can vary from the type of crashes (Favarò et al., 2017) to factors influencing the likelihood of crashing (Boggs et al., 2020). An examples of the latter is the study by Kalra & Paddock (2016) as discussed in section 1.2 and figure 3 which aims to calculate how many kilometres need to be driven by vehicles with driving automation systems in order to demonstrate their reliability relative to human driven vehicles.

Road test data analysis is the only ex-post safety evaluation method discussed is this paragraph. All the other methods are ex-ante safety evaluation methods which aim to evaluate the safety effect of ADAS before they are released on the market.

Driving Test Studies (DST)
Driving Test Studies (DST) take a different approach by organising various types of tests. Yue et al. (2018) classifies Field Operational Tests (FOTs), Naturalistic Driving Study (NDS) and Driving simulator tests as DSTs. In all three of these types of experiments, drivers drive vehicles outfitted with specific ADAS and with measuring equipment so the impact of these ADAS can be measured (Yue et al., 2018). Both FOTs and NDS take place on real roads in daily life. The difference between the two is that in FOTs, drivers get certain instructions like when to use the ADAS and when to turn them off. This usually follows an A-B test design in which the drivers first drive without the ADAS (test A) and then with the ADAS (test B). In this way, the change in traffic safety, in whatever chosen metric, can be measured. (Yue et al., 2018).
As can be seen in figure B1, there is an overlap between FOT and NDS. In an NDS, people can drive in any way they prefer, without specific instructions while in NDS people drive in any way they prefer (FESTA, 2018; SWOV, 2012). Note that Vlakveld (2019) does not make the distinction between FOTs and NDS. The advantage of NDS over FOTs is that NDS resembles reality closer and that factors such as usage are considered (SWOV, 2012).
A disadvantage of NDS is the high costs of outfitting vehicles with measuring equipment resulting in relatively short duration of these studies (Vlakveld, 2019). This is problematic because crashes are rare and thus might not occur during the tests (Vlakveld, 2019).

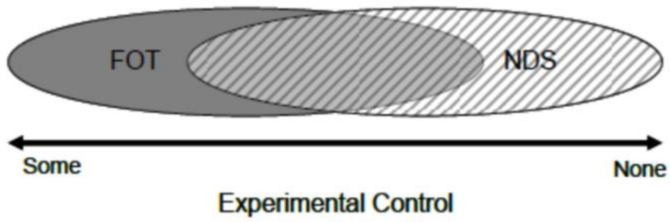


Figure B1 The partial overlap between FOT and NDS (FESTA, 2018)

This problem can be overcome by using a driving simulator. A driving simulator test is similar to FOTs, except that it takes place in a virtual environment. This allows the researchers to control many variables and create specific situations that are rare in reality. Yue et al. (2018) for example, use a driving simulator to test the effect of fog on the effectiveness of FCW. The disadvantage of driving simulators is that there might be bias in for example the participants of in the simulated environment leading to less valid results (Sohrabi et al., 2021). This makes it more difficult to generalise results from driving simulator test to real world driving (Vlakveld, 2019).

Test Track tests
ADAS can also be tested on a test track, a circuit or road section closed off to other drivers. Here, specific situations that might be rare in reality can be created to see how ADAS and/or the drivers respond to this situation (Vlakveld, 2019). This is similar to the driving simulator but in some case more valid because the drivers are in actual cars instead of a simulator (Vlakveld, 2019).

Traffic simulation
To assess the impact of ADAS on traffic safety on a higher level, traffic simulations can be used. These can for example be used to explore the impact of market penetration rates of ADAS on safety (Sohrabi et al., 2021). Like driving simulator tests, traffic simulations might suffer from similar biases. Additionally, these simulations have to make assumptions on the effectiveness of specific ADAS, which

are based on previous research. In this way, the simulation inherits the shortcoming of those studies and make it difficult to validate the simulations (Sohrabi et al., 2021).

Target crash population

The target crash population approach aims to quantify the number of preventable crashes if ADAS were implemented. According to Sohrabi et al (2021), it consists of three steps:

1. Identify the functionality of the ADAS
2. Match this functionality with a specific target crash type
3. Calculate the number of preventable crashes in the crash dataset

Vlakveld (2019) classifies this as the least valid safety assessment method because potential changes in driving behaviour when driving with these ADAS are not considered and because 100% effectiveness of these systems are assumed. Sohrabi agrees and adds that potential new types of crashes due to interaction in mixed traffic are not considered. This together may lead to an overestimation of the effect of ADAS on safety and thus is more likely to represent the theoretical upper bound (Sohrabi et al., 2021). This method is mostly used for ADAS that have just entered the market or are still in development, making the much more valid road test safety analysis not possible (Vlakveld, 2019).
In the interpretation of Yue et al. (2018), the effectiveness of the ADAS is not assumed to be 100% but based on previous NDS. A step is added in which the information about the functionality of the ADAS and the crash database is used to reconstruct a pre-crash scenario in kinematic models. This is then used to calculate the number of crashes avoided. On one hand, this is more valid because it does not assume a 100% effectiveness of ADAS but on the other hand the single near-crash event has no meaning because it is the result of a simulation, not an actual event (Yue et al., 2018).

System failure risk assessment

Bhavsar et al. (2017) have tried to calculate the reliability of vehicles with driving automation systems by determining the failure rate of individual components from the literature. These are then synthesised according to a certain hierarchical model. This method is criticised by Sohrabi et al. (2021) because the failure risks can be overestimated due to redundancy in the system and because it is difficult to determine failure rates accurately.

Safety effectiveness

Safety effectiveness is a way to synthase results from different studies by using several formulas to normalise the effect of the ADAS on safety (Sohrabi et al., 2021). In this way, several studies can be compared. The use of method may be limited because it can still only compare studies based on the same metrics. Additionally, this method uses other driver or traffic simulation studies and thus inherits their limitations and flaws. (Sohrabi et al., 2021)

B.2 Data collection for safety evaluation methods

Several ways of collecting data for safety evaluation methods exist. A common method is to use crash statistics, as is done in road test data analysis (Young et al., 2014). This data is collected after a crash and these crash statistics databases are often incomplete. There are however more fundamental issues. According to Tarko (2012) the crash causality cannot be properly understood based on crash statistics, only contributing factors to the crash. This is because of the lack of detailed driving data. As a result, basing research on these crash statistics can sometimes lead to results opposite of those in reality (Talal et al., 2020).

A second way in which data can be collected is with observational data (Young et al., 2014) or as Talal et al. (2020) call it, with fixed point sensors. This can include already present loop-detectors and surveillance cameras, purposely placed cameras and/or trained observers (Talal et al., 2020; Young et al., 2014). This method can observe and help understand driver behaviour, and can collect data on near-crashes which crash data cannot do (Young et al., 2014). It is however not without its issues: it is difficult to collect data on the characteristics of the driver and the vehicle (Young et al., 2014) and the sensors could be costly and maintenance intensive (Talal et al., 2020). Another major drawback of this method is that these sensors are in a permanent location resulting in data gathered from limited locations and with less variety (Talal et al., 2020; Young et al., 2014).

This leads to the third possible way of collecting data: on-board data collection as is often done in NDS. Collecting the necessary data can be done with several different data acquisitions systems (DAS). The first DAS to be discussed is in-vehicle research-designated DAS, in which vehicles are equipped/retrofitted with a large variety of fixed sensors like GPS, forward-facing radar and several cameras (Grimberg et al., 2020). This allows for low measurement error, high sampling rate and also the possibility of researchers to customise the DAS to their research question (Grimberg et al., 2020; Young et al., 2014). This is however a highly costly method, the Australian Naturalistic Driving Study (ANDS) with 379 drivers costed $4 million. This is not just the cost of the equipment but also the costs of processing the data as the video footage often still needs to be processed and coded manually (Grimberg et al., 2020; Talal et al., 2020).
These high costs have resulted in a second DAS: smartphone-based DAS. Smartphones might be a cheaper alternative to dedicated in-vehicle DAS given that smartphones are widely spread among the population and that these often have built-in GPS, accelerometer, gyroscope, and are capable of storing and transmitting data from these sensors (Grimberg et al., 2020; Young et al., 2014). According to Talal et al. (2020) these sensors are too unreliable although this is disputed by Grimberg et al. (2020) who state that the measurement error is not too different from dedicated in-vehicle DAS and that computational corrections can help compensate for the small measurement errors. Nonetheless, both papers point to issues around data incompleteness and validity due to limited available storage, battery drainage, or the driver's concern about these two issues. Additionally, using only smartphones as DAS does almost always not allow for video footage. This together makes smartphone DAS not suitable for crash and near-crash analysis (Grimberg et al., 2020)

So, dedicated in-car DAS are capable of providing detailed data suited for the safety evaluation of several ADAS but comes at high costs. Using smartphones as DAS is much cheaper but this method is not capable of providing detailed and viable enough data. This poses a problem as this data is necessary as is shown in this review. As Talal et al. (p25, 2020) concludes about DAS: "providing a low-cost, reliable and easy-to-implement system is a tremendous step towards research advancement".

To conclude this paragraph, data for the safety evaluation of ADAS is collected in various ways, depending on the safety evaluation method used. On-board data collection is found to have several key advantages over the two other discussed methods, crash statistics and observational data. The former is not capable of identifying crash causality and does by definition only concern actual crashes while the latter can only collect data on fixed locations and is not able to identify driver characteristics. On-board data collection is able cover all these factors but only with - costly - dedicated research equipment.

Appendix C Interview summary on RoadMonitor (RoMo)

This is a summary of an interview conducted with Erik Vrijens (Ministry of Infrastructure and Water management) on the Road Monitor (RoMo) (06-12-21)

The Road Monitor (RoMo) is a project of the Dutch Ministry of Infrastructure and Water management in which vehicle data is processed into useful information for the road authorities. The project focusses on three areas: asset management, winter-maintenance, and detecting unsafe situations.
Detecting unsafe situations is done based on data collected by the sensors present in the vehicles. This can be the input of drivers like sudden braking or abrupt steering wheel motions, or it can be the intervention of specific ADAS like Forward Collision Warning (FCW). An algorithm is developed which combines these inputs to detect potentially unsafe situations. What is special about this project that it not only detects potentially dangerous situations but also aims to provide context. This context comes from an algorithm that analyses a clip made by the vehicle's cameras of the 10 seconds leading up to the situation. From this clip, the cause of the unsafe situation can be determined.
The information is collected by all Mercedes vehicles of which the owners have given consent and also have the necessary sensors and communication technology. Mercedes is also the party that processes and analyses the data. This is then aggregated or anonymised and delivered to the Ministry of I&W in form the form of an interactive dashboard. In this way it is easier to guarantee the privacy of the drivers.
Even though RoMo only uses vehicle data of Mercedes, there is still enough data to sufficiently cover most of the Netherlands. The main issue with using data from a single manufacturer is that only the specific ADAS of that specific manufacturer are evaluated. This means that it may not be possible to generalise results to all vehicles. In addition, other manufacturers produce vehicles with slightly different ADAS or even ADAS with different functions.

Appendix D Delphi survey round 1

Part 1

For this first part of this questionnaire, assume that vehicle manufactures are willing to supply the necessary vehicle data to the Dutch ministry of Infrastructure and Water Management and that they are capable of doing so. The goal of this part is to evaluate different types of indicators that could be used to measure traffic safety at network level. The goal of measuring traffic safety at network level here means to measure the level of road safety at a network, for example the network managed by Rijkswaterstaat (*hoofdwegennet).*

This part will consist of five questions. In the first four questions, you will be asked to evaluate a type of indicator to measure traffic safety on four different criteria. Examples of the types of indicators are provided to give you an idea of what could potentially be used as indicator.

Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree) and explain why you make this assessment. Table 1 below shows the meaning of each score. Please answer a 0 in case you do not have an answer.

All parts that require a response from you are marked in blue.

Question 1

Type of indicator: Dutch Safety Performance Indicators (SPIs), like Safe Speeds and Safe Participants.

Examples 1 Safe Speeds: The share of motorized traffic that does not exceed the speed limit (per road type) 2 Safe Participants: Share of vehicles (per type) that has their correct lights on (per visibility condition) 3 Safe Participants: Share of drivers that wears their seatbelt	
a. This type of indicator reflects traffic safety well	*Score*
b. This type of indicator could be measured in a reliable way	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	*Score*
Explanation	

Table 1 Meaning of the scores in part 1

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

Question 2

Type of indicator: Proximity based Surrogate Measures of Safety (SMoS) such as time-to-collision (TTC).

SMoS = pro-active (leading) indicator that use traffic conflicts as surrogate for actual collisions.
TTC = the time remaining until a collision between two vehicles would occur if they maintained their course and speed

Examples 1. The number of times TTC exceeds a certain threshold per 100.000 kilometers travelled (per road type)	
a. This type of indicator reflects traffic safety well	*Score*
b. This type of indicator could be measured in a reliable way	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	*Score*
Explanation	

Table 1 Meaning of the scores in part 1

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

Question 3

Type of indicator: Kinematic Surrogate Measures of Safety (SMoS), like strong deceleration, acceleration or swerving.

SMoS = pro-active (leading) indicator that use traffic conflicts as surrogate for actual collisions.

Examples 1. The number of strong decelerations per 100.000 kilometers travelled (per road type) 2. The number of strong accelerations per 100.000 kilometers travelled (per road type) 3. The number of strong swerving movements per 100.000 kilometers travelled (per road type)	
a. This type of indicator reflects traffic safety well	*Score*
b. This type of indicator could be measured in a reliable way	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	*Score*
Explanation	

Table 1 Meaning of the scores in part 1

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

Question 4

Type of indicator: Engagement of ADAS, like Forward Collision Warning (FCW), Autonomous Emergency Braking (AEB) or Blind Sport Warning (BSW).

Engagement of ADAS = the moment an ADAS acts, either by warning the driver or by actively intervening

Examples 1. The number of warnings of Forward Collision Warning (FCW) per 100.000 kilometers travelled (per road type) 2. The number of interventions of Autonomous Emergency Braking (AEB) per 100.000 kilometers travelled (per road type) 3. The number of warnings of Blind Sport Warning (BSW) when the driver changes lanes per 100.000 kilometers travelled (per road type)	
a. This type of indicator reflects traffic safety well	*Score*
b. This type of indicator could be measured in a reliable way	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	*Score*
Explanation	

Table 1 Meaning of the scores in part 1

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

Question 5

Are there any other types of indicators based on vehicle data not mentioned above that you think could potentially be a suitable indicator for measuring traffic safety at network level? If this is the case, please explain below.

Explanation

End of part 1

Part 2

This part of the questionnaire is aimed at identifying what factors could potentially form a barrier for successfully implementing a system of measuring traffic safety at network level based on vehicle data. This does not specifically refer to any of the in part 1 discussed types of indicators but is meant in a more general way.

This part consists of seven questions. In these questions, you will be asked to evaluate a potential barrier to successful implementation of a system of measuring traffic safety at network level based on vehicle data.

Please answer the questions below by scoring the potential barrier from 1 (no barrier at all) to 7 (Insurmountable barrier) and explain why you make this assessment.

Table 2 below shows the meaning of each score. Please answer a 0 in case you do not have an answer. All parts that require a response from you are marked in blue.

Table 2 Meaning of the scores in part 2

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 1

Technical feasibility

1a. Is the technical feasibility of **collecting the data within one vehicle** a barrier to successful implementation?

This refers to the technical feasibility of collecting the data necessary for reporting any of the indicators discussed in part 1, within a single vehicle.

Score
Explanation

1b. Is the technical feasibility of **extracting and processing the data** of a fleet of vehicles a barrier to successful implementation?

This refers to the technical feasibility of the entire process of extracting, processing, and storing the data necessary to report one of the indicators discussed in part 1 for a fleet of vehicles.

Score
Explanation

1c. Would the answers given at 1a and 1b be different for any of the specific types of indicators discussed in part 1? If this is the case, please explain below.

Explanation

Question 2

Is **Legal feasibility** a barrier to successful implementation?

This refers to how (EU) laws or directives on for example privacy or fair market competition could be a potential barrier.

Score
Explanation

Table 2 Meaning of the scores in part 2

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 3

Is **Economic feasibility** a barrier to successful implementation?

This refers to the question of whether or not a business case exists for relevant stakeholders (OEMs, government, service providers, etc.) that would make them willing to participate or organise such a system? Would it thus be possible to organise such a system with several stakeholders?

Score
Explanation

Table 2 Meaning of the scores in part 2

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 4

Is **Cybersecurity** a barrier to successful implementation?

This refers to the cybersecurity during the whole process: within a vehicle, during the over-the-air communication, and during the processing of the data.

Score
Explanation

Table 2 Meaning of the scores in part 2

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 5

Willingness of stakeholders

Is Willingness of **OEMs/vehicle manufacturers** a barrier to successful implementation?

Score
Explanation

Is Willingness of **suppliers** a barrier to successful implementation?

Score
Explanation

Is Willingness of **Service Providers** a barrier to successful implementation?

Score
Explanation

Table 2 Meaning of the scores in part 2

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 6

Is **Going from pilot to reality** a barrier to successful implementation?

This refers to the organisational difficulty of stakeholders to go from small scale pilots to the full incorporation of such a system in their organisations.

Score
Explanation

Table 2 Meaning of the scores in part 2

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 7

Are there any other factors not mentioned above that you think could potentially be a barrier for implementing a system of using vehicle data to measure traffic safety? If this is the case, please explain below.

Explanation

Question 8

Do you have any other comments or thoughts that you wish to share?

Explanation

End of the Delphi survey

Thank you for your participation! Please save this file and sent it back to me. By the end of June, you will receive the second and final round of the survey. Please fill in that survey as well.

Appendix E Delphi survey round 2

Instruction for round 2

Dear expert,

Thank you for participating in the second round of the Delphi study.

This second and final round of the survey will consist of the same questions as round 1. Additional information is given, and you are asked to re-evaluate your answers based on this information. Each question has the following:

1. The original question of round 1 with your own scores
2. A histogram showing the distribution of scores given by all the experts in round 1
3. Arguments against the type of indicator/potential barrier made by experts in round 1
4. Arguments in favour of the type of indicator/potential barrier made by experts in round 1
5. The question to re-evaluate you score from round 1
6. The meaning of the scores

You are asked to re-evaluate your answer and see if you want to change your answer based on the additional information. It is important to note that you do not have to change your answer.

Similar to round 1, there is room to explain the score you give. All parts that require a response from you are marked in *blue*.

Please fill in all the questions of this second round as well.

Instruction for round 2

For this first part of this questionnaire, assume that vehicle manufactures are willing to supply the necessary vehicle data to the Dutch ministry of Infrastructure and Water Management and that they are capable of doing so. The goal of this part is to evaluate different types of indicators that could be used to measure traffic safety at network level.

The goal of measuring traffic safety at network level here means to measure the level of road safety at a network, for example the network managed by Rijkswaterstaat (*hoofdwegennet).*

This part will consist of five questions. You will be asked to evaluate a type of indicator to measure traffic safety on four different criteria. Examples of the types of indicators are provided to give you an idea of what could potentially be used as indicator.

Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree) and explain why you make this assessment. Table 1 below shows the meaning of each score. Please answer a 0 in case you do not have an answer.

All parts that require a response from you are marked in *blue*.

Table 1 Meaning of the scores in part 1

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

Clarification criteria c: sensitivity

Several experts asked for clarification on the third criterium:

- c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions

This the goal of this criterium is to answer the question if the indicator would change if policies aimed at improving traffic safety are implemented in the future? So, it aims to clarify if the indicator can measure traffic safety in a relative way.

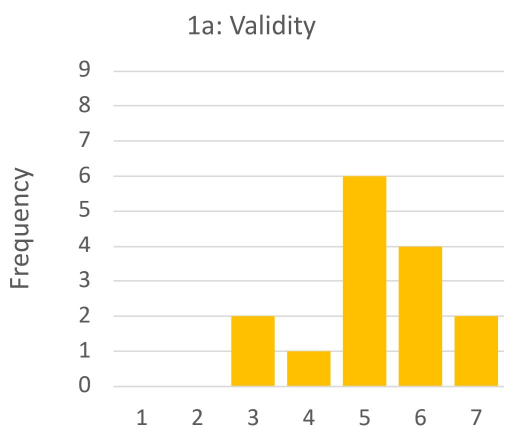
An example about an indicator focussed on speed: let's say that in the coming years action is taken to reduce speeding in any way (adapting infrastructure, the further deployment of ISA, or by increasing enforcement, etc.), will the indicator show a change?

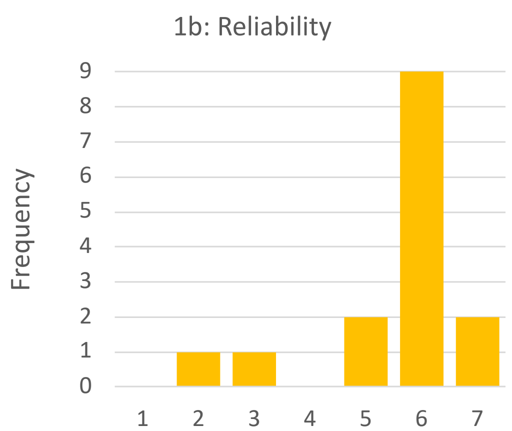
Question 1 Dutch Safety Performance Indicators (SPIs)

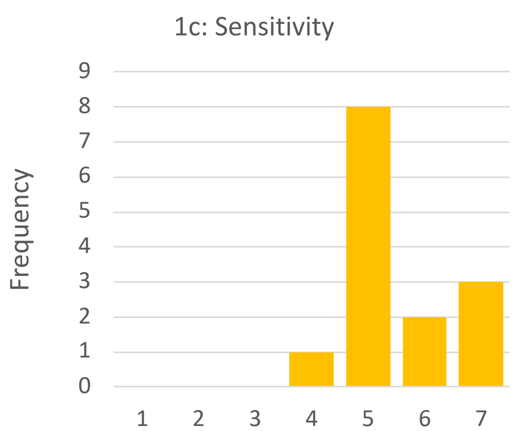
1. The original question 1: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

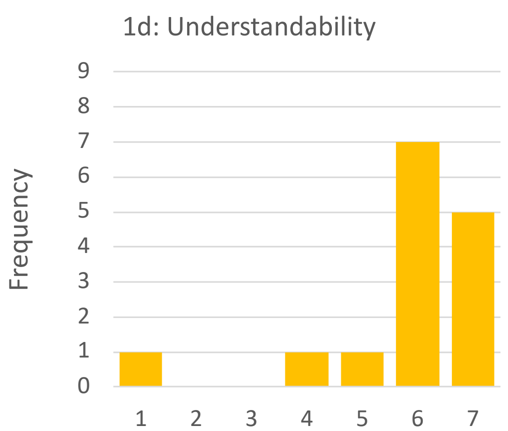
Type of indicator • Dutch Safety Performance Indicators (SPIs)	
Examples of specific metrics • Safe Speeds: The share of motorized traffic that does not exceed the speed limit (per road type) • Safe Participants: Share of vehicles (per type) that has their correct lights on (per visibility condition) • Safe Participants: Share of drivers that wears their seatbelt	
	Your score round 1
a. This type of indicator reflects traffic safety well	5
b. This type of indicator could be measured in a reliable way	6
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6

2. Distribution of scores in round 1









3. Arguments made by experts against this type of indicator:

- A safe speed is not always comparable with the speed limit: think of the external conditions of safe speed, such as weather or road works.
- Monitoring all road users and measuring such a proportion is very time consuming and so assumptions are needed (e.g., extrapolation). This may lead to measurement error in data.
- This type of indicator describes the safety risks from the perspective of policy makers, which does not consider the operational features of traffic dynamics. So, they are more for describing the general background/conditions of safety, instead of pinpointing the safety itself.

4. Arguments made by experts in favour of this type of indicator

- This kind of indicators represent basic safety conditions. They are pretty easy to collect and analyse.
- Traffic speed is a good indicator of safety and severity of crash outcomes. It is also understandable for end users.

5. Re-evaluate question 1: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

Type of indicator • Dutch Safety Performance Indicators (SPIs)		
Examples of specific metrics • Safe Speeds: The share of motorized traffic that does not exceed the speed limit (per road type) • Safe Participants: Share of vehicles (per type) that has their correct lights on (per visibility condition) • Safe Participants: Share of drivers that wears their seatbelt		
	Score R1	**Score R2**
a. This type of indicator reflects traffic safety well	5	*Score*
b. This type of indicator could be measured in a reliable way	6	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6	*Score*
Explanation (optional)		

6. Meaning of the scores

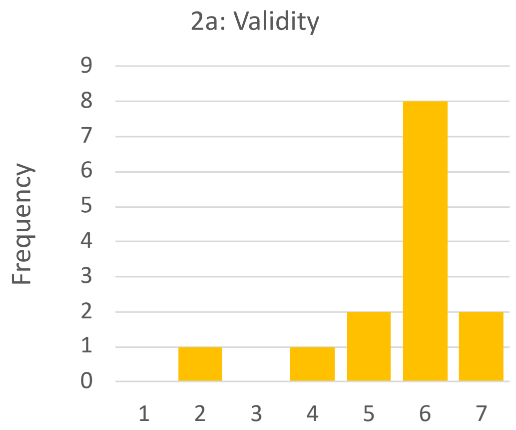
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

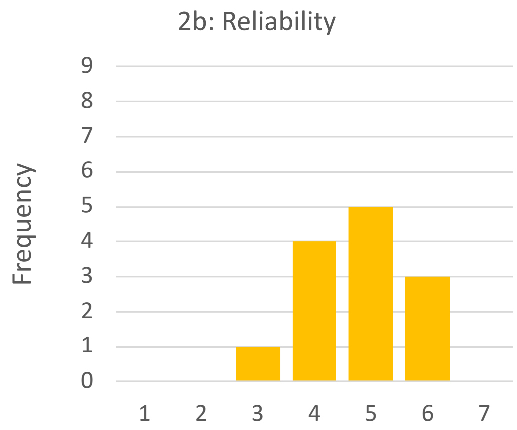
Question 2 Proximity based Surrogate Measures of Safety (SMoS)

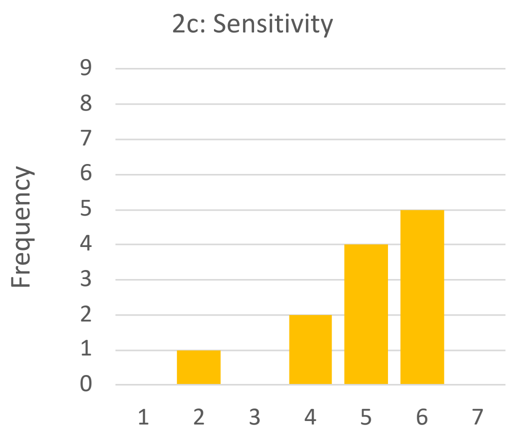
SMoS = pro-active (leading) indicator that use traffic conflicts as surrogate for actual collisions.
TTC = the time remaining until a collision between two vehicles would occur if they maintained their course and speed

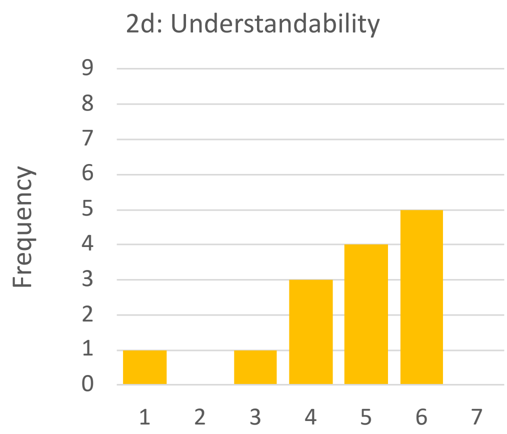
1. The original question 2: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

Type of indicator • Proximity based Surrogate Measures of Safety (SMoS) such as time-to-collision (TTC).	
Example of a specific metric • The number of times TTC exceeds a certain threshold per 100.000 kilometres travelled (per road type)	
	Your score round 1
a. This type of indicator reflects traffic safety well	5
b. This type of indicator could be measured in a reliable way	6
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6









3. Arguments made by experts against this type of indicator:

- TTC sounds nice on paper, but in practice can be cumbersome. TTC is often ill-defined. For example, what objects or road users will be included/excluded in this metric? It may be difficult to apply this homogenously and in such a way that it is robust across different road types etc.
- All these SMoS are context dependent. They rely heavily on assumptions of driving behaviour.
- It relies on vehicle's sensors and those have their limitations (for instance for lane changing manoeuvre).

4. Arguments made by experts in favour of this type of indicator

- SMoS have been shown to be very good precursors of crashes in recent years. There are many ongoing studies to correlate them with crashes.
- Especially with a low threshold (TTC being under the threshold), this measure represents critical situations well.
- The aggregated data could generally show the occurrence of safety risks at the network level. Also, this type of indicator is often measurable/computable from vehicle data and still is applicable when automated driving is promoted in the future.

5. Re-evaluate question 2: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

Type of indicator • Proximity based Surrogate Measures of Safety (SMoS) such as time-to-collision (TTC).		
Example of a specific metric • The number of times TTC exceeds a certain threshold per 100.000 kilometres travelled (per road type)		
	Score R1	**Score R2**
a. This type of indicator reflects traffic safety well	5	*Score*
b. This type of indicator could be measured in a reliable way	6	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6	*Score*
Explanation (optional)		

6. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

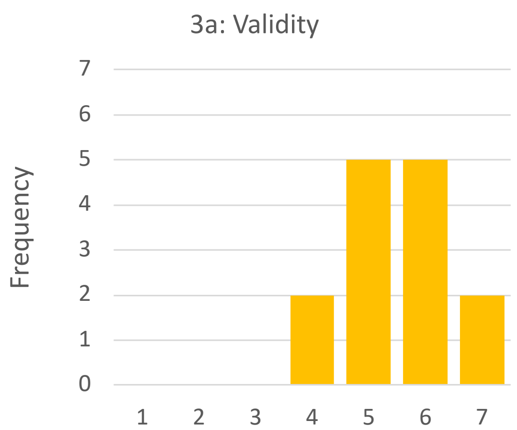
Question 3 Kinematic Surrogate Measures of Safety (SMoS)

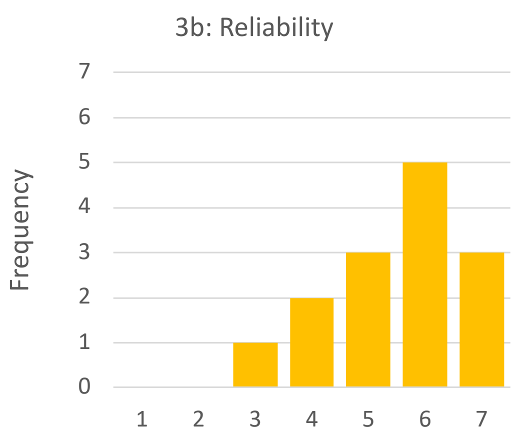
SMoS = pro-active (leading) indicator that use traffic conflicts as surrogate for actual collisions.

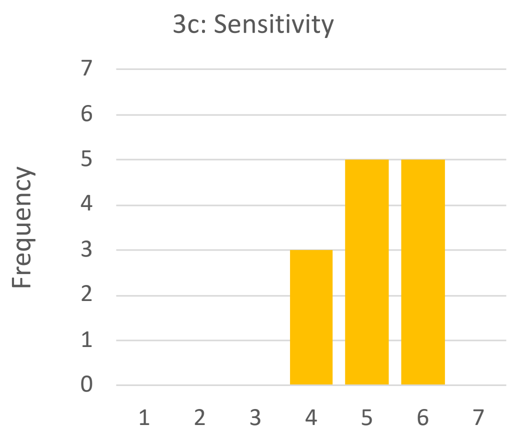
1. The original question 3: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

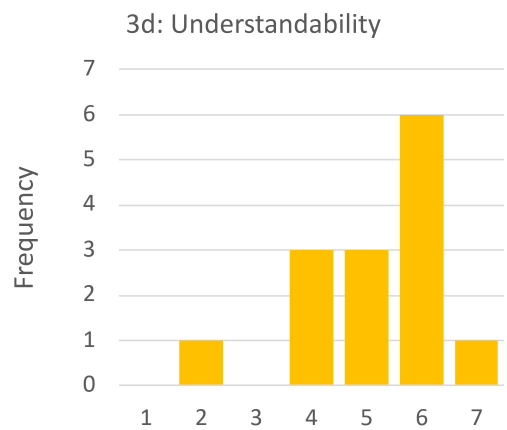
Type of indicator • Kinematic Surrogate Measures of Safety (SMoS)	
Examples of specific metrics • 1. The number of strong decelerations per 100.000 kilometres travelled (per road type) • 2. The number of strong accelerations per 100.000 kilometres travelled (per road type) • 3. The number of strong swerving movements per 100.000 kilometres travelled (per road type)	
	Your score round 1
a. This type of indicator reflects traffic safety well	5
b. This type of indicator could be measured in a reliable way	6
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6

2. Distribution of scores in round 1









3. Arguments made by experts against this type of indicator:

- Due to the nature of these variables, there are many errors and noises in measuring them via accelerometers, and other devices.
- Context matters: It may reflect distraction or a preference of the driver for aggressive driving, but it may also just represent any other energetic driving.
- The measurements could be different for different vehicle types and traffic compositions.

4. Arguments made by experts in favour of this type of indicator

- These indicators provide insight into behaviours that might never show up in crash statistics as they are corrected before critical situations. Yet, they still influence road safety.
- It is not matched to a location or traffic situation, so you don't know why the vehicle had to brake and also not the conditions of the surrounding vehicles. But probably in an aggregate way it is a good indicator for the (statistical) safety at network or road (type) level.
- I would say that accelerations/decelerations (swerving less so, but lane changes yes) is measurable relatively easily (unlike TTC, it does not require sensors that look ahead, but it can be measured from the ego-vehicle behaviour only)

5. Re-evaluate question 3: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

Type of indicator • Kinematic Surrogate Measures of Safety (SMoS)		
Examples of specific metrics • 1. The number of strong decelerations per 100.000 kilometres travelled (per road type) • 2. The number of strong accelerations per 100.000 kilometres travelled (per road type) • 3. The number of strong swerving movements per 100.000 kilometres travelled (per road type)		
	Score R1	**Score R2**
a. This type of indicator reflects traffic safety well	5	*Score*
b. This type of indicator could be measured in a reliable way	6	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6	*Score*
Explanation (optional)		

6. Meaning of the scores

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

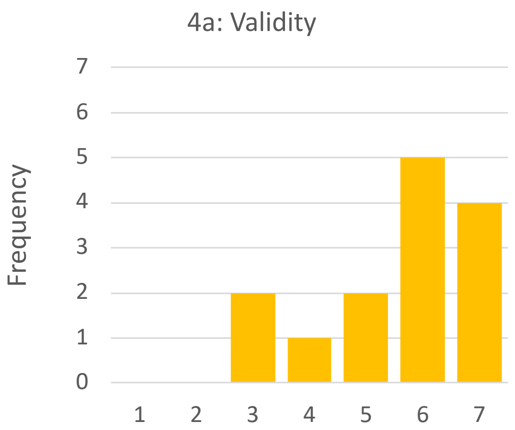
Question 4 Engagement of ADAS

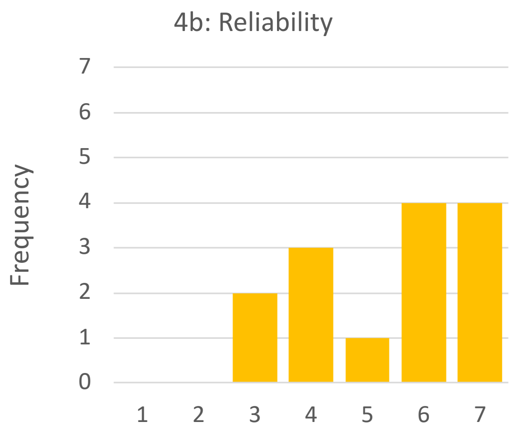
Engagement of ADAS = the moment an ADAS acts, either by warning the driver or by actively intervening

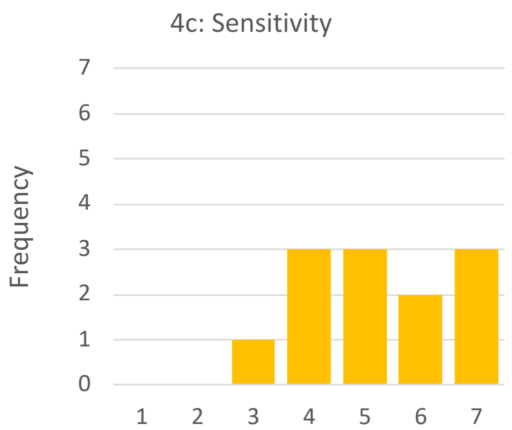
1. The original question 4: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

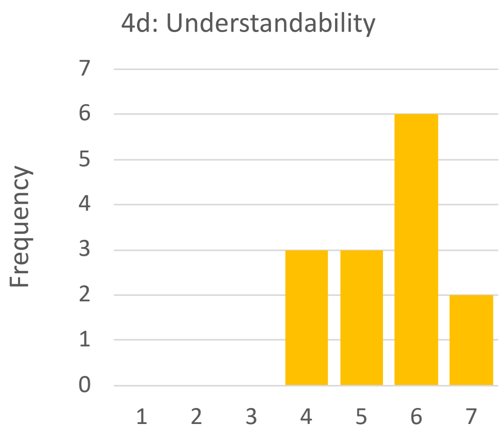
Type of indicator • Engagement of ADAS	
Examples of specific metrics • The number of warnings of Forward Collision Warning (FCW) per 100.000 kilometres travelled (per road type) • The number of interventions of Autonomous Emergency Braking (AEB) per 100.000 kilometres travelled (per road type) • The number of warnings of Blind Sport Warning (BSW) when the driver changes lanes per 100.000 kilometres travelled (per road type)	
	Your score round 1
a. This type of indicator reflects traffic safety well	5
b. This type of indicator could be measured in a reliable way	6
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6

2. Distribution of scores in round 1









3. Arguments made by experts against this type of indicator

- Systems are improving, which makes it a difficult indicator to compare over time as activation is not same.
- Every manufacturer can define an ADAS system as they want to: a warning from FCW will be different from one OEM to another OEM.
- The reliability of such systems may be a limitation (e.g., AEB/FCW false positive)
- These warnings are triggered based on some pre-defined thresholds on indicators of safety like TTC or acceleration. So, using these warnings as a measure of safety is a chicken-egg problem; it is not exogenous but endogenous with safety.

4. Arguments made by experts in favour of this type of indicator

- These indicators reflect more serious conflicts and can be measured reliably by the vehicle.
- Activation of AEBs, discarding false positives, indicate an imminent collision. Thus, if all cars had AEBs one could count the number of frontal car near collisions reliably.
- These should be easy to measure in vehicle. It is important to know what the boundaries of the systems are and when exactly they activate. Differences between systems should be well understood and considered.
- If the thresholds for activation of the ADAS as set by the OEMs can be harmonized, then yes these will be good indicators

5. Re-evaluate question 4: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

Type of indicator • Engagement of ADAS		
Examples of specific metrics • The number of warnings of Forward Collision Warning (FCW) per 100.000 kilometres travelled (per road type) • The number of interventions of Autonomous Emergency Braking (AEB) per 100.000 kilometres travelled (per road type) • The number of warnings of Blind Sport Warning (BSW) when the driver changes lanes per 100.000 kilometres travelled (per road type)		
	Score R1	**Score R2**
a. This type of indicator reflects traffic safety well	5	*Score*
b. This type of indicator could be measured in a reliable way	6	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	5	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	6	*Score*
Explanation (optional)		

6. Meaning of the scores

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

Question 5 Driver distraction like sleepiness

This type of indicator is added in this second round based on suggestions made by the experts.

1. Evaluate question 5: Please rate the statements on a scale from 1 (strongly disagree) to 7 (strongly agree)

Type of indicator • Driver distraction like sleepiness		
Examples of specific metrics • The number of times a driver is at risk of falling asleep per 100.000 kilometres travelled (per road type), as measured by the driver drowsiness and attention warning (DDAW)		
	Score R1	**Score R2**
a. This type of indicator reflects traffic safety well	n/a	*Score*
b. This type of indicator could be measured in a reliable way	n/a	*Score*
c. This type of indicator is sensitive to external changes, i.e., it will change with future traffic safety interventions	n/a	*Score*
d. This type of indicator is understandable for different end-users such as researchers and policy makers	n/a	*Score*
Explanation (optional)		

End of part 1

Information part 2

This part of the questionnaire is aimed at identifying what factors could potentially form a barrier for successfully implementing a system of measuring traffic safety at network level based on vehicle data. This does not specifically refer to any of the in part 1 discussed types of indicators but is meant in a more general way.

This part consists of six questions. In these questions, you will be asked to evaluate a potential barrier to successful implementation of a system of measuring traffic safety at network level based on vehicle data.

Please answer the questions below by scoring the potential barrier from 1 (no barrier at all) to 7 (Insurmountable barrier) and explain why you make this assessment.

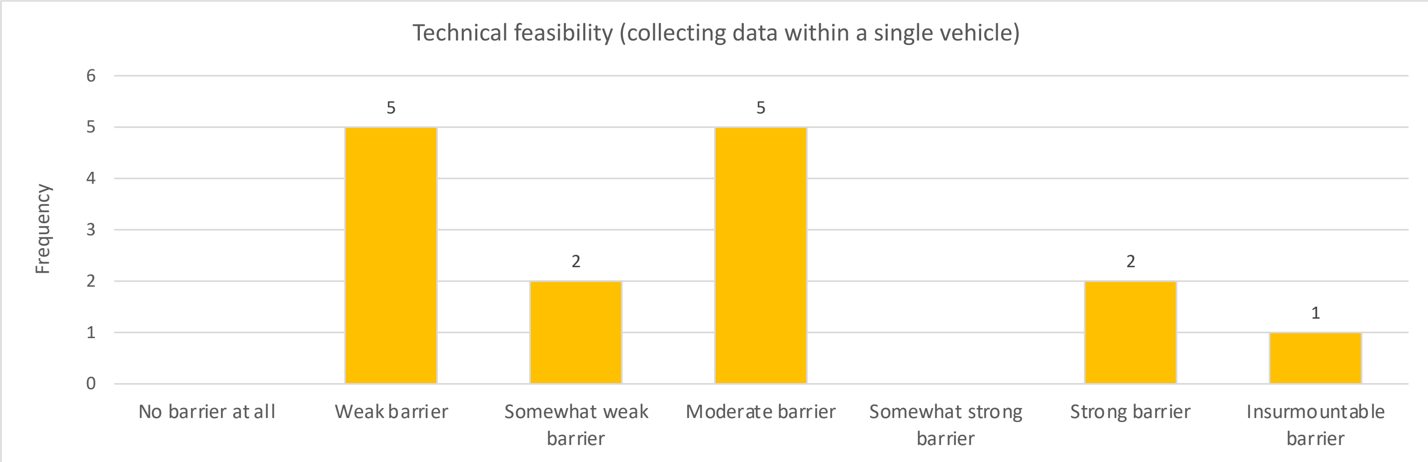
Table 2 below shows the meaning of each score. Please answer a 0 in case you do not have an answer. All parts that require a response from you are marked in blue.

Table 2 Meaning of the scores in part 2

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 1a Technical feasibility (collecting data within a single vehicle)

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- Many signals are available; however, others are not (yet) available or require processing/calculation.
- More complex is sensor data for TTC, not all vehicles are equipped with the necessary sensors.

3. Arguments for a weaker barrier

- Most of the indicators can already be collected technically within a vehicle
- Depending on the indicator of interest and the sensor suit in the vehicle, and assuming the right algorithms exist to acquire, process, and transmit the data, collecting it should not be that difficult.

4. Re-evaluate question 1a: Is the technical feasibility of collecting the data within one vehicle a barrier to successful implementation?

This refers to the technical feasibility of collecting the data necessary for reporting any of the indicators discussed in part 1, within a single vehicle.

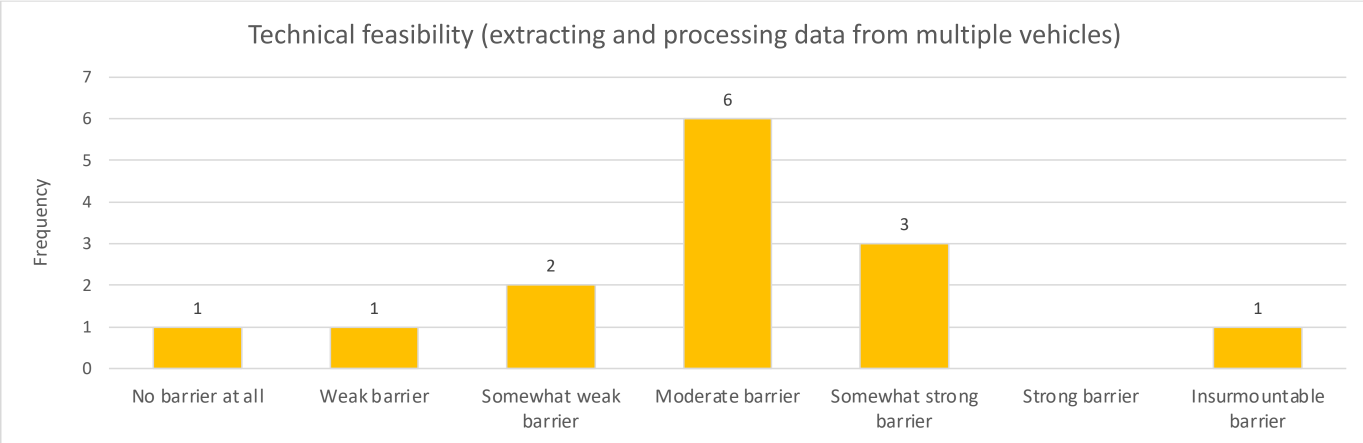
Score round 1	Score round 2
3	score
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 1b Technical feasibility (extracting and processing data from multiple vehicles)

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- The problem is to communicate all that data via wireless communication to a central location periodically. It is not impossible, but it requires effort and cost.
- On fleet level, data should be combined and processed, which is a little more difficult than at vehicle level.
- I see the storing of data as a difficult given the volume.

3. Arguments for a weaker barrier

- With the advances in data management, storing, and processing, there is no barrier in these tasks anywhere in the world.
- Storing and exporting all data requires storage and somewhat regular extraction. Both are easy enough technically. Ensuring the data is comparable between different manufacturers is more difficult but not from a technical standpoint.

4. Re-evaluate question 1b: Is the technical feasibility of extracting and processing the data of a fleet of vehicles a barrier to successful implementation?

This refers to the technical feasibility of the entire process of extracting, processing, and storing the data necessary to report one of the indicators discussed in part 1 for a large number of vehicles.

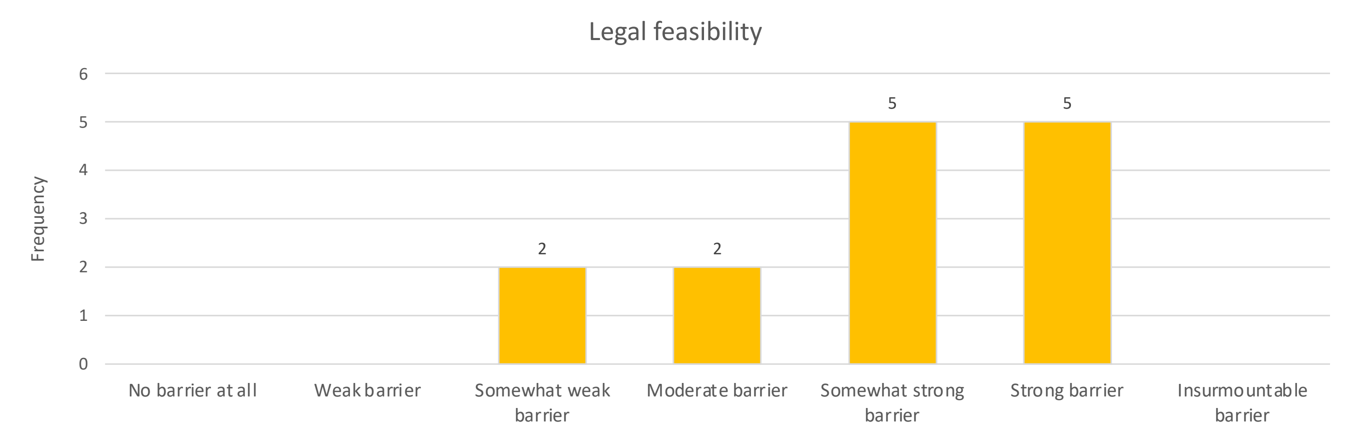
Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 2 Legal feasibility

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- I think there is a legal barrier in terms of privacy if the data are to be collected using unobtrusive technologies inside the vehicles.
- Legal requirements limit the allowance to share information. Also, because who is the 'owner' of the data: user or manufacturer?
- GDPR seems an important issue here and may require strong EU support. A possible solution would be to anonymize all data as soon as possible, which would require OEMs to accept to do the job.

3. Arguments for a weaker barrier

- With the new Data Act there could be some improvement. This new Act will put the driver "in charge" who is able to extract data from his vehicle and who is permitted to do something with it. This will solve a part of the competition question and also solve (partially) the privacy question.
- Laws can be changed to rule in favour of collecting the required data as long as safeguards are in place.

4. Re-evaluate question 2: Is legal feasibility a barrier to successful implementation?

This refers to how (EU) laws or directives on for example privacy or fair market competition could be a potential barrier.

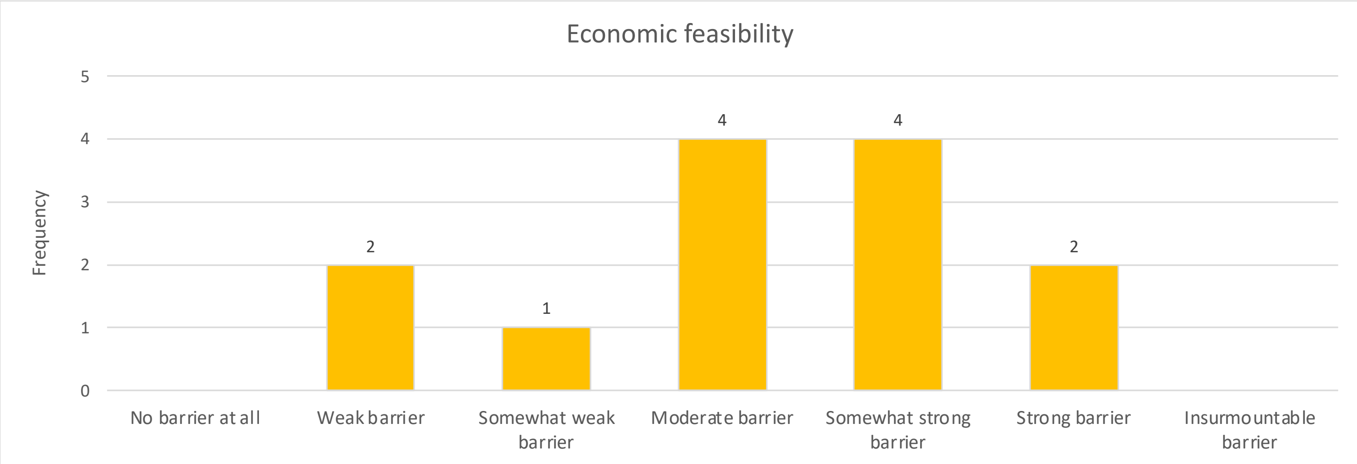
Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 3 Economic feasibility

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- The business case for traffic safety is very difficult, since costs and benefits are divided over different stakeholders.
- There is no obvious business model for OEMs. What would be the gain for OEMs to participate and engage in costs? The question is even more important if it requires adding (and paying for) new systems in the vehicle.

3. Arguments for a weaker barrier

- As long as there is an economic benefit for the companies, they might be willing to participate.
- The cost of a system for extracting, collecting, and processing the data is quite high, but the ecosystem of the ITS/SRTI regulation has shown that there are possibilities.
- If the indicators can be measured with sensors already onboard, then I think OEMs would not have too much of an economic problem. The government certainly benefits by doing this (evidence of improving safety). Service providers like cloud service providers, 5G networks, etc. may also stand to gain with this.

4. Re-evaluate question 3: Is economic feasibility a barrier to successful implementation?

This refers to the question of whether or not a business case exists for relevant stakeholders (OEMs, government, service providers, etc.) that would make them willing to participate or organise such a system? Would it thus be possible to organise such a system with several stakeholders?

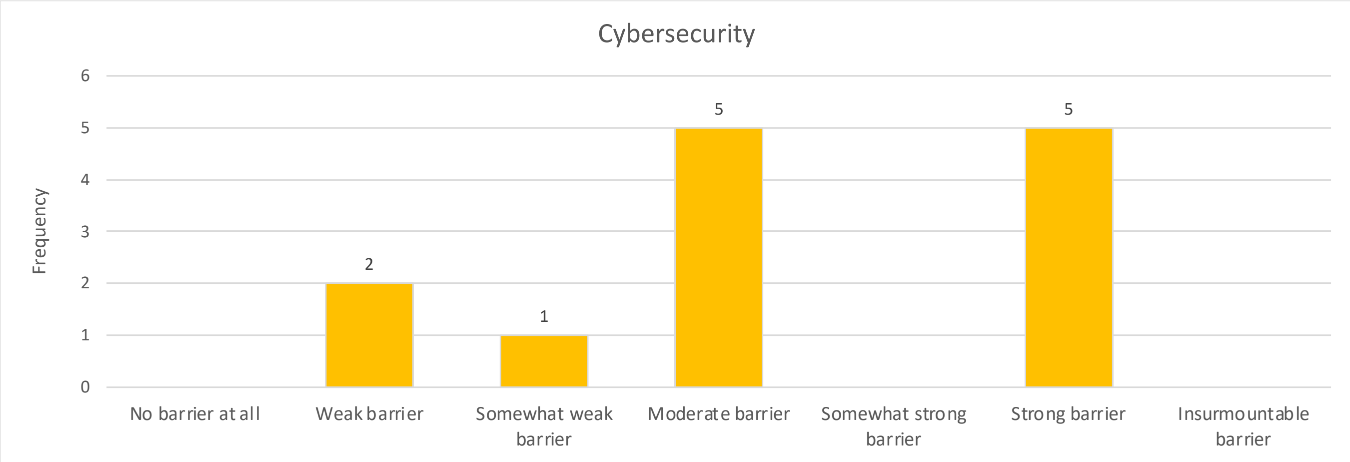
Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 4 Cybersecurity

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- Cybersecurity attacks will be possible. It strengthens even more the importance of data privacy and the solidity of the GDPR measures
- Cyber security is a hot topic. This is mainly of concern for the OEM but also for aftermarket suppliers of devices, especially when these devices are somehow connected to the CAN-network in the vehicle.
- During the processing of data, it is important to decide who has access to the data, how it is stored and when data are deleted.

3. Arguments for a weaker barrier

- Secure communication from vehicle should be possible. Especially as communication is not time critical and no vehicle decisions are based on it.
- If over the air communication is used this becomes a bigger problem. If this is not the case, cybersecurity is less of an issue.
- I don't think it should be very hard to ensure a sufficient level of cybersecurity, although it may not always get the attention it deserves.
- There is a potential safety risk if in-vehicle or traffic systems could be accessed from outside, but the safety indicators in itself do not contain a security risk.

4. Re-evaluate question 4: Is cybersecurity a barrier to successful implementation?

This refers to the cybersecurity during the whole process: within a vehicle, during the over-the-air communication, and during the processing of the data.

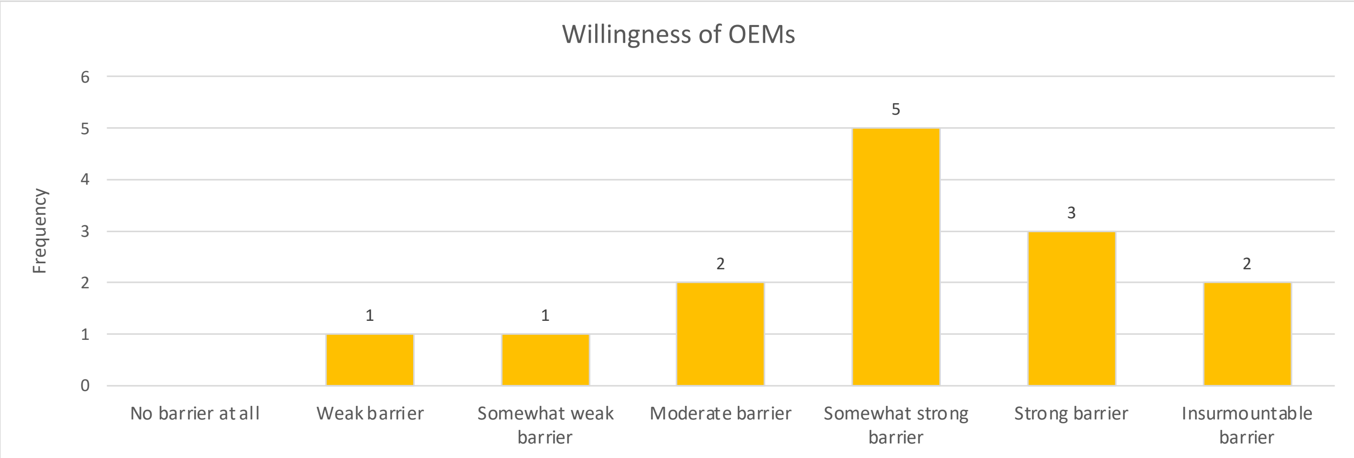
Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 5a Willingness of OEMs

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- OEMs tend to keep the data for themselves.
- It depends on the request for information of the OEM. If it contains confidential information that is a threat to their competitive position, they will be reluctant to share it. It is also an additional effort (money), so it should not weaken their position.
- The data is owned by the consumer (as defined in the Data Act), so the OEM can only share this data with consent by the consumer, or on a legal basis.

3. Arguments for a weaker barrier

- They may not want to share their data, but this is a point of economic benefit and regulation.
- Market parties act where the money is.

4. Re-evaluate question 5a: Is Willingness of **OEMs/vehicle manufacturers** a barrier to successful implementation?

Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 5b Willingness of suppliers

1. Distribution of scores in round 1

2. Arguments for a stronger barrier

- Similar to OEMs, implementing this needs consensus on how they could protect their IP and information

3. Arguments for a weaker barrier

- If there is an economic incentive for them, it is no barrier
- I don't see this. They only supply.

4. Re-evaluate question 5b: Is Willingness of **suppliers** a barrier to successful implementation?

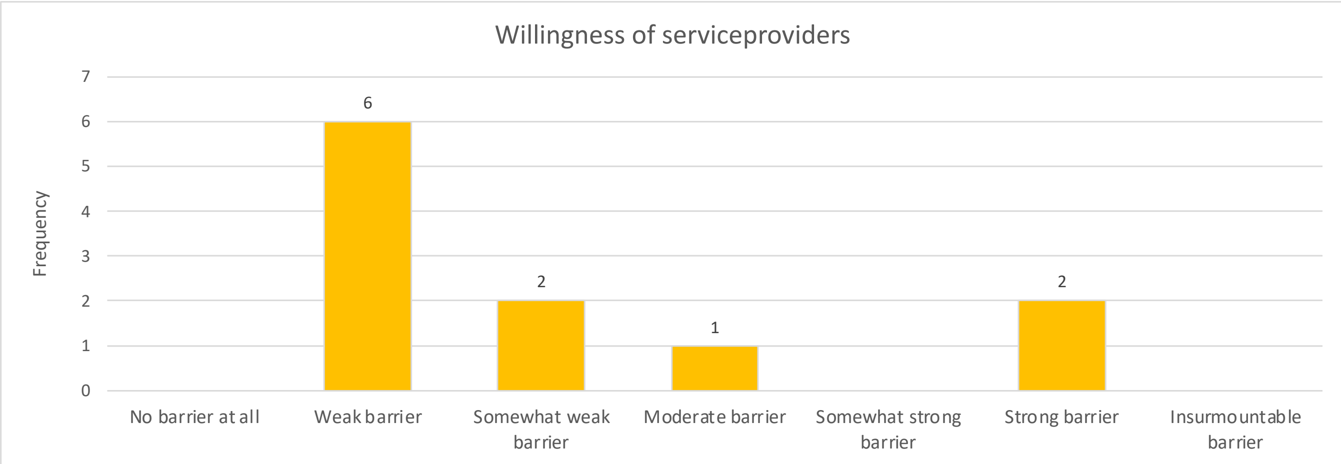
Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 5c Willingness of service providers

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- It is a barrier as long as there is no feasible business case for them to invest in a (part of a) network

3. Arguments for a weaker barrier

- I do not see a big issue here. They will see this as new business
- I think service providers might be more open to the idea.

4. Re-evaluate question 5c: Is Willingness of **Service Providers** a barrier to successful implementation?

Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 5d Willingness of people

This potential barrier is included based on suggestions made by experts and does therefore not have any information about prior answers.

1. Evaluate question 5d: Is Willingness of **people** a barrier to successful implementation?

People refers to the owners of the vehicle that have to give permission to share their data.

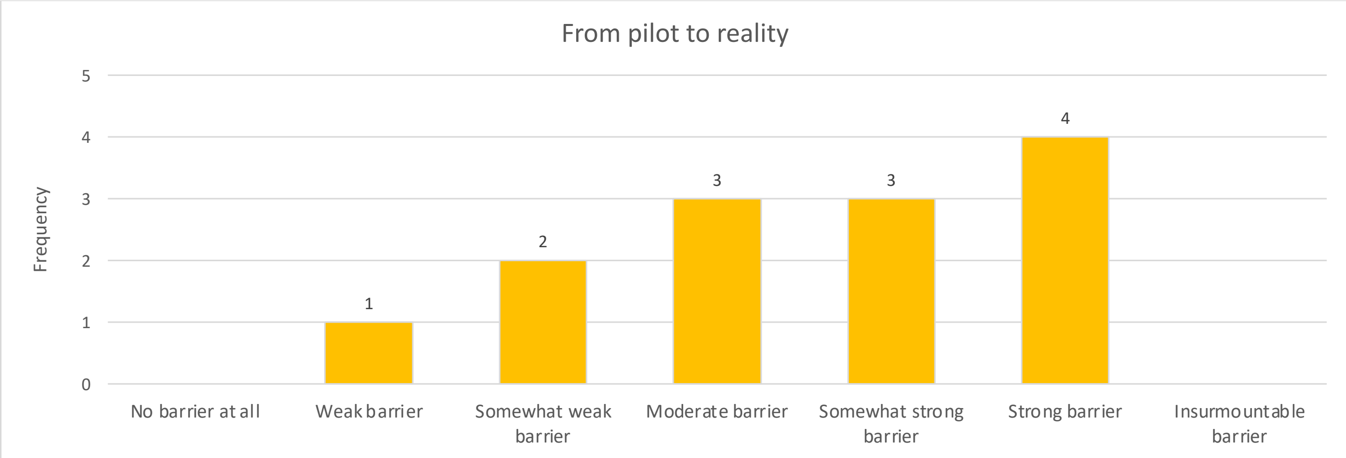
Score round 1	Score round 2
n/a	*score*
Explanation	

2. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier

Question 6 Going from pilot to reality

1. Distribution of scores in round 1



2. Arguments for a stronger barrier

- The logistics of setting this up will be huge and political. We can compare this to the Dutch Rekeningrijden (road pricing), a nice idea but received political resistance.
- It is always difficult to go from innovation to operation.
- I think there is a barrier in adapting these technologies by the stakeholders mostly because there is still lack of strong and sufficient evidence showing that these technologies are useful, and that they will work.
- Experience from the ecosystem around the ITS/SRTI directive has shown that it takes a long time to realise such an implementation, in particular if this includes harmonisation of the data/interface etc.

3. Arguments for a weaker barrier

- I think this is a question of feasibility rather than going from pilot to reality.
- I guess this would depend on the particular stakeholders, their experience with this process, are people involved with the correct skill sets?
- Starting from small scale pilots to larger scale pilots can help with the jump from small scale pilots to actual implementation.

4. Re-evaluate question 6: Is Going from pilot to reality a barrier to successful implementation?
This refers to the organisational difficulty of stakeholders to go from small scale pilots to the full incorporation of such a system in their organisations.

Score round 1	Score round 2
3	*score*
Explanation (optional)	

5. Meaning of the scores

0	1	2	3	4	5	6	7
I do not know	No barrier at all	Weak barrier	Somewhat weak barrier	Moderate barrier	Somewhat strong barrier	Strong barrier	Insurmountable barrier