Assessing Human Drivers
From Raw Data to Context-Aware Interpretations

Driessen, T.

DOI
10.4121/uuid:be5e6366-d881-4ab4-8522-42416efab787

Publication date
2025

Document Version
Final published version

Citation (APA)
Driessen, T. (2025). *Assessing Human Drivers: From Raw Data to Context-Aware Interpretations*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4121/uuid:be5e6366-d881-4ab4-8522-42416efab787

Assessing Human Drivers

From Raw Data to Context-Aware Interpretations

Tom Driessen

Assessing Human Drivers

From Raw Data to Context-Aware Interpretations

Tom Driessen

Assessing Human Drivers: From Raw Data to Context-Aware Interpretations

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof. dr. ir. T. H. J. J. van der Hagen,
Chair of the Board for Doctorates
to be defended publicly on
Wednesday, 10 September 2025 at 15:00 o'clock

by

Tom DRIESSEN

Master of Science in Mechanical Engineering,
Delft University of Technology, the Netherlands
born in Leiden, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof. dr. ir. J. C. F. de Winter	Delft University of Technology, promotor
Dr. D. Dodou	Delft University of Technology, promotor
Prof. dr. D. de Waard	University of Groningen, promotor

Independent members:

Prof. dr. M. A. Neerinx	Delft University of Technology
Prof. dr. ir. M. Mulder	Delft University of Technology
Dr. ir. H. Farah	Delft University of Technology
Prof. dr. P. A. Hancock	University of Central Florida
Prof. dr. ir. M. Wisse	Delft University of Technology, reserve member

This research is funded by Transitions and Behaviour grant 403.19.243 ("Towards Safe Mobility for All: A Data-Driven Approach"), provided by the Dutch Research Council (NWO).

Printed by: Ridderprint
Cover design: Tom Driessen
ISBN: 978-94-6384-811-4

An electronic version of this thesis is available at https://repository.tudelft.nl/.

Contents

Summary

Road traffic accidents are a large public health concern, causing 1.19 million deaths annually and ranking as the leading cause of death among young people aged 5–29. While substantial progress in reducing road fatalities was made over the past 50 years, in many countries such as the Netherlands this progress is stagnating.

Recent technological advances, particularly in vehicle automation and artificial intelligence, have transformed modern vehicles into machines equipped with advanced sensing and computing capabilities. While full automation remains a distant goal, these technological developments present new opportunities for improving human driving performance. Additionally, the widespread adoption of smartphones, with their built-in sensors and computing power, offers potential for data collection and analysis even in vehicles without advanced safety features.

This dissertation aims to develop and evaluate algorithms capable of detecting patterns in human driving behavior, focusing specifically on car and truck drivers. The research investigates both practical, accessible methods using existing technology and more advanced approaches using artificial intelligence and automated driving systems. The dissertation is structured in three parts:

Part 1: Perspectives on data use and technology in driver testing and the trucking industry
The first part of the dissertation provides groundwork by examining the perspectives of stakeholders in driving assessment and professional driving. Through interviews with driving examiners and a survey of truck drivers, *Chapter 2* and *Chapter 3* explore the current state of assessment of prospective drivers and the role of technology in professional driving, respectively.

It is found that both driving examiners and professional truck drivers express openness to data-driven tools, but with a condition: these tools must support rather than supplant professional autonomy and judgment. Driving examiners welcome data that can help them explain their decisions and standardize assessments, but stress that human judgment must remain central to the evaluation process. Similarly, truck drivers appreciate technological aids that genuinely support their work, such as adaptive cruise control, but resist systems that are overly constrictive, make errors, or create a sense of excessive monitoring.

Part 2: Measuring driving behavior: practical approaches
The second part focuses on developing practical, accessible methods for assessing driving behavior using commonly available sensors and technology. This part investigates how

relatively simple data collection methods can provide insights into driving behavior and performance.

Chapter 4 presents an algorithm to detect lane changes from mobile GPS data. The performance is accurate for local analyses of group-level traffic behavior but has limitations in real-time detection systems due to a relatively high rate of false alarms. *Chapter 5* presents methods to distinguish between cautious, normal, and aggressive driving styles as portrayed by experienced driving examiners, based on accelerometer and GPS data. *Chapter 6* examines driving data from truck drivers to predict damage incidents, traffic fines, and fuel consumption, finding that the number of harsh braking events per hour is predictive of the number of fines and damage incidents, while engine torque exceedances are predictive of higher fuel consumption.

An overarching conclusion of the three chapters in *Part 2* is that while readily available sensors can provide valuable population-level insights into driving behavior, their application to individual assessment requires careful consideration of context. The research demonstrates that basic sensors like GPS and accelerometers can effectively detect patterns in driving behavior and identify risk factors across large populations. However, these same metrics can be misleading when applied to individual cases without contextual information. For instance, frequent harsh braking events might indicate risky driving behavior in some contexts but be entirely appropriate in others, such as urban environments or challenging traffic conditions. This conclusion introduces the need for assessment systems that can incorporate contextual factors when evaluating driving behavior of individuals.

Part 3: Measuring driving behavior: towards context-aware methods
The final part explores more sophisticated approaches to driving assessment, using advanced artificial intelligence and automated driving systems. This part represents a step towards developing context-aware methods of evaluating driving behavior. The research demonstrates new applications of AI technology, including the use of automated driving systems to analyze human driving decisions (*Chapter 7*) and the application of large language models with vision capabilities to assess risk in traffic situations (*Chapter 8*).

The conclusion of *Part 3* is that modern AI-driven techniques show promise regarding more context-aware evaluation of driving behavior. The research demonstrates that AI systems developed for automated driving can help distinguish between justified and unjustified driving actions by considering the full context of traffic situations. Besides that, we show that visual large language models can be used to assess risk in images of traffic situations.

Conclusion
Finally, reflecting on the dissertation's three central ideas: First, data-driven tools should reinforce, rather than displace, professional autonomy, whether in driver testing or professional trucking. Second, accelerometer, GPS, and telematics data alone, while powerful at scale, require contextual enrichment for individual-level insights. Third, advanced

AI, from modern automated driving algorithms to vision-language models, can supply missing context and thereby improve the validity of safety assessments.

Thereby, through developing and evaluating methods to understand driving behavior, this dissertation has contributed towards the main aim of advancing assessment approaches and improving road safety.

Samenvatting

Verkeersongevallen vormen een groot probleem voor de volksgezondheid. Ze zorgen jaarlijks voor 1,19 miljoen sterfgevallen en zijn de belangrijkste doodsoorzaak onder jongeren van 5 tot 29 jaar. Hoewel er de afgelopen 50 jaar aanzienlijke vooruitgang is geboekt in het terugdringen van het aantal verkeersdoden, stagneert die vooruitgang in veel landen, waaronder Nederland.

Recente technologische ontwikkelingen, vooral op het gebied van voertuigautomatisering en kunstmatige intelligentie, hebben moderne voertuigen veranderd in machines met geavanceerde sensoren en rekenkracht. Hoewel volledige automatisering nog niet in zicht is, bieden deze ontwikkelingen nieuwe mogelijkheden om de rijprestaties van mensen te verbeteren. Daarnaast biedt het wijdverbreide gebruik van smartphones, met hun ingebouwde sensoren en rekenvermogen, kansen om gegevens te verzamelen en te analyseren, zelfs in voertuigen zonder geavanceerde veiligheidssystemen.

In dit proefschrift worden algoritmen ontwikkeld en geëvalueerd die patronen in menselijk rijgedrag kunnen detecteren, met een focus op automobilisten en vrachtwagenchauffeurs. Hiervoor is onderzoek gedaan naar zowel praktische en laagdrempelige toepassingen met bestaande technologie als naar meer geavanceerde methoden met kunstmatige intelligentie en geautomatiseerde rijsystemen. Het proefschrift is opgebouwd uit drie delen:

Deel 1: Perspectieven op datagebruik en technologie in het rijexamen en de transportsector
Het eerste deel van dit proefschrift legt de basis door de perspectieven van verschillende belanghebbenden in het beoordelen van bestuurders en in beroepsmatig rijden te inventariseren. Aan de hand van interviews met rijexaminatoren en een enquête onder vrachtwagenchauffeurs wordt in *Hoofdstuk 2* en *Hoofdstuk 3* onderzocht hoe de beoordeling van beginnende bestuurders nu wordt uitgevoerd en hoe de technologie een rol speelt in het beroepsmatig rijden.

Uit deze hoofdstukken blijkt dat zowel rijexaminatoren als beroepschauffeurs openstaan voor datagedreven hulpmiddelen, maar alleen als deze het professionele oordeel en de autonomie ondersteunen in plaats van vervangen. Rijexaminatoren verwelkomen data die hun beslissingen kan onderbouwen en beoordelingen kan standaardiseren, maar benadrukken dat menselijk inzicht leidend moet blijven. Beroepschauffeurs waarderen technologische hulpmiddelen zoals adaptieve cruisecontrol, mits die daadwerkelijk helpen bij het werk en niet leiden tot te veel beperkingen, fouten of een gevoel van overmatige controle.

Deel 2: Het meten van rijgedrag: praktische benaderingen
Het tweede deel richt zich op het ontwikkelen van praktische, toegankelijke methoden om rijgedrag te meten met behulp van algemeen beschikbare sensoren en technologie. Hierbij is onderzocht hoe relatief eenvoudige vormen van dataverzameling gebruikt kunnen worden om meer inzicht te krijgen in rijgedrag en -prestaties.

In *Hoofdstuk 4* wordt een algoritme beschreven dat rijstrookwisselingen detecteert op basis van mobiele GPS-gegevens. De nauwkeurigheid is goed voor lokale analyses van verkeersgedrag, maar voor realtime-detectie is het aantal foutieve meldingen te hoog. *Hoofdstuk 5* presenteert methoden onderscheid te maken tussen verschillende rijstijlen op basis van gedrag geacteerd door ervaren rijexaminatoren. Hierbij maken we gebruik van accelerometer- (versnellingsmeter) en GPS-data. In *Hoofdstuk 6* wordt gekeken naar gegevens van vrachtwagenchauffeurs om schade-incidenten, verkeersboetes en brandstofverbruik te voorspellen. Hieruit blijkt dat het aantal harde remacties per uur verband houdt met het aantal boetes en schadegevallen, terwijl overschrijdingen van het motorkoppel samengaan met een hoger brandstofverbruik.

Een belangrijke conclusie uit de drie hoofdstukken in *Deel 2* is dat algemeen beschikbare sensoren waardevolle inzichten kunnen geven in rijgedrag op groepsniveau, maar dat toepassing bij individuele bestuurders om zorgvuldige overweging van de context vraagt. Hoewel eenvoudige sensoren als GPS en accelerometers risicofactoren en patronen in groot en gevarieerd rijgedrag kunnen herkennen, kunnen dezelfde maatstaven misleidend zijn zonder de juiste context. Frequente harde remacties kunnen bijvoorbeeld in sommige situaties–zoals in druk stadsverkeer–volledig gerechtvaardigd zijn, terwijl ze in andere situaties kunnen wijzen op gevaarlijk rijden. Dit illustreert de noodzaak van beoordelingssystemen die rekening houden met contextuele factoren bij een individuele beoordeling.

Deel 3: Het meten van rijgedrag: naar contextbewuste methoden
Het laatste deel beschrijft geavanceerdere benaderingen voor de beoordeling van rijgedrag, waarbij kunstmatige intelligentie en geautomatiseerde rijsystemen worden ingezet. Deze benaderingen vormen een stap richting contextbewuste methoden voor het evalueren van rijgedrag. Zo wordt in *Hoofdstuk 7* ingegaan op het gebruik van geautomatiseerde rijsystemen voor het analyseren van menselijke rijbeslissingen. *Hoofdstuk 8* beschrijft hoe grote taalmodellen met visuele mogelijkheden gebruikt kunnen worden om risico's in verkeerssituaties te beoordelen. De conclusie van *Deel 3* is dat moderne, door AI ondersteunde technieken veel potentie hebben voor contextbewuste evaluatie van rijgedrag. Het onderzoek laat zien dat AI-systemen, ontwikkeld voor geautomatiseerd rijden, kunnen helpen onderscheid te maken tussen gerechtvaardigde en ongerechtvaardigde rijacties door de volledige verkeerscontext in de beoordeling mee te nemen. Ook blijkt dat visuele taalmodellen kunnen worden ingezet om risico's in afbeeldingen van verkeerssituaties te identificeren.

Conclusie
Ten slotte, reflecterend op de drie centrale ideeën van dit proefschrift: Ten eerste moe-

ten data-gedreven tools de professionele autonomie versterken in plaats van verdringen, of het nu gaat om rijexaminering of de vrachtwagenindustrie. Ten tweede hebben accelerometer-, GPS- en telematicagegevens, hoewel krachtig op schaal, contextuele verrijking nodig voor inzichten op individueel niveau. Ten derde kunnen geavanceerde AI-technieken, van moderne geautomatiseerde rijalgoritmen tot vision-taalmodellen, de ontbrekende context leveren en daarmee de validiteit van veiligheidsbeoordelingen verbeteren.

Door het ontwikkelen en evalueren van methoden om rijgedrag te begrijpen, heeft dit proefschrift bijgedragen aan het hoofddoel om beoordelingsmethoden te verbeteren en de verkeersveiligheid te vergroten.

1

Introduction

1.1. The health burden of road accidents and technological promises

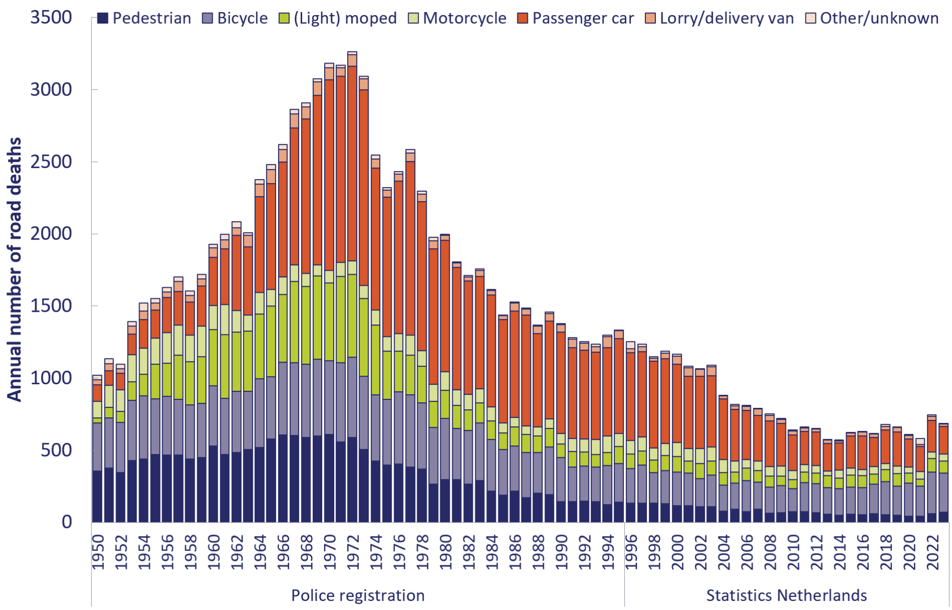


Figure 1.1: Annual number of road deaths in the Netherlands. From SWOV (2024).

Road traffic accidents are an important issue in public health. Currently, they result in 1.19 million deaths annually and leave many more disabled (World Health Organization, 2023). Among children and young adults aged 5-29 years, they are now the leading cause of death (World Health Organization, 2023).

Over the past 50 years, progress has been made in reducing road fatalities. In the Netherlands, road deaths were on the rise until the 1970s but have steadily declined until about 2010 (Figure 1.1). This improvement can be attributed to the implementation of road safety measures often categorized under the "three E's": Education, Engineering, and Enforcement (Learoyd, 1950; McKenna, 2012). These measures include public awareness campaigns, improvements in vehicle safety and infrastructure improvements, and stricter enforcement of traffic laws, respectively.

However, the downward trend in traffic fatalities seems to have stalled (SWOV, 2024), indicating that existing measures may have reached their limit in effectiveness. This stagnation threatens ambitious goals such as Vision Zero, the long-term vision endorsed by the European Union to achieve a fatality-free transport system (European Commission, 2019), or the Sustainable Development Goals formulated by the United Nations (2015), which aim to halve global road traffic fatalities and injuries by 2030, relative to 2015 levels.

Engineering efforts to improve road safety have intensified in the past decade or two towards the development of automated driving systems. The progress in automated driving can, in large part, be attributed to advances in the field of machine learning and vision. This technology has come very far: many new vehicles are capable of driving without human input, especially on long sections of the highway. In some areas, automated taxis can now be booked without a physical driver in the vehicle (e.g., USA: Waymo, 2024; China: The Economist, 2024).

Yet, all current forms of driving automation still require human supervision. For example, Tesla's "Autopilot" and "Full self-driving" features, despite their names, come with warnings that they "require active driver supervision and do not make the vehicle autonomous" (Tesla, 2024a), and warn that drivers should keep their hands at the steering wheel at all times and that they maintain responsibility for the control of the vehicle. Automated taxis rely on human supervision as well, albeit through remote operators, and there are ongoing concerns about the vehicles safety, public image and legislative issues. For instance, in October 2023, Cruise withdrew their automated taxi fleet from San Francisco after losing its permit following an accident involving a pedestrian, to "rebuild public trust" (Cruise, 2023; see also Cruise, 2024). And Tesla now increasingly faces legal resistance in cases where it claims drivers are accountable for crashes due to ignoring their supervision instructions (Thadani, 2024). Similar issues are playing out in China (e.g., Yang, 2024), a country with an emerging automated driving industry.

Whether these are minor flaws or large barriers to the general introduction of fully automated vehicles in our lives is beyond the scope of this dissertation, but what is clear is that the promise that full self-driving automation is nearby may be overly optimistic; it remains to be seen when these vehicles will become commonplace.

A consequence of the recent advances in automation is that modern vehicles have transformed into cognitive robots on wheels, packed with sensing and computing abilities. New vehicles may now contain multiple cameras and high computational resources. In this dissertation, I propose that these resources should not just aim to make cars better at driving themselves, but should also be leveraged to improve human driving abilities. Of additional use for this goal are the capabilities of mobile phones that most drivers now bring on board: basic models have at least GPS, accelerometers, and simple cameras, and newer models are equipped with more advanced sensors such as LiDAR (see Teppati Losè et al., 2022 for a technical evaluation of iPhone's LiDAR and Wang, 2022 for an example of speed estimation using the LiDAR of an iPhone mounted on a bicycle) and chips optimized for machine learning and vision applications (such as Google's Tensor chip, see Amadeo, 2021). The use of mobile phones for recording driving behavior also opens up possibilities for those who do not possess vehicles with the latest safety features. This is important as 92% of all fatal accidents occur in low- and middle-income countries, and within high-income countries, fatalities are also skewed towards lower incomes (World Health Organization, 2023).

1.2. Assessing individual drivers

The notion of recording data to assess one's ability to drive dates back to the early twentieth century. The industrializing economy and more motorized traffic meant the burden

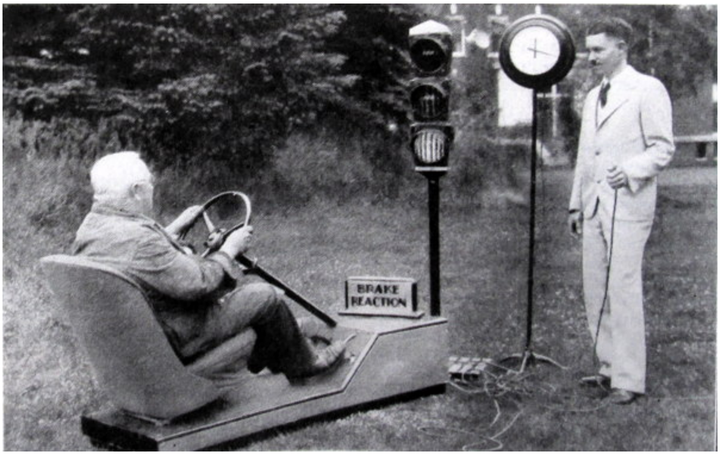


Figure 1.2: Brake reaction time apparatus used by De Silva (1936).

of accidents increased, laying the foundation for psychological research into individual differences in accident proneness. Setups to record individual differences typically involved basic tests such as brake reaction time tests (Figure 1.2; De Silva, 1936), or Münsterberg's (1913) setup involving cards with simulated streets and tracks populated with dots representing different moving objects (pedestrians, horses, automobiles) at varying speeds. However, standard psychometric tests have limited validity for driving competence. It is known, for example, that young people do well on psychometric tests, such as tests involving reaction time, but young drivers are also overrepresented in accident statistics (SWOV, 2024).

The modern example of individual assessment is the driving test, which was introduced around the same time period (Figure 1.3), and is now mandatory in most places before being allowed to drive a vehicle. However, research on driving tests is scarce and shows no conclusive evidence about their validity. Moreover, there are issues with standardization and consistency across different regions and between test centers (Baughan et al., 2005). Additionally—and this is more of a criticism of driving education approaches in general—instruction and testing stops abruptly; no feedback is received by a typical driver anymore after obtaining the license except for interventions by law enforcement or by unsolicited advice from well-meaning passengers. Only truck drivers are required to undergo periodic training (in the EU named 'code 95'; Directive 2003/59/EC, 2003), though these courses are typically generalistic and offer no personalized advice.

The advent of sensors in vehicles can possibly overcome these problems, as they open up possibilities for prolonged tutoring, feedback, or rewards based on displayed driving behavior. In simple forms, this idea already made its way to the market: many vehicle insurance companies now offer personalized discounts based on driving performance measures such as harsh braking or cornering (e.g., Admiral, 2024; Allianz, 2024; Allstate, 2024; ANWB, 2024) and examples of data-driven driver coaching for truck drivers ex-

Figure 1.3: Still from a 1935 instruction video by Ford on the new compulsory UK driving test (Pathe Films & Ford Motor Co Ltd, 1935). In the scene, the narrator explains signals are to be given by raising an arm out of the window (turn signal lights were popularized throughout the 1940s).

ist (e.g., Geotab, 2024; NEXTdriver, 2024). However, the scientific grounding of these measures and research about the consequences of such reward systems are still in their infancy.

1.3. Aim of this dissertation

In summary, the recent developments in engineering (automation, smartphones) create new opportunities for improving human driving (education). The main aim of this dissertation is therefore to contribute to the development of algorithms that can detect patterns in human driving behavior (maneuvers, style, risk).

1.4. Scope of this dissertation

The focus of the dissertation will be on car drivers and truck drivers. Car drivers are of interest as they cause the most fatal injuries (European Road Safety Observatory, 2021; NHTSA, 2021). Truck drivers are of interest as they drive professionally, have been exposed to modern ADAS systems, and commonly drive with fleet management systems that aggregate data on driving performance.

1.5. Outline of this dissertation

This dissertation consists of three parts, with the overarching goal to contribute to the development of algorithms that can be used for the assessment of human driving. The first part consists of qualitative analyses with the aim to inform us about the practice of human driving examination (through interviews with driving examiners) and the practice of truck driving (through a survey among truck drivers). The second part aims to develop methods to assess human driving using accessible data recording techniques. The final part explores more advanced and upcoming methods that are first steps towards obtaining context-aware assessments of human driving performance. Figure 1.4 gives a graphical overview.

Part I: Perspectives on Data Use and Technology in Driver Testing and the Trucking Industry

In Chapter 2, Dutch driving examiners were interviewed about their profession, and specifically their views on data-supported examination of driving test candidates. They provided advice and opinions on what data should be collected, how it should be presented, in what ways it can support examiners in their work and what they believe the limitations are. This chapter introduces the reader to the current practice of driver assessment by human examiners and important challenges towards assessing human drivers with the support of data.

Chapter 3 contains a survey study among Dutch truck drivers and is more broadly focused on the experiences truck drivers have in general. A part of the survey contains specific questions about the technology that truck drivers encounter in their daily work such as ADAS and fleet management solutions.

Part II: Measuring Driving Behavior: Practical Approaches

The second part aims to develop and test various methods to assess human driving. In Part II, the focus is on data recording methods that are already available in most cars or trucks today. The sensors and computations in these chapters can be executed locally on mobile devices or on-board computers. The advantage is that this makes these methods widely applicable and accessible. The three chapters cover the measurement of maneuvers (specifically lane changes), measuring driving style (e.g., aggressive/calm driving), and predicting incidents and fuel consumption from truck telematics data.

Chapter 4 begins with a study on detecting lane changes on highways. We look specifically at GPS data recorded by external devices such as smartphones.

Chapter 5 focuses on acceleration data. In the chapter, experienced driving examiners were asked to demonstrate how exam candidates typically perform driving exams. The trips were recorded and the acceleration data was studied, to see to what extent one can draw conclusions about the driving style from measuring accelerations alone.

Chapter 6 analyzes telematics data from a fleet of truck drivers logged over a period of two years, such as harsh brake counts, engine idle time and cruise control usage statistics, and investigates their relation to damage occurrence, fines received, and fuel consumption.

Part III: Measuring Driving Behavior: Towards Context-Aware Methods

Part III consists of two chapters where we use emerging technology to analyze driving data. *Chapter 7* demonstrates how existing automated driving systems can be used in the analysis of human driving data. An adapted version of Openpilot, an open-source automated driving system, is given prerecorded video and vehicle data of human driving.

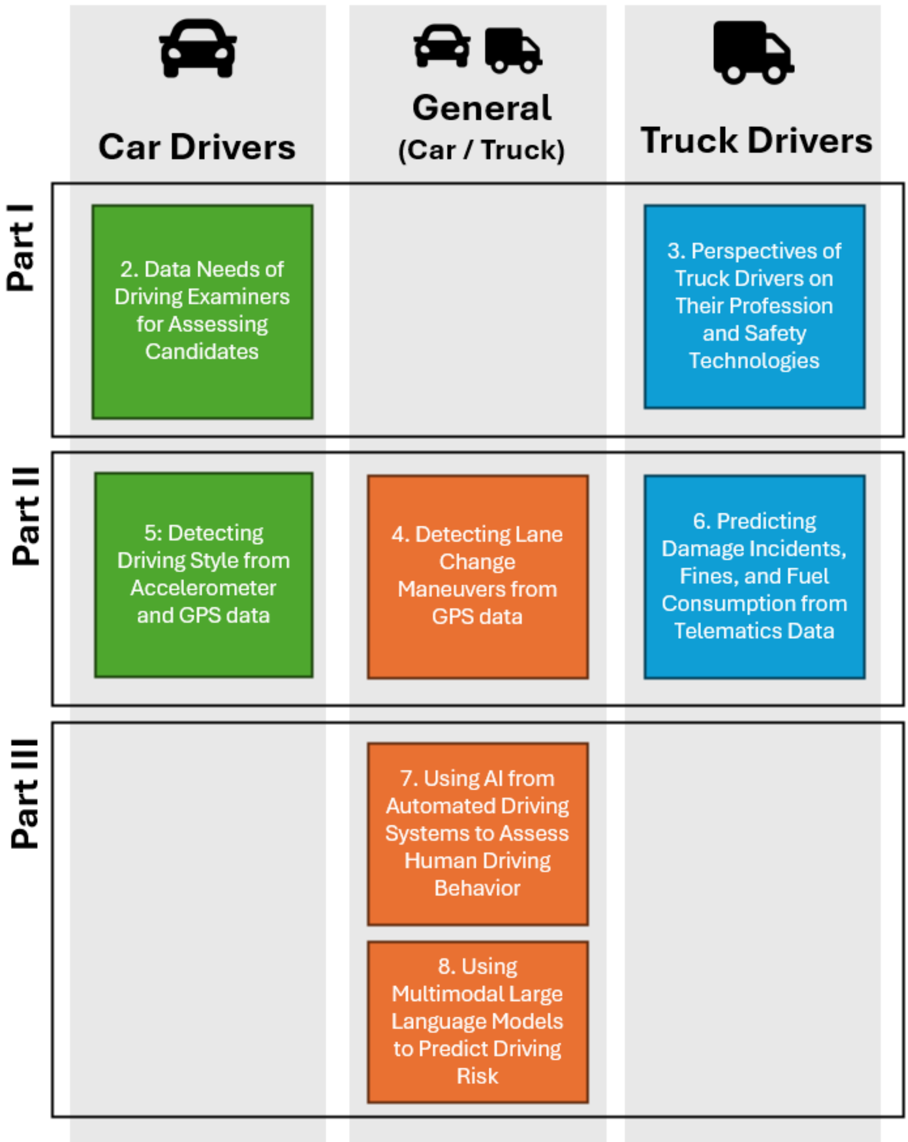


Figure 1.4: Schematic overview of research chapters in this dissertation.

Openpilot produces continuous predictions about the desired state of the vehicle for the event that it has to take control of the vehicle. We use these predictions as a reference to compare human driving performance to, and we show several situations where we can use these predictions to say something about the manner in which the human drove the vehicle.

In *Chapter 8*, we directly "ask" large vision-language models (ChatGPT-4v) to assess the risk in dashcam images and show that such general-purpose models can infer a basic notion of risk in these images with a strong correlation to human ratings.

In summary, this dissertation aims to contribute to an understanding of what the information need is in the practice of car and truck driving (part I) and how the raw information from various data sources (CAN-bus data, accelerometers, cameras) can be processed using both low-cost and accessible techniques (part II) as well as emerging technology such as generative AI and automated driving systems (part III). The main aim is to contribute towards the future of driving assessment and to improving road safety.

References

Admiral. (2024). Black Box Insurance. https://www.admiral.com/black-box-insurance

Allianz. (2024). Telematik-Tarif: Allianz BonusDrive [Telematics tariff: Allianz BonusDrive]. https://www.allianz.de/auto/kfz-versicherung/telematik-versicherung

Allstate. (2024). Drivewise. https://www.allstate.com/drivewise

Amadeo, R. (2021, October 19). The "Google Silicon" team gives us a tour of the Pixel 6's Tensor SoC. Ars Technica. https://arstechnica.com/gadgets/2021/10/the-google-silicon-team-gives-us-a-tour-of-the-pixel-6s-tensor-soc/

ANWB. (2024). Hoe werkt de ANWB Veilig Rijden Autoverzekering? [How does ANWB Safe Driving Car Insurance work?]. https://www.anwb.nl/verzekeringen/autoverzekering/veilig-rijden/hoe-werkt-het

Baughan, C. J., Sexton, B., Maycock, G., Chin, L., & Quimby, A. (2005). Novice driver safety and the British practical driving test (TRL Report No. RR652). Crowthorne: Transport Research Laboratory.

Cruise [@Cruise]. (2023, October 27). (1/3) The most important thing for us right now is to take steps to rebuild public trust. Part of this [Post]. X. https://x.com/Cruise/status/1717707807460393022

Cruise. (2024, January 1). Cruise releases third-party findings regarding October 2. https://www.getcruise.com/news/blog/2024/cruise-releases-third-party-findings-regarding-october-2

De Silva, H. R. (1936). On an investigation of driving skill — I. Human Factors, X, 1–13.

Directive 2003/59/EC. Directive 2003/59/EC of the European Parliament and of the Council of 15 July 2003 on the initial qualification and periodic training of drivers of certain road vehicles for the carriage of goods or passengers, amending Council Regulation (EEC) No 3820/85 and Council Directive 91/439/EEC and repealing Council Directive 76/914/EEC. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003L0059

European Commission. (2019). EU Road Safety Policy Framework 2021-2030 - Next steps towards "Vision Zero". https://transport.ec.europa.eu/system/files/2021-10/SWD2190283.pdf

European Road Safety Observatory. (2021). Facts and figures – Buses / coaches / heavy goods vehicles - 2021. https://road-safety.transport.ec.europa.eu/document/download/953ba67a-7408-4d04-b78d-559d489c67d2_en?filename=FF_buses_hgv_20220209.pdf

Geotab. (2024). Driver coaching with telematics. https://www.geotab.com/fleet-management-solutions/driver-coaching

Learoyd, C. G. (1950). The carnage on the roads. The Lancet, 255, 367–369. https://doi.org/10.1016/S0140-6736(50)90124-3

McKenna, F. P. (2012). How should we think about the three E's: Education, engineering and enforcement? Proceedings of the 5th International Conference on Traffic and Transport Psychology, Groningen, the Netherlands.

Münsterberg, H. (1913). Psychology and industrial efficiency. Boston, MA: Houghton Mifflin. https://doi.org/10.1037/10855-000

NEXTdriver. (2024). Rijgedragcoaching voor transportondernemingen [Driving behavior coaching for transport companies]. https://nextdriver.nl

NHTSA. (2021). Summary of motor vehicle traffic crashes (DOT HS 813 515). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813515

Pathe Films (Producer), & Ford Motor Co. Ltd. (Sponsor). (1935). Your driving test [Film]. Pathe Films. Available on FordHeritage YouTube Channel. https://www.youtube.com/watch?v=BbbERUEsQ4Q

SWOV. (2024). Verkeersdoden in Nederland [Traffic fatalities in the Netherlands]. SWOV-factsheet, April 2024. https://swov.nl/sites/default/files/bestanden/downloads/FS%20Verkeersdoden.pdf

Teppati Losè, L., Spreafico, A., Chiabrando, F., & Giulio Tonolo, F. (2022). Apple LiDAR sensor for 3D surveying: Tests and results in the cultural heritage domain. Remote Sensing, 14, 4157. https://doi.org/10.3390/rs14174157

Tesla. (2024a). Autopilot and Full Self-Driving (Supervised). https://www.tesla.com/support/autopilot

Tesla. (2024b). Cabin camera. https://www.tesla.com/ownersmanual/modely/en_us/GUID-EDAD116F-3C73-40FA-A861-68112FF7961F.html

Thadani, T. (2024, April 28). Lawsuits test Tesla claim that drivers are solely responsible for crashes. The Washington Post. https://archive.ph/NGkbf#selection-463.0-463.73

The Economist. (2024, June 24). China's robotaxis are racing ahead of Tesla's. https://archive.ph/ZFCWp#selection-945.0-945.45

United Nations. (2015). Resolution adopted by the General Assembly on 25 September 2015. https://documents.un.org/doc/undoc/gen/n15/291/89/pdf/n1529189.pdf

Wang, M. (2022). Estimating vehicle speed with consumer grade mobile LiDAR [MSc thesis, Massachusetts Institute of Technology]. https://hdl.handle.net/1721.1/144817

Waymo. (2024). Waymo One. https://waymo.com/waymo-one

World Health Organization. (2023). Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517

Yang, H. (2024). China's robotaxis need regulation. Nature, 632, 259. https://doi.org/10.1038/d41586-024-02551-x

Part I

Perspectives on Data Use and Technology in Driver Testing and the Trucking Industry

2

Data Needs of Driving Examiners for Assessing Candidates

This chapter was published as: Driessen, T.*, Picco, A.*, Dodou, D., De Waard, D., & De Winter, J. (2021). Driving examiners' views on data-driven assessment of test candidates: An interview study. Transportation Research Part F: Traffic Psychology and Behaviour, 83, 60–79. https://doi.org/10.1016/j.trf.2021.09.021

* Joint first authors

Abstract

Vehicles are increasingly equipped with sensors that capture the state of the driver, the vehicle, and the environment. These developments are relevant to formal driver testing, but little is known about the extent to which driving examiners would support the use of sensor data in their job. This semi-structured interview study examined the opinions of 37 driving examiners about data-driven assessment of test candidates. The results showed that the examiners were supportive of using data to explain their pass/fail verdict to the candidate. According to the examiners, data in an easily accessible form such as graphs of eye movements, headway, speed, or braking behavior, and color-coded scores, supplemented with camera images, would allow them to eliminate doubt or help them convince disagreeing test-takers. The examiners were skeptical about higher levels of decision support, noting that forming an overall picture of the candidate's abilities requires integrating multiple context-dependent sources of information. The interviews yielded other possible applications of data collection and sharing, such as selecting optimal routes, improving standardization, and training and pre-selecting candidates before they are allowed to take the driving test. Finally, the interviews focused on an increasingly viable form of data collection: simulator-based driver testing. This yielded a divided picture, with about half of the examiners being positive and half negative about using simulators in driver testing. In conclusion, this study has provided important insights regarding the use of data as an explanation aid for examiners. Future research should consider the views of test candidates and experimentally evaluate different forms of data-driven support in the driving test.

2.1. Introduction

The last decade has seen a vast amount of research on automated driving, spanning areas such as sensor systems (Marti et al., 2019; Schoettle, 2017), computer vision (Ranft & Stiller, 2016; Rangesh & Trivedi, 2019), path planning (González et al., 2015; Marin-Plaza et al., 2018), and control (Farag, 2020; Lima et al., 2018). At the same time, there is a growing realization that fully automated driving may not be achieved within the next three to five decades (Litman, 2021; Shladover, 2016; Tabone et al., 2021). While there have already been compelling demonstrations of automated driving without human intervention, even the most advanced prototypes to date need occasional human intervention or behave in unexpected manners (Boggs et al., 2020; Goodall, 2021), suggesting that for the coming decades, drivers still need to be trained and licensed.

Although fully automated driving is not within immediate reach, cars are becoming increasingly computerized data-collection machines. Modern cars collect not only data about ego-vehicle state (e.g., speed, heading) and driver input (throttle, brake, and steering) but also data about the environment (presence of other road users and their speed, lane boundaries, infrastructure) via cameras, lidar, radar, or ultrasonic sensors as part of advanced driver assistance systems (ADAS). In addition, data from the traffic environment are now also often collected via nomadic devices such as MobilEye (Chen et al., 2017) and via dashcams and smartphones (Ahmad et al., 2021; Tummala et al., 2019).

The growing proliferation of computers in cars makes it possible to use these computers to assess driver behavior. The literature shows an increasing number of applications that use forward-facing or driver-facing cameras, sometimes combined with acceleration-based triggers, to detect drowsiness and distraction (Chowdhury et al., 2018; Kashevnik et al., 2019; Lechner et al., 2019; Ramzan et al., 2019; Sikander & Anwar, 2018) and unsafe driving behavior (Hickman & Hanowski, 2011; Mase et al., 2020). Other types of systems rely on in-vehicle data recorders (Shimshoni et al., 2015) or smartphones for driver assessment (e.g., Bergasa et al., 2014; Farah et al., 2014; Musicant & Lotan, 2016; Shanly et al., 2018), and see Michelaraki et al. (2021) for a review on post-trip feedback solutions, including smartphone apps, gamification approaches, and reward schemes. Schöner et al. (2021) proposed a concept where a norm-referenced driver safety score was computed relative to the time headway and time-to-collision distributions of a large highway traffic dataset. Similarly, usage-based insurance, also known as pay-as-you-drive insurance, commonly uses smartphones or dongles to obtain measures of driver risk such as speeding, hard braking, or other kinematic events (Arumugam & Bhargavi, 2019; Händel et al., 2013; Vavouranakis et al., 2017) and rewards safe behavior with reduced insurance premiums. Relatedly, in motorsports, data acquisition for assessing driving performance is the norm (Segers, 2014).

With the increasing capabilities of in-vehicle sensors and computers, it may become viable to flag deviant driving behavior automatically. This notion can be traced back to the Generic Intelligent Driver Support (GIDS) project, which proposed a tutor that functioned based on the difference between observed driving behavior and reference driving behavior (Michon, 1993). Adaptive training, intelligent tutoring, and driver profiling have already been available in simulator-based driver training for many years (Boelhouwer et al., 2020; De Winter et al., 2008a; Graesser et al., 2005; Karvonen et al., 2006; Ropelato et al., 2018; Wassink et al., 2006; see Zahabi et al., 2020 for a review on adaptive training in simulators). Today, these tutoring concepts are becoming feasible in actual cars. Fridman et al. (2019) demonstrated a real-time intelligent driving system that supervised a second intelligent driving system: disagreements between the two steering angles were found to be predictive of critical situations (automation-to-manual hand-overs). By extension, it should also be possible to have a similar system detect unexpected manual driving behaviors. In the same vein, researchers have performed on-road studies with a personal assistant for fuel saving (Magaña & Muñoz-Organero, 2015) and with an intelligent driving assistant that used an accident risk map and vehicle telemetry as inputs (Terán et al., 2020). In summary, automated assessment of driving tasks seems within technological reach.

The concept of automated driver testing is not far-fetched, at least when it comes to basic driving skills. The Roads and Transport Authority of Dubai has recently implemented a driving test that uses instrumented cars on a driving range, and where the pass/fail verdict is supposedly provided automatically (Government of Dubai, 2019). In 2018, Microsoft introduced Harnessing AutoMobiles for Safety (HAMS), an automated driver license testing system that relies on a smartphone mounted on the windshield and which produces an assessment without human intervention (Nambi et al., 2018), in an attempt to eliminate bribery of the examiners (Giridharan, 2019; Microsoft, 2021).

The above developments may be of strong interest to driving license organizations, which face long-standing challenges regarding the reliability and validity of their driving tests. One issue is test-retest reliability, estimated by performing the driving test twice with different examiners, and the other is inter-examiner reliability, estimated by using two driving examiners in the car assessing the same candidate. Experiments in these areas are scarce but suggest only low test-retest reliability, presumably because traffic conditions change from test to test (pass/fail congruence of 64% in Baughan & Simpson, 1999, and 63% in Olweus, 1958; the latter as cited in Alger & Sundström, 2013), and logically higher inter-examiner reliability of the same test (72% in Bjørnskau, 2003, and 93% in Alger & Sundström, 2013). The high inter-examiner reliability may reflect high-quality assessment procedures, or as noted by Alger and Sundström, *"One possible explanation for the high examiner agreement in Sweden is that quality in the driving test and consistency of assessment are continuously discussed among examiners."* (p. 28). But even in the study by Alger and Sundström, which found very high inter-examiner reliability, there were occasional disagreements between the two examiners. For example, there were cases where the interpretation of the severity of the candidate's faults or speed adjustments differed between the examiners, or where there were disagreements about whether the candidate should be penalized for faults on specific tasks or should be assessed more holistically. It is noted that low test reliability may be expected if the driving test admits candidates who are *just* good enough to pass, if there is variability in the testing conditions (traffic, weather, road types), or if the driving test is only short (Baughan & Simpson, 1999; De Winter & Kovácsová, 2016).

Another issue is that of predictive validity. Driving test outcomes are not necessarily good predictors of safe post-license driving, as males have been found to perform better on the driving test than females (Crinson & Grayson, 2005; Mynttinen et al., 2011), even though males are overrepresented in post-license crashes (SWOV, 2016). That said, a recent interview study with 13 driving instructors found that instructors often have a sense about whether the learner driver has a risky attitude, lack of concern for safety, or overconfidence (Watson-Brown et al., 2021). These findings are consistent with a study that found that risky pre-license driving in a simulator can predict self-reported post-license traffic violations 3.5 years later (De Winter, 2013). The above factors suggest that driving instructors and examiners may benefit from driver performance data to complement their verdict in a predictive-valid way, pinpoint driving deficiencies, or contribute to the inter-examiner and interregional calibration of driving norms.

In the Netherlands, prospective drivers follow, on average, 40 hours of training at a private driving school before applying for the driving test (Roemer, 2021). The Netherlands uses a test-led model, where the driving test implicitly determines the content of the preceding driver training (Helman et al., 2017). Next to an exam on theoretical knowledge, the driving test, organized by the Dutch Central Office of Driving Certification (CBR), involves 35 minutes of driving, of which 10 to 15 min using a route navigation system. The candidate is assessed based on seven elements of participation in traffic: driving off, driving on straight and curvy road sections, behavior near and at intersections, merging/exiting, overtaking/moving sideways, behavior near and on special road sections, and special maneuvers. The Dutch driving test has undergone various recent modifications, such as the introduction of hazard perception in the theory test and a self-reflection form to be completed before the on-road test (consistent with the Goals for Driver Education; Hatakka et al., 2002). Supervised driving has been introduced as well since 2011 for drivers who obtained their license between their 17th and 18th birthdays (2todrive, 2021). After their 18th birthday, licensed drivers are allowed to drive independently. In introducing further modifications to the driving test, such as the possible introduction of data-driven assessment, it is important to consider the users of such systems, that is, the examiners. User acceptance is crucial, as was also pointed out by De Waard and Brookhuis (1999) in the context of driver support systems: *"A system may function perfectly in the technical sense, if it is not accepted by the public, it will not be used"* (p. 50). In the context of driver testing, acceptance by examiners is crucial.

In the current study, semi-structured interviews were conducted to examine what driving examiners think about the prospect of data-driven assessment. A broad perspective was taken, where we first asked the examiners how they view the current driving test. Subsequently, the interviews went into depth about specific forms of data-based assessment, starting with simple concepts such as automated recordings of speed infringements. However, we also asked the examiners whether they think their task could be replaced by a computer entirely. The interviews also addressed how and when the assessments should be delivered, e.g., during or after the driving test. Additionally, it was asked whether sharing the driving data with different stakeholders would be a welcome idea, an important topic in the era of computerized cars (De Winter et al., 2019; Pugnetti & Elmer, 2020). Finally, we asked some open-ended questions about whether the examiners think that their organization is open to technological change and whether they think that driving simulators, i.e., tools that allow for accurate data recording, could have a role in driver testing.

2.2. Methods

2.2.1 Participants and recruitment

A total of 39 driving examiners were recruited, of whom 2 (P21 & P23) canceled their participation, leaving 37 examiners who participated in this interview study. They all were examiners of the driving license "B", which allows driving cars of up to 3500 kg. Twenty-eight participants were male, and nine were female. The average age of the examiners was 46.8 years (*SD* = 9.0 years), ranging from 31 to 62 years. They had on average 9.0 years of experience as an examiner (*SD* = 7.8, min = 1, max = 29), and 51% (*n* = 19) had worked as a driving instructor before (for an average of 12.8 years, *SD* = 7.0, min = 4, max = 27). The examiners reported performing an average of 37.3 driving tests per week (*SD* = 7.7, min = 0.5, max = 47.5). The examiners were recruited from all 12 provinces of the Netherlands, with at least two examiners per province. They were all employed by the Dutch Central Office of Driving Certification.

An invitation email was sent to 17 driving test managers across the Netherlands, together with a one-page description of the study and its aims. The managers then provided the contact details of examiners willing to participate. Before the interviews, the researchers sent the examiners the informed consent form in Dutch via email. It included the following description of the research aim: "... *to investigate driving examiners' views on the validity of current assessment methods, the possibilities of data-driven assessment, and the type of vehicle data they would like to have to support their judgment*". The research was approved by the Human Research Ethics Committee of the Delft University of Technology (approval number 1418).

2.2.2 Procedure

The interviews were conducted online via Zoom and Microsoft Teams between the 15th of February and the 1st of March 2021. The interviews were conducted by two first authors. Author 1 conducted the interviews in Dutch, whereas Author 2 interviewed in English because she was not a Dutch speaker. Participants willing and able to be interviewed in English, based on self-evaluation of mastering the English language, were interviewed in English by Author 2, whereas the rest were interviewed in their mother tongue (Dutch) by Author 1. As a result, 11 of the interviews were conducted in English and 26 in Dutch.

Each interview lasted approximately one hour. Consent from the participants was recorded orally before the start of each interview. The video recordings of the interviews were stored separately from the consent recordings, in compliance with the data management plan of the project and privacy regulations. The participants were interviewed during their working hours and did not receive additional compensation.

2.2.3 Interview structure

The interviews were semi-structured according to an interview guide (see Appendix A). The questions were divided into three parts: (1) examiners' opinions about the current driving test, (2) examiners' opinions about a data-driven driving test, and (3) general questions. The interviewers occasionally asked follow-up questions based on the topics mentioned by the participants.

Examiners' opinions about the current driving test

Examiners were asked what the strengths and weaknesses of the driving test are today. Additionally, they were asked whether the test allows them to assess if a candidate would drive safely later on—to obtain a general idea of the perceived effectiveness of the driving test. Furthermore, it was asked whether the examiners' intuition plays a role in establishing the verdict—to better understand the process examiners go through when evaluating a candidate.

Examiners' opinions about a data-driven driving test

The second part of the interview concerned the possible implementation of data in the driving test. The questions were divided into three topics: (1) examiners' opinions about the use of data, (2) examiners' opinions about the characteristics the data should have, and (3) examiners' views about the future of the driving test.

Examiners' opinions about the use of data
Examiners were provided with a short explanation: *"With the development of new technologies in the vehicle, it is possible to monitor drivers' behaviors and obtain data regarding their driving performance"*. They were then asked if they thought data of any form could be of help in the driving test. They were encouraged to provide any example they could come up with. After allowing some time for the examiner to come up with examples themselves, the interviewer screen-shared a PowerPoint slide, revealing seven suggestions (Figure 2.1) one by one. After an item appeared on the slide, examiners rated its usefulness from 1 (not useful at all) to 5 (very useful) and were encouraged to explain their rating. The presentation order of the seven suggestions was randomized for each participant.

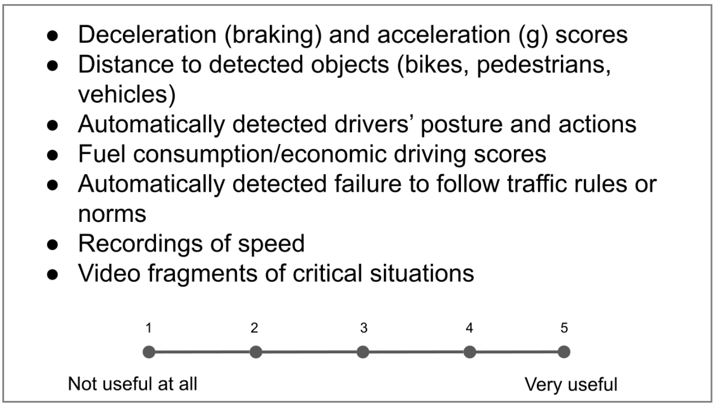


Figure 2.1: Slide with seven suggestions of data forms presented.

Examiners' opinions about the characteristics the data should have
Next, the interviewers asked when data should be collected and when and how the collected data should be presented to them. Examiners' opinions about sharing the collected data were also asked.

Examiners' views about the future of driver testing
The examiners were asked whether a part of their task could be automated. Furthermore, they were asked if they thought that artificial intelligence (AI) would one day be able to evaluate a driver completely and what it would take for them to rely on such an automated evaluation. Examiners were also asked whether simulators could play a role in the driving test and how they see the future of driver testing.

Closing questions

The interview was concluded with generic questions about the possible views of candidates and the licensing organization about data-driven assessment, whether COVID-19 would affect the future of driver testing, and the use of data in other types of tests, such as for motorcycles or trucks.

2.2.4 Data processing and analysis

The interviews were fully transcribed in their original language. This was done automatically by submitting the audio channel of the recordings to AmberScript (2021) for the Dutch interviews and by using built-in transcription services of Zoom and Microsoft Teams for the English interviews. Errors in the transcripts were manually corrected by Authors 1 and 2. Qualitative analysis While conducting and transcribing the interviews, the interviewers familiarized themselves with the participants' responses, similar to the 'familiarization phase' described by Braun and Clarke (2006). Authors 1 and 2 discussed common responses given by the examiners to decide how to subsequently structure the results.

After further discussions with Authors 3–5, it was decided to retain the structure of the interview guide (Appendix A) for presenting the results. Several questions were omitted (i.e., a question about the differences between driving skill vs. driving style, and the closing questions listed above) since these questions did not appear to yield substantive new insights. Furthermore, the responses to several questions were merged to prevent repetition. For example, responses to questions regarding the recording and presentation of data were grouped in a section entitled: "Introducing data in the driving test: how?" One topic that was not in the interview guide emerged clearly during the interviews. This concerned the examiners' motivations for using data (i.e., as an explanation aid vs. decision aid), addressed in a new section "Introducing data in the driving test: why?"

For each interview topic (e.g., strong points about the current driving test, weak points about the current driving test, future of simulators, etc.), the corresponding quotes were grouped in separate documents. From these documents, quotes were selected to be presented in the narrative of the Results section. The quote selection was done using input from the above-mentioned familiarization phase, based on their potential to explain the topics in a concise manner and by ensuring no over-representation of specific participants. The selection and translation of the presented quotes was done by Authors 1 and 2. Authors 3 and 5 reviewed the selected quotes and corrected the translations using the transcripts.

The above process was applied iteratively, where the results structure was subjected to minor revisions during the quote selection process.

Quantitative analysis

In addition to the qualitative analysis, a quantitative analysis was performed. Firstly, means and standard deviations of the examiners' usefulness ratings of the presented data examples were computed. Secondly, counts were reported for questions that yielded categorizable responses and where counts would illustrate the degree of consensus among participants. More specifically, participant counts were reported of the strong and weak aspects of the exam, examiner-generated data-usage examples, motivations for using data, views about data sharing, and views about the future of driver testing. Responses were tagged (e.g., #strong-aspect.examiner-freedom) in the raw transcripts in Microsoft Word by Authors 1 and 2, a process that was counterchecked by Author 3. Subsequently, the tagged responses were automatically counted using a custom-built Python script (available in Appendix B). This script avoids multiple counts of participants who raised the same item more than once.

2.3. Results

2.3.1 Examiners' opinions about the current driving test

On average, the examiners mentioned 1.22 (SD = 0.47, min = 1, max = 3) strong aspects and 1.35 (SD = 0.67, min = 0, max = 3) weak aspects of the current on-road driving test. Table 2.1 summarizes the responses. Responses mentioned only once were grouped in the category 'Other'.

Table 2.1: Strengths and weaknesses of the current driving test. n is the number of examiners (responses given more than once by the same examiner count as one)

	Response	n	%
Strong aspects	Examiner flexibility	14	38
	Human aspect	10	27
	Candidate independence	7	19
	Objective basis	6	16
	Examiner training	2	5
	Other	3	8
Weak aspects	Lack of time	20	54
	Test variability	8	22
	Test is a snapshot	4	11
	Nervous candidates	3	8
	Poor candidate level	3	8
	Test is too strict	2	5
	Other	6	16

Strong aspects

Examiner flexibility

Examiners mentioned as a strong aspect that they are allowed some degree of freedom to make judgments and decisions.

> "*... I find it nice that, as an examiner, I am not too bound by all kinds of rules about the assessment. … There is, of course, a framework in which I have to operate, I have to comply with, but I also have my own responsibilities, and I can use my own knowledge and experience to, yes, weigh certain parts. … I also experienced how it was done in the past; it was all with fault codes. Well, I think that is much better now; for example, an item in the past was: 'examiner intervention meant a failed exam'. Well, now, I can think about that for myself. What is my opinion? I am very happy with this.*" (P7, Translated)

Relatedly, examiners were positive that they are expected to make a holistic assessment of the candidate's ability to drive independently.

> "... *in the past, we simply had fault scores. Four strikes, or two, three strikes simply meant: failed. Now we have something like an overall image, which means that if the overall image is better than the mistakes made, someone can still pass.*" (P2, Translated)

> "*…you are not sitting in a car with an abacus, like: that did not go right, not right, right, right, right, not right, and eventually a number will come out, as it has been in the past. … Now an overall picture emerges. And if someone braked too late, ok, that can happen. If they do that once, compared to the whole ride which they do very well ... then I can live with that.*" (P27, Translated)

Besides having flexibility in the assessment, the examiners pointed out they have the freedom to investigate. Candidates may be guided to different situations based on their performance so far.

> "... *it is just nice that you currently have the freedom in the exams to adjust your route. The moment I see, for example, [that] you have problems with roundabouts. ... Well, then I just take a section with many roundabouts to re-test.*" (P1, Translated)

> "... *sometimes people do not drive as they should according to the procedure, but you still feel safe. And you can test it, of course. You take some extra routes, junctions, or anything, and well, I think the strong point is that we have the freedom …*" (P10)

Human aspect
A strong aspect mentioned by ten examiners is the 'human aspect' of the test. Examiners have to make the candidates comfortable, trying to make them less anxious.

> "... *it is really important for the examiner; it is our job to get the candidates reassured and make them feel at ease so they can drive how they normally do in driving lessons with their instructor.*" (P12)

The examiners noted that this comforting facet is new; they mentioned that the guidelines for examiners evolved positively over the past years.

> "*I think the strong thing about the exams now is that we try to comfort students more now.*" (P10)

Testing candidate independence
Seven examiners mentioned that a strong aspect of the driving test is that it tests for driver independence.

> "*Well, I think it is strong to the extent that they at least have to show a decent degree of independence.*" (P29, Translated)

During the exam, candidates are asked to drive independently to a destination specified by the examiner. The candidates use a navigation system for this task. Four of the seven examiners who praised candidate independence mentioned the independent route driving part specifically.

> *"At the moment, I think one of the strongest aspects [is] that they have to navigate themselves. That was not in it [the driving test] before, and now it is. Because I am still of the opinion that this best approximates how a candidate will eventually behave on the road ..."* (P15, Translated)

Objective standard

Six examiners found the procedure of the driving test a strong aspect. They mentioned that the procedures contribute to objectivity. Some brought up the driving procedure document (CBR, 2020), sometimes referred to as their "Bible" (P39, P4, P36), and noted its positive impact on the driving test.

> *"It is called the driving procedure. What we expect of the candidate ... is all written out in this procedure. All aspects. So, this is the most objective way that we can let candidates take the driving exam. It is very clear: what we expect ... and how we judge it."* (P4)

Weak aspects

Lack of time

Twenty examiners mentioned lack of time as a weak point of the driving test. They indicated that due to high traffic density and decreased speed limits, it has become challenging to (re-)test the desired skills within the 35 minutes of driving time.

> *"... when I look at 25 years ago and now, when looking at the traffic intensity, but also the residential areas that are now all 30 km/h zones, then time is sometimes short."* (P6, Translated)

> *"... traffic is so busy that you cannot always test everything. You are constantly thinking: 'okay, ... I also have to be back in time for the next test.'"* (P3, Translated)

> *"... I think we need more time. When we want to test well, we need more time to assess [candidates]. … Within the short amount of time, we cannot always do long stretches of highway, stretches outside urban areas ..."* (P15, Translated)

> *"[A] shortcoming is that … due to the time, the area in which you drive is restricted, the radius around your place [examination office]."* (P29, Translated)

Test variability

Eight examiners reported that the variability in testing conditions is a weak point of the driving test. They mentioned that traffic conditions are variable and dependent on the time of day or testing location.

> *"... we try to do every exam the same, but traffic situations can be completely different. For instance, an exam on Tuesday morning at eight o'clock is completely different from an exam on a Saturday morning at eight."* (P18)

> "*... traffic all around the country is different from place to place. It is a lot harder to do a test over here in the east of the Netherlands.*" (P8)

Furthermore, individual differences between examiners are a source of variability.

> "*Of course, because it is a human effort, there will be people [examiners] who make the route more difficult than may actually be necessary.*" (P1, Translated)

> "*I think that my colleagues and I can be really all-determining. How I create my atmosphere, or things I could say, well-meant, or maybe not well-meant; with that, you can get somebody to a certain verdict [pass/fail], I think that that may be the weaker aspect of the test.*" (P2, Translated)

Predicting safe driving

It was asked whether the current driving test allows examiners to predict safe driving later on. Some examiners interpreted this question as to whether the test helps them assess candidates effectively. From this perspective, the answers were generally positive, with a few mentions of the lack of time.

> "*In general, yes, it does. But well, every now and then, there are times, like I said, that time-wise, [it is] always a bit tight. If you would have more time, the verdict is probably going to still be the same. But sometimes you just need some more time to check; you reevaluate.*" (P18)

The second interpretation regarded the ability to predict whether a driver will drive safely after having passed the driving test. The examiners generally indicated that they are not well able to predict what will happen in the future.

> "*Well, in order to give an honest answer to that question, [If] I would be able to know how they drive after the exam. I do not.*" (P14)

The examiners specifically pointed out that candidates can pretend: they may adopt an appropriate driving style but reveal themselves as aggressive drivers or risk-takers when driving independently.

> "*... I am sure that the candidates that pass the test because they drive in the way we would like him to drive at that time. But for sure, later on, they will change their attitude in traffic.*" (P8)

Examiner intuition

The examiners indicated that intuition plays a role but not to the point of deciding on a verdict. Their intuition may, for example, help them assess situations more quickly.

> "*I do not know if it is intuition or if it is knowledge. Because when you have done more exams, you can recognize sooner where the problem might be or what was good and what is not.*" (P13)

Furthermore, based on their intuition, or 'gut feeling', examiners may formulate hypotheses to be tested by gathering additional information.

> *"You sometimes have a particular gut feeling about behaviors or from certain expressions they give. This can mean that you sometimes want to check a certain thing, an intersection or perhaps a highway, by which the behavior you have a gut feeling about, that is, the so-called intuition, may or may not come out positively."* (P29, Translated)

Although intuition can influence the route driven, according to the examiners, the final judgment is always based on facts and procedures.

> *"In principle, we base ourselves on what we see, so the actual facts. We have to judge based on that."* (P34, Translated)

> *"So you cannot let somebody fail because your intuition tells you it is not good enough ... If one fails, you have to tell them facts. And if you do not have any facts, you cannot let them fail."* (P39)

2.3.2 Examiners' opinions about a data-driven driving test

After introducing the basics of driver monitoring, the interviewers asked the participants to express their views about the use of data in the driving test. Some examiners were positive and enthusiastic about the idea:

> *"I think so; well, I am quite positive."* (P7, Translated)

> *"Yes, absolutely; I totally agree."* (P34, Translated)

A small number of examiners based their answers on previous experience with video recordings or driving simulators, when they were still driving instructors:

> *"That is why we used it [video recordings]. It helps a lot. We saw a big difference between when we were using it and before we were using it."* (P39)

> *"We had a camera ride ... you can see a lot of things … it was really helpful because I had some students who were very, well, a bit naughty or stubborn, and they saw themselves, and they were like: 'Whoa! Am I doing that?'"* (P10)

Others were a bit more hesitant or asked for clarification.

> *"... I think … my feeling says no … "* (P15, Translated)

> *"I have been thinking since I received the invitation for this interview what kind of data would that be."* (P14)

During the interviews, discussions emerged about different uses of data. The interviews addressed the different purposes of data mentioned during the interviews ('why'), what data examiners would like to use ('what'), and how the specifics of data recording should be arranged, such as delivery, data sharing, and moment of recording ('how').

Introducing data in the driving test: why?

The interviews addressed whether data could help examiners come to their verdict. Two main motivations for the use of data became apparent from the interviews: using the data for explaining the verdict to the candidate (explanation aid) and using the data to support the examiner in arriving at the verdict (decision aid).

Explanation aid

The examiners saw merit in the use of data as an explanatory tool. As many as 36 of the 37 examiners mentioned they would want to use data for this purpose. They mentioned encountering candidates who refuse to accept a fail verdict or who even become aggressive or file a complaint. Thus, the examiners would like to have objective data to back up their assessment.

> *"I already make my judgment without all these things, of course. ... The only way it would be useful is avoiding the discussion and avoiding the aggressiveness and the angriness, and the one who is going to threaten you or file a complaint. Because they do not have any grounds if you can show them 'look here'. ... I would use this after the exam to back up my story."* (P16)

> *"If you tell a certain person or candidate that their [following] distance is too short, they will often defend it using the motto 'I think this is sufficient', so to speak. If you can show based on the equipment, how often they have not kept a sufficient distance, for example, that would be an addition."* (P29, Translated)

> *"If you assume that an examiner is competent, then you can actually tell, regardless of the data, whether someone is fit to drive ... No data is necessary, I would say, because I can just see they have a too short following distance. I can also see whether they are driving too fast. I think that data is very useful to get the candidate to feel: 'yes, that examiner is in fact right; I indeed was not safe there … I indeed did not look properly there. I indeed drove too fast there.'"* (P33, Translated)

Decision aid

The interviewers often followed up with questions such as *"And could the data also help you establish the verdict?"* or *"Could the data also help you in making decisions?"*, to assess the use of data as a decision aid. Overall, the examiners believed they do their job well already and do not need data to judge a candidate. More than half of the examiners indicated they would not want to use data to reach the pass/fail verdict.

> *"Yes, but not to make the decision. I know how to decide if somebody passes or fails. I do not need data for that… I do not see the possibility, really, yet, to help me make my decision."* (P14)

> *"We are at this moment strong enough to come to a verdict."* (P6, Translated)

Several examiners were positive about the use of data as a decision aid, however. They mentioned that it could be useful to obtain extra information, as they cannot pay attention to everything. Data availability could also support examiners' memory and let them review situations through video recordings. It was also argued that data could be used to improve objectivity.

"... we cannot follow candidates' eyes during 100% of the test. ... We miss some things, I think, sometimes in really important moments. Because we have to be aware and pay attention to the traffic." (P8)

"... those technologies are developing rapidly at the moment. … and I think we might be able to be assisted by those systems. Possibly with exam results or as support. ... support systems that help us establish a proper result [of the driving test]." (P24, Translated)

"... certain things stick with you [in memory]. And why wouldn't it be the case that actually too many negative things stick, and that the positive things do not stick enough, or vice versa. So, I think, if you just set objective data to that, and you can preview it before you present the test outcome ... I think you might well be in for some surprises. You might think: 'darn, my own view was different after all.'" (P11, Translated)

Other uses of data

Besides explanation aids and decision aids, the examiners brought up other potential uses of data. Data could also improve the way new drivers are taught, for example, by learning from previous mistakes.

"You don't have any discussion because it is all clear; everybody can see it. For the candidate but also for the instructor. And they can, for instance, if it is not good, they can practice with it because they know exactly what happened." (P18)

A recurring issue the examiners mentioned is that they often encounter students who clearly lack the skills required to pass the driving test. Data about the number of hours of training or the training conditions encountered could be used to preselect candidates for the driving test.

"Certain schools often just come with candidates who are far from ready for the driving test." (P7, Translated)

"that you know ... that [the candidate] fulfills minimum requirements like 'this many training hours, training at different times of the day, … and this many hours … on the highway.'" (P38, Translated).

It was further mentioned that data could help create uniform norms and assess the effectiveness of driving test elements, such as special maneuvers. Another suggestion from the examiners was to analyze the routes driven. By integrating this information with traffic intensity data, optimal routes could be generated that bypass congested intersections.

Introducing data in the driving test: what?

Data suggestions by examiners

Table 2.2 lists the examples brought forward by the examiners when asked what data they could use in their work. Furthermore, Figure 2.2 shows the ratings of the concepts provided by the interviewers (additional statistics can be found in Appendix A).

Table 2.2: Examples of the types of data that could be used, brought forward by examiners. *n* is the number of examiners who brought up the response (responses given more than once by the same examiner count as one).

Response	*n*	%
Gaze behavior	22	59
Recordings of speed	20	54
Distances	10	27
Position on road	8	22
Braking	8	22
Eco-driving	5	14
Video recordings	5	14
Reaction time	3	8
Traffic signs	3	8
Vehicle handling	3	8
Other	12	32

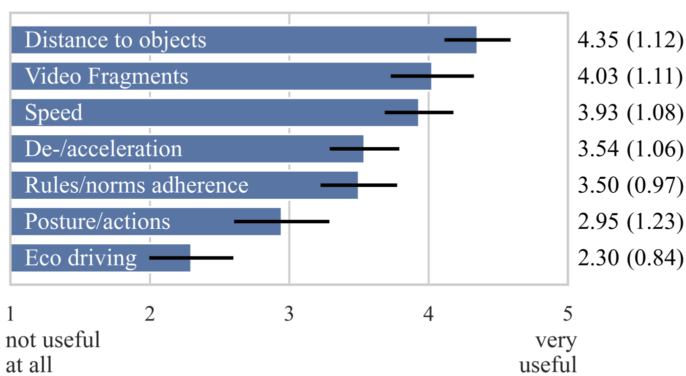


Figure 2.2: Means (*SD*) of the ratings of the data examples provided by the interviewers ($n = 37$). Error bars represent 95% bootstrapped confidence intervals computed using Morey's (2008) method for within-subjects designs.

Items that received a high rating were typically discussion points between examiners and candidates. Most frequently, the examiners suggested capturing the candidates' eye-gaze behavior.

> *"Another thing is looking, observing, ... we can pretty well see in a mirror what somebody is looking at, but it could be supported by data ..."* (P19, Translated)

Eye-gaze data could help prevent discussion or misunderstandings.

> *"Gaze behavior, you could look … at what the candidate is looking towards … then [I think] that you can get a lot of misunderstandings out of the way."* (P5, Translated)

> "*We call it 'viewing technique', and that is always a discussion point. When you have a candidate, and you say 'you are not looking, or you are not looking right or enough' ... [With] access to data, how they [the candidates] look, … you avoid that discussion because you can show it."* (P12)

The examiners were aware that the eyes pointing somewhere does not imply that the driver perceived the event.

> *"But well, I have seen people looking into a street, looking straight at the car, and who still continue driving."* (P22, Translated)

> "*I cannot see if you really perceive something. I can only tell from the action you make."* (P1, Translated)

The opposite was also noted: things can be perceived even when the candidate does not appear to have moved the eyes towards the object.

> *"I do not see him look anywhere, and still he responds to everything he should respond to. So, well ..."* (P9, Translated)

The examiners frequently indicated they would like to have access to recordings of driving speed. They saw value in such recordings, especially if the recordings could be related to the driving context, such as oncoming intersections, curves, or before merging onto the highway.

> *"A simple example is speed. An important item in the driving test is an intersection within the built-up area. These should proceed safely. When you drive too fast there, that's not safe."* (P33, Translated)

> "*I could use this [speed recording] in many cases to explain that someone is approaching too fast, for example, approaching that intersection too fast, approaching too fast on roundabouts."* (P36, Translated)

> "*We are sometimes having a discussion about it [about speed], you know; we have a very nice speedometer we can look at, and yet there is still discussion about how fast someone has driven."* (P19, Translated)

Measuring the distance between the vehicle and other road users/objects was regarded as valuable as well. Distance was seen as relatively easy to measure via sensors.

> *"Take, for example, keeping distance, that two-second rule. You can, of course, measure very well whether … enough distance is kept ..."* (P17, Translated)

> *"What average drivers — I am not even talking about novices but also existing drivers — find difficult: ... keeping distance. And the modern car of today can show, register, what your current headway is."* (P29, Translated)

An index of lane-keeping behavior was also regarded as valuable.

> *"Position on the road on straight roads, I could use that, because there are candidates who really zigzag on their way; they just have no feeling of staying in the middle of a lane."* (P28, Translated)

Cars used for driver training and testing have a double control system. In the case of a hazard, the examiner can (preemptively) press the brakes. Such an intervention can be a source of debate, where data may be of help.

> *"We hear from candidates almost daily, like: 'I also braked myself' [after an examiner intervention]. We are generally professional enough to be able to see whether somebody braked or not, or if we have done that [ourselves]. Data can just point that out. Data could simply tell: 'you have or have not braked at that moment.'"* (P19, Translated)

Video recordings were deemed useful for providing feedback to the candidates about traffic violations or aberrant behaviors.

> *"Yes, sure, that would be … after the driving test, just like in 'Blik op de weg' [Dutch TV program showing drivers committing traffic violations], where they are able to rewind the violations very quickly, so that you can see: '… in this situation ... here you did that, and I did this.'"* (P3, Translated)

> *"What you sometimes see when … being stopped by the police, is like 'Please come along, and then we could show you some video footage.'"* (P6, Translated)

The examiners indicated that body position was not of much interest to them.

> *"If you can drive a car [sitting] backward better than with your nose forward, then you should do that."* (P11, Translated)

However, the examiners were relatively enthusiastic about the measurement of head posture to extract eye gaze behavior, which may explain the overall medium rating of posture/actions (Figure 2.2).

Even though eco-driving is formally part of the assessment criteria (CBR, 2020), eco-driving was rated the lowest. Examiners reported that, in practice, candidates do not fail the test because of bad eco-driving. Examiners noted that factors causing bad eco-driving are usually bad vehicle handling, such as over-revving or jerky driving, which are factors that are obvious to an examiner without data support.

Introducing data in the driving test: how?

When should data be recorded?

Examiners were asked whether data should be collected before, during, or after the driving test. Some examiners answered that the driving test should only consider the ride itself: they would not want to take data from the driving lessons into account, for example.

> *"No, I do not think that is of any added value for my final test. Each test is on its own."* (P5, Translated)

Obtaining additional information about the candidate still appeared interesting to some examiners. They explained that receiving data from previous driving lessons may help reduce the 'snapshot' issue:

> *"[It is] often that they tell me that it is due to nerves, and I think that if the data show that it is true, that they have done things differently [in the exam] than they have done during the lessons, then it could be helpful in some way ..."* (P13)

The examiners pointed out there is a conflict of interest with the driving instructor and that they have no control over how the pre-examination data are collected.

> *"Perhaps I want to say very carefully that I do not trust them [instructors] enough for that. You can manipulate it now; it can be set up. I want it to be objective; it is too important."* (P4)

> *"The difficulty with using the data from the lessons, I cannot see if it is the instructor who says 'oh, you have to slow down' or 'you have to wait' or 'you have to …'. So I do not know if it is from [the candidates] themselves."* (P35)

Benefits of data were also identified for the candidates: data about the driving lessons could help them get familiarized with data and facilitate self-reflection.

> *"I think that if there will be data in the lessons, they will be better prepared. It gives them some self-reflection."* (P14)

> *"I remember seeing images from a certain driver training course of how I acted in traffic, on a motorcycle ... I think that for driving schools, this can be a very useful tool, simply because you have immediate feedback."* (P36, Translated)

Some benefits of data collection after the driving test were mentioned as well. The examiners pointed out that post-license data collection may help candidates self-reflect and support them in the first years of independent driving.

When should data be delivered?

Three options were discussed with the examiners: to receive data before, during, or after the driving test. A clear outcome was that receiving data during the test was judged as impractical. It would only be possible for small amounts of information or information that is very easy to process.

> *"In essence, most of the data we collect ourselves in real-time as a human being. I do not think it is useful to have one more extra stimulation. In that regard, the job is sometimes already busy enough: all your senses are used."* (P36, Translated)

> *"During the exam, we have enough to do. Paying attention to the traffic, paying attention to the candidates, maybe to the instructor—if he or she is allowed to come with us. So, well, maybe about speed, that can be data that you can just read from a screen, maybe, or you can see in a blink of an eye."* (P10)

Real-time feedback based on the data collected was also considered. It could assist examiners in making split-second decisions such as taking the wheel to ensure safety. In that sense, it would be acceptable to receive data while driving.

> *"So this would be helpful if there is, like a sound in the car at the moment, someone, a pedestrian, or a bike is too close, and I would just take the wheel."* (P16)

Some examiners stated it would be useful to adapt the driving test based on the information recorded during the preceding driving lessons. However, there were concerns about not being objective anymore.

> *"I think you could be influenced if you already know someone, what their weak spots are. I would not want to know myself."* (P18)

Receiving the data after the driving test was the most accepted option. In the current driving test, the examiners give their verdict only a few minutes after the driving part. Analyzing the data might require extra time, which means a change in the test setup may be needed.

> *"If data were to be used to make a decision, ... then maybe a moment will have to be inserted just after the actual driving where I retreat for a few minutes. I am going to deliberate with myself and with the data I have, to see: 'What do I actually think of that?'"* (P36, Translated)

How should data be delivered?

Regarding hardware, the examiners explained that they already use tablets and that these could be very suitable for receiving and presenting data. Receiving information in an auditory way, via headphones, was also mentioned.

According to the examiners, data could take the form of color-coded results or percentages. Lack of time to analyze the raw data was often listed as a reason why examiners preferred processed data.

> *"Because time is always an issue. So it must be easily accessible, easy to read."* (P13)

> *"... if you see those seven elements that we grade, and you would have a stoplight principle, and so red, yellow, and green, where you could actually see back, 'These parts are all green. Only this part is red, and this part is orange. Or yellow' as the law requires. You could still retest that yellow part or that red one."* (P38, Translated)

Some examiners preferred less-processed data, yet still easy to read. Graphs were often given as examples.

> "*... I think something that is at least very simple, that you can see at a glance, and I do not know how that would have to be worked out. But yes, graphs, indeed, often then you can see something pretty quickly without looking at numbers in great detail. I think that might be useful.*" (P28, Translated)

However, a few examiners preferred having access to a large amount of information, insisting that it is their task, not the computer's task, to analyze the information.

The examiners further indicated that having access to single variables only (such as presented in Figure 2.2) is not particularly helpful, as single variables lack the necessary context. The examiners pointed out that this problem could be resolved, in part, by presenting a combination of data types or through additional information (e.g., location, traffic density, video).

> "*I would like to see the combination, with a map, so the driven route. ... Combination with data you have, deceleration, distance, so that you get a piece of video.*" (P25, Translated)

> "*Whether the candidates ... should fail or pass should be our judgment. I think graphics would help. So regarding the G forces or brake and acceleration forces and the video fragments. So I think it is a combination of that.*" (P39)

With whom should data be shared?
The possibility of data sharing was discussed during the interviews. Sharing with the candidate, the instructor, other examiners, or the testing organization were proposed. The examiners were generally in favor of data sharing, especially with the candidate. Out of the 35 examiners who mentioned the possibility of sharing data with the candidate, 34 were in favor. The main arguments were that the data belongs to the candidate and that data could have educational purposes.

> "*I think anyway that if I am allowed to see it, that the candidate should be able to see it as well, because it is his behavior, his exam ...*" (P2, Translated)

> "*Well, of course we all do it for traffic safety, so it makes little sense to only share that [data] with me. Of course [it is] very useful … for the candidate to take note [of the data] and learn from it.*" (P31, Translated)

When discussing the possibility of sharing data with other stakeholders, some privacy issues were raised, and sharing was only considered viable if the data were made anonymous.

> "*I would not have a problem with that [sharing], as long as it does not haunt the person themselves. ... because it is a snapshot, people are very vulnerable. And then it would not be fair.*" (P22, Translated)

The interviewers suggested that, once anonymized, data could be used to improve uniformity across driving tests, for example, by identifying discrepancies between examiners and to train them. This type of data use was generally agreed upon, although some examiners expressed concerns regarding potential misuse and an increasing number of complaints. The examiners also noted that it is hard to compare individual driving tests in the attempt to achieve uniform assessment criteria.

2.3.3 Examiners' views about the future of driver testing

The interviewers asked questions regarding the future of driver testing. This topic was first addressed with an open question, followed by more specific questions regarding the possibility of automating some parts of the driving test, the use of artificial intelligence to assess candidates, and the use of simulators.

The future of driver testing

The examiners typically mentioned technological developments, such as increased ADAS usage, the increasing presence of electric cars, as well as the replacement of the manual gearbox by an automatic one.

> *"... we should start using more of the assistance systems that are available in cars."* (P4)

> *"I expect that ... in about ten years, we will have moved to a driving test with automatic transmission."* (P15, Translated)

The main topic of the interview, the data-supported driving test, was often repeated when asked about the future of the driving test.

> *"Indeed, I think the research that you are doing is very positive and that it is indeed moving in that direction that it is becoming data-driven."* (P6, Translated)

Most examiners were not worried about being replaced by a computer.

> *"I am not so afraid about that [being replaced]. [Airplane] pilots are also still needed ... I think it is just a shift. You may then need fewer examiners. But something else will take its place. No, I am not afraid of this at all."* (P19, Translated)

> *"There are colleagues who are a bit afraid to lose their jobs. But I do not believe it ... I think [the systems] will be more supportive systems that can help us establish a verdict."* (P24, Translated)

Automating parts of the driving test

When asked what parts of the driving test could be automated, examiners had difficulty coming up with suggestions. Occasionally, they mentioned parts that they found time-consuming, such as waiting for the participant to enter a destination in the navigation system. However, no noteworthy suggestions for automation were provided.

Driver assessment by artificial intelligence

The interviewers asked if it would be possible, in the future, to have artificial intelligence assessing candidates partially or completely. Opinions were mixed, but a common ground for all examiners was that they would be difficult to replace.

> *"... What we actually test of course is whether they have traffic insight, and that word is so elusive, because what is it? But if that were possible [to measure traffic insight], I think it should be possible too [to replace the human examiner]."* (P2, Translated)

> *"... you [would] have to collect a huge amount of data, because it is not only the driver and what happens in the car. It is also about the whole environment. Road safety has to do with everything that happens on the road. Everything in his head, in his behavior and in his actions. That is very complex."* (P4)

> *"It is always dependent on situations and conditions, and that is the human factor that we add. If someone happens to drive 60 [kph] once, where 50 is allowed, but it is necessary that that speed is driven for a while, … [this does not automatically mean] a failed test."* (P31, Translated)

While it was agreed upon that it would be complex to achieve complete driver assessment by AI, opinions were divided on whether it would be possible or not. Sixteen examiners said it would one day become possible, mentioning a time range between 15 and 40 years.

> *"I think that it is possible, but way in the future. In, like, 40 years or something like that, to do it only with artificial intelligence. Today, no."* (P30)

> *"We need a lot of development for that ... I think maybe we are 30 years down the road before we can really start to trust the system because it all looks really nice, but there are so many uncertain factors in everything."* (P15, Translated)

On the other hand, eight examiners did not believe it would be feasible, mentioning typically that a computer cannot predict or assess everything.

> *"Making mistakes is human, but the computer systems say 'Yes: correct; this and that went wrong'. Such a computer is, of course, much more black-and-white than we are."* (P37, Translated)

> *"Feeling also plays a lot of a role. It just does. And you have those gray cases ... and pure data will not be able to make that distinction."* (P11, Translated)

Future role of simulators

Examiners' responses to the question *"Do you think simulators can play a role in the exam?"* were mixed. Out of the 37 examiners, 19 saw opportunities for simulators in testing (e.g., *"I surely think so."*; P26, Translated), and 15 examiners were negative about this (*"Absolutely not."*; P10). Two examiners were ambivalent, and one provided no clear answer.

The examiners saw the benefit of simulators to preselect candidates by testing the basics, to lower the influx of students with poor basic skills. It was noted that simulators cannot fully replace the driving test but that parts of the driving test could be done in the simulator.

> "... *if in a simulator it turns out that somebody really misses all kinds of things, then you do not even have to go to the driving test.*" (P1, Translated)

> "*Basic things can already be … [tested]. I think you can capture 80% reasonably well on a simulator.*" (P7, Translated)

> "*For example, [you could simulate] a narrow street and there are cars parked on the left and the right side. You can make these standard situations ... and see how our candidates react ...*" (P4)

A frequently heard argument in favor of introducing simulators in the driving test was that uniform situations can be tested. Currently, outside rush hours, a candidate may pass multiple intersections without any other traffic. Standardized testing may contribute to the fairness of the driving test across districts or times.

> "... *you can more easily present the same situations to people, that allows you to measure more fairly candidate-to-candidate …*" (P2, Translated)

> "*You try choosing your route to test all aspects, but sometimes certain situations will not occur. And then I fantasize about simulators, like: 'I would like to have a car coming from the right, now'...*" (P7, Translated)

Examiners noted that limiting factors are simulator sickness and low realism.

> "*When you are already used to driving in a car, and you go to the simulator, you get really nauseous. And you get really sick, and you are not able to drive like you should ...*" (P10)

> "*You feel nothing, you hear nothing. Yes, you sit still, you do not move; it is very different.*" (P37, Translated)

Negative replies to simulator testing were sometimes followed up by mentioning that there could be a greater role for simulators in driver training. Out of the 18 examiners who were not positive about simulator testing, 10 saw opportunities for training drivers in simulators.

> "*I think it is a very good educational tool. I do not think that it is a useful thing from an assessment point of view.*" (P15, Translated)

2.4. Discussion

This study aimed to assess the views of driving examiners for newly licensed drivers about using data as part of the driving test. The interviews started with questions about the current driving test and the factors that examiners take into consideration when coming to a pass/fail verdict. Subsequently, the interviews went into detail about the why, what, and how of data-driven assessment by discussing examples of presentation and delivery modes of the test results. Simulator-based testing, and offline use of performance data, such as sharing with other examiners, were also addressed in the interviews.

According to the examiners, an important advantage, and a source of job satisfaction of the current driving test is that examiners have a certain freedom to arrive at a holistic assessment of the candidate's capabilities. For example, the examiner can guide the candidate along an alternative route if the examiner believes that a driving task requires re-assessment. Furthermore, examiners are not obliged to fail an overall competent candidate who made benign errors (CBR, 2020). In the same vein, a stated advantage of the driving test was that the candidates are expected to show independence, for example, by driving to a particular destination themselves with the help of a route navigation device. These characteristics of the driving test correspond to the Goals for Driver Education, which were created some 20 years ago (Hatakka et al., 2002; Keskinen, 2007) and which are increasingly embedded in driver training and testing worldwide (e.g., Alger & Sundström, 2013; Molina et al., 2014; Rodwell et al., 2018; Senserrick et al., 2017). This trend is in line with research showing that safe driving is not attributed to vehicle-handling and maneuvering skills; rather, higher-order skills, such as choosing the appropriate route, insight, and self-reflection, are regarded as essential determinants of safe driving (Gregersen, 1995; Isler et al., 2011; Watson-Brown et al., 2019).

It may be hard for a computer to assess a candidate's higher-order driving skills, for the same reason that automated vehicles have difficulty understanding traffic context and predicting what other road users will do (Rudenko et al., 2020; Vinkhuyzen & Cefkin, 2016). Using Endsley's (1995) terminology: computers may excel at low-level situation awareness (i.e., perception via sensors) but have difficulty achieving high-level situation awareness (comprehension, anticipation of the traffic situation). Consistent with this viewpoint, the examiners pointed out that they cannot rely solely on data for obtaining an overall picture of the candidate. Reliance on data, in a sense, goes against the holistic approach examiners tend to have nowadays. According to the examiners, data should only be used as an aid and should be interpreted in context, for example, by relating the data to surrounding traffic or by combining the data with geographical and real-time traffic information obtained via connected smart-mobility applications (and see Roemer, 2021; and Vissers & Tsapi, 2020, who recommend the integration of smart mobility in the curriculum).

The examiners saw value in measuring proximity to other road users and driving speed (4.4 and 3.9, respectively, on a scale of 1 to 5), which are critical components of safe driving (SWOV, 2012). The measurement of eye movements was considered important as well, while assessments of driver posture or eco-driving were regarded as of less importance. Poor eco-driving is not a reason for failing a candidate and can often be noticed directly from engine sound or dashboard readings without needing supplementary data. However, the interviews made clear that there are limits to what a driving test can test (be it a data-driven test or not), an observation consistent with the literature. For example, candidates may be susceptible to the looked-but-failed-to-see phenomenon (Herslund & Jørgensen, 2003), make errors because they are nervous (Fairclough et al., 2006), or show rule compliance during the test but reveal themselves as risk-takers once licensed

(Baughan et al., 2005).

The interviews showed that examiners are under considerable time pressure and have little time to assess the candidate's driving ability. Of note, as early as 1992, Meijman et al. assessed the workload of Dutch driving examiners and concluded that *"the examiners' job must be characterized as a high stress job"* (pp. 255–256), based on which a recommendation was adopted to reduce the number of driving tests per examiner per day (for a similar study, see Parkes, 1995). Workload and shortage of time were important factors for examiners to accept or reject certain forms of data-driven support. For example, it became clear that, apart from direct warning signals, there is little opportunity to process data during the driving test, as examiners are busy monitoring safety, giving instructions, and making sure the candidate is at ease. Also after the driving test, only little time is available, and hence data would need to be available in a straightforward format, or the test structure would have to be changed. At the same time, the examiners emphasized the need to have transparent access to the raw data or graphs, since it is the examiner's role to explain how a verdict is reached. Of note, in France, the test outcome is communicated a few days after the exam (Sécurité Routière, 2021), an approach that would allow the examiners to spend more time analyzing the data before formulating the verdict.

The above findings can be related to the levels of automation proposed by Sheridan (1992, p. 358). On a scale of 1 *"the computer offers no assistance, human must do it all" to 10 "the computer decides everything and acts autonomously, ignoring the human"*, the examiners would accept Level 2 or 3 at maximum: *"the computer offers a complete set of action alternatives" and "narrows the selection down to a few"*. In other words, the examiners were favorable towards having access to computer-generated material such as graphs or scores but would not want higher levels of support. Indeed, perhaps the most striking result from the interviews was that the examiners did not want the computer to make the pass/fail decisions for them. Instead, they wanted to use data and video material to clarify and justify their verdict or rule out doubts about the candidate's viewing behavior, headway to the car in front, speed, or braking behavior. This includes the use of data to convince candidates who strongly disagree with the examiner's verdict and who, in some cases, display aggression towards examiners, a problem also noted by others (Foxe, 2020; Roemer, 2021). In a way, the proposed use of data resembles how police patrol uses speed measurements and cameras to show offenders that they violated the traffic rules (Young & Regan, 2007), a concept referred to by some examiners. Body cams have been proposed for the UK driving test to curb violence against examiners (GOV.UK, 2017). Whether body cams are effective in reducing assaults or the number of complaints filed is an ongoing topic of debate (Ariel et al., 2018; Lum et al., 2019). Some examiners expressed concerns that if data were to be shared for evaluation purposes, this could cause an increase in formal complaints.

The scientific literature offers various ideas for introducing higher levels of automation into the driving test. In Fridman et al. (2019), deviant driving was automatically flagged by a computer and then passed to a human supervisor. Such a concept would correspond to Level 5 automation according to Sheridan's (1992) ten-level taxonomy: *"the computer … suggests one [action alternative], and executes that suggestion if the human approves"*.

The notion of event-triggered data- and camera-based monitoring as part of graduated driver licensing (GDL) has been discussed and studied extensively in the literature (Baker et al., 2020; Klauer et al., 2016; McGehee et al., 2007; Williams & Shults, 2010). Even higher levels of automation are possible, such as in Dubai, where the verdict is supposedly made by a computer with human involvement (e.g., Level 7: *"the computer …executes automatically, then necessarily informs the human"*, albeit in a controlled driving range (Government of Dubai, 2019). From the interviews, it became clear that the examiners were hesitant and skeptical about fully automated assessments, noting that computers are unable to make a holistic assessment. The examiners did recognize, however, that automated driver assessments may have a role in specific subtasks, such as special maneuvers or acceleration behavior. In summary, the examiners appeared to be open-minded about the use of data in their current job (Level 2 and 3 automation), while the notion of a fully automated driving test was regarded as unfeasible for the coming decades.

The present interview study concerned the use of in-vehicle technology for assessing driver behavior. A related topic, assessing how drivers use in-vehicle technology, becomes of increasing interest to licensing organizations as well. How drivers of different experience levels change gear was once a topic of considerable academic interest (e.g., Duncan et al., 1991; Shinar et al., 1998), but with the growing popularity of automated gearboxes and electric cars, this component of expertise may disappear, as pointed out by some examiners. Furthermore, newly sold cars contain various ADAS, including blind-spot warning, forward collision warning, adaptive cruise control, lane assist, or other forms of shared control (Oviedo-Trespalacios et al., 2021; Ziebinski et al., 2017). Driving instructors and licensing organizations face growing challenges regarding the training and testing of drivers' interaction with ADAS and automated driving systems (Heikoop et al., 2020; Sturzbecher et al., 2015; Van den Beukel et al., 2021). An increasing body of research aims to examine which training methods are suitable for learning how to interact with assisted and automated driving technology (Ebnali et al., 2019; Manser et al., 2019; Merriman et al., 2021; Noble et al., 2019; Payre et al., 2017; Shaw et al., 2020). For several years in the Netherlands, it has been permitted to use ADAS in the driving test (Claesen, 2018), but according to a questionnaire study among driving instructors and examiners, driver assessment of ADAS use is not yet incorporated in driver training and testing in a structured manner (Vlakveld & Wesseling, 2018).

One of the challenges in using ADAS in the driving test is that ADAS availability differs between vehicle models and that different ADAS have different purposes (e.g., comfort/luxury option vs. safety benefits; Tsapi, 2015; Vlakveld & Wesseling, 2018). Similar challenges can be expected in future data-supported driver testing, as variability in vehicles and sensors may compromise the fairness of the assessment. Therefore, attention must be paid to standardization and legislation of data-driven assessment technologies. Regarding legislation, while in the UK, for example, it is not allowed for candidates to record audio or video during the driving test (GOV.UK, 2021), while Poland (Kamiński et al., 2008) and Pakistan (Government of Pakistan, 2019) are reported to video-record their driving tests. The examiners argued that the data should be made accessible to the candidates and, provided that privacy is properly taken into consideration, were in favor of sharing data with their employer to improve the quality and uniformity of the driving test. Of note, the Dutch Central Office of Driving Certification already adheres to some open data principles by making the pass rates of all driving schools and examination locations available (CBR, 2021).

A limitation of this study is that it is possible that the examiners' responses were influenced by the familiarity heuristic (Metcalfe et al., 1993). The examiners may have brought up particular possibilities of data-driven assessment because they encountered similar technology in their job (examiners are often seated in modern vehicles and appeared very knowledgeable about ADAS) and may have had difficulty envisioning new ways of data use. In particular, the data type most frequently suggested by the examiners were eye-gaze measurements (Table 2.2). This may legitimately be a crucial element, as incorrect viewing behavior is a common reason for failing the test (De Winter et al., 2008b). However, many examiners had recently received on-the-job training on eye movements, which may also explain why they brought up this topic. Similarly, the examiners' views about simulators may have been shaped by the fact that simulators are used in driver training in the Netherlands for many years already (De Winter et al., 2019; Kappé & Van Emmerik, 2005) and are a well-known topic of discussion (United States: Allen et al., 2010; Australia: Rodwell et al., 2019; Norway: Sætren et al., 2018). The examiners regarded simulators as promising for training and screening in standardized conditions but not as a suitable full replacement of the current driving test. Simulator fidelity and simulator sickness in some drivers remain bottlenecks in the acceptance of simulators (De Winter et al., 2012; Kappé & Van Emmerik, 2005).

Another limitation is that the present interviews were conducted with a specific sample: all participants were driving examiners and of Dutch nationality. Future research should assess the views of other participants, such as test candidates (e.g., novice drivers, but also older drivers and professional drivers) as well as examiners and candidates from other nationalities. The Netherlands is a country that adopts a test-led model without obligatory driver training modules and without fixed routes of the driving test. Although the examiners were open-minded about the use of data before and after the driving test, the scope of this research was limited to the driving test itself. Differences in the setup of driver training and testing between countries (e.g., training-led models, multi-phase models) can be expected to lead to different opinions about data use (for overviews of national differences, see Genschow et al., 2014; Helman et al., 2017). Research on other possible data uses, such as whether data could support lifelong learning, should still be performed.

2.5. Conclusion

Cars are becoming 'computers on wheels', and an increasing number of mobile devices are available that produce driving-related data. These developments raise the question of whether data-driven assessments could have a role in formal driver testing. Interviews were conducted with 37 driving examiners from all testing regions in the Netherlands. The interviews examined if and why examiners would like to use data and what data format would be most useful for them.

It is concluded that examiners are positive about receiving data in the driving test, especially if the data could help them explain their verdict to the candidate. Frequently suggested data types were recordings of the candidates' eye movements and data that describe the car's state in relation to its surroundings, such as speed relative to traffic, distance to surroundings, and position on the road. Examiners were also positive about the use of video fragments, flagged at critical situations. Data should be presented in an easily accessible format, allowing the examiner to obtain an overview in the limited time available between the driving test and the presentation of the verdict. Another key finding was that examiners emphasized the human element in testing drivers and the importance of establishing an overall picture of the candidate.

Our observations are relevant in the context of recently published recommendations stating that the Dutch driving education system needs a fundamental overhaul from a test-led system towards a test- and education-driven system (Roemer, 2021; see Helman et al., 2017 for similar recommendations in a European perspective). For example, it has been recommended that the Netherlands should introduce a modular curriculum and a student monitoring system. The same report recommends conducting experiments with instrumented vehicles to take steps towards a more competency-based assessment (Roemer, 2021). It is expected that the current interview study provides a suitable basis for determining what type of data-driven technology could be used in this experimental phase. Finally, there is a need for knowledge on data-driven assessment in a broader perspective. Future interview studies and experiments could be performed as part of an international consortium that takes into account other target groups, such as truck drivers, as well.

References

2todrive. (2021). Begeleid rijden vanaf je 17e [Supervised driving from your 17th]. https://2todrive.nl

Ahmad, F. H., Zhang, Y., & Qiao, F. (2021). Comparative study of dashcam-based vehicle incident detection techniques. Proceedings of the 28th Conference of FRUCT Association (pp. 539–544). Moscow, Russia. https://doi.org/10.5281/zenodo.4514947

Alger, S., & Sundström, A. (2013). Agreement of driving examiners' assessments–Evaluating the reliability of the Swedish driving test. Transportation Research Part F: Traffic Psychology and Behaviour, 19, 22–30. https://doi.org/10.1016/j.trf.2013.02.004

Allen, R. W., Park, G. D., & Cook, M. L. (2010). Simulator fidelity and validity in a transfer-of-training context. Transportation Research Record: Journal of the Transportation Research Board, 2185, 40–47. https://doi.org/10.3141/2185-06

AmberScript. (2021). Transform your audio and video to text and subtitles. https://www.amberscript.com

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., Megicks, S., & Henderson, R. (2018). Paradoxical effects of self-awareness of being observed: Testing the effect of police body-worn cameras on assaults and aggression against officers. Journal of Experimental Criminology, 14, 19–47. https://doi.org/10.1007/s11292-017-9311-5

Arumugam, S., & Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. Journal of Big Data, 6, 86. https://doi.org/10.1186/s40537-019-0249-5

Baker, S. A., Klauer, C., Baynes, P., & Dingus, T. A. (2020). Parent perceptions of real-time and non-video feedback from the Driver Coach Study (Report No. 20-UM-082). Blacksburg, VA: National Surface Transportation Safety Center for Excellence.

Baughan, C., & Simpson, B. (1999). Consistency of driving performance at the time of the L-test, and implications for driver testing. In G. B. Grayson (Ed.), Behavioural research in road safety IX (pp. 206–214). Crowthorne: Transport Research Laboratory.

Baughan, C. J., Gregersen, N. P., Hendrix, M., & Keskinen, E. (2005). Towards European standards for testing (Final Report). Brussels: The International Commission for Driving Testing. http://www.cieca.eu/sites/default/files/documents/projects_and_studies/EU_TEST_Project_Final_Report.pdf

Bergasa, L. M., Almería, D., Almazán, J., Yebes, J. J., & Arroyo, R. (2014). Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. 2014 IEEE Intelligent Vehicles Symposium Proceedings

(pp. 240–245). Dearborn, MI. https://doi.org/10.1109/IVS.2014.6856461
Bjørnskau, T. (2003). Stryk eller stå. En undersøkelse av faktorer som påvirker resultatene av praktisk førerprøve [Pass or fail. A study of factors that affect the results of the driving test] (TØI report No. 662/2003). Oslo: Department of Transport Economics.
Boelhouwer, A., Van den Beukel, A. P., Van der Voort, M. C., Verwey, W. B., & Martens, M. H. (2020). Supporting drivers of partially automated cars through an adaptive digital in-car tutor. Information, 11, 185. https://doi.org/10.3390/info11040185
Boggs, A. M., Arvin, R., & Khattak, A. J. (2020). Exploring the who, what, when, where, and why of automated vehicle disengagements. Accident Analysis & Prevention, 136, 105406. https://doi.org/10.1016/j.aap.2019.105406
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3, 77–101. https://doi.org/10.1191/1478088706qp063oa
CBR. (2020). Rijprocedure B [Driving procedure B]. https://www.cbr.nl/nl/voor-rijscholen/nl/partner-in-verkeersveiligheid-1/rijprocedure.htm
CBR. (2021). Open data. https://www.cbr.nl/nl/over-het-cbr/over/cbr-in-cijfers/open-data.htm
Chen, B., Zhao, D., & Peng, H. (2017). Evaluation of automated vehicles encountering pedestrians at unsignalized crossings. Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (pp. 1679–1685). Los Angeles, CA. https://doi.org/10.1109/IVS.2017.7995950
Chowdhury, A., Shankaran, R., Kavakli, M., & Haque, M. M. (2018). Sensor applications and physiological features in drivers' drowsiness detection: A review. IEEE Sensors Journal, 18, 3055–3067. https://doi.org/10.1109/JSEN.2018.2807245
Claesen, R. (2018). Automated driving. The consequences for future driver training and testing. Zeitschrift für Verkehrssicherheit, 64, 98–99.
Crinson, L. F., & Grayson, G. B. (2005). Profile of the British learner driver In G. Underwood (Ed.), Traffic and transport psychology, theory and application (pp. 157–170). London: Elsevier.
De Waard, D., & Brookhuis, K. A. (1999). Driver support and automated driving systems: Acceptance and effects on behavior. In M. W. Scerbo & M. Mouloua (Eds.), Automation technology and human performance: current research and trends (pp. 49–57). Norfolk, VA: Old Dominion University Research Foundation.
De Winter, J. C. F. (2013). Predicting self-reported violations among novice license drivers using pre-license simulator measures. Accident Analysis & Prevention, 52, 71–79. https://doi.org/10.1016/j.aap.2012.12.018
De Winter, J. C. F., De Groot, S., Dankelman, J., Wieringa, P. A., Van Paassen, M. M., & Mulder, M. (2008a). Advancing simulation-based driver training: lessons learned and future perspectives. Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 459–464). Amsterdam, The Netherlands.
De Winter, J. C. F., De Groot, S., Van Loenhout, M. J., Van Leeuwen, A., Do, P., Wieringa, P. A., & Mulder, M. (2008b). Feedback on mirror-checking during simulation-based driver training. In J. C. F. de Winter, Advancing simulation-based driver training [Doctoral dissertation] (pp. 153–161). Delft, The Netherlands: Delft University of Technology. http://resolver.tudelft.nl/uuid:d5f04e70-37aa-4a53-8e8d-af17cc402d1e
De Winter, J. C. F., Dodou, D., Happee, R., & Eisma, Y. B. (2019). Will vehicle data be shared to address the how, where, and who of traffic accidents? European Journal of Futures Research, 7. https://doi.org/10.1186/s40309-019-0154-3
De Winter, J. C. F., & Kovácsová, N. (2016). How science informs engineering, education, and enforcement: A message for driving instructors. In D. L. Fisher, J. K. Caird, W. J. Horrey, & L. M. Trick (Eds.), Handbook of teen and novice drivers: research, practice, policy, and directions (pp. 31–45). Boca Raton, FL: Taylor & Francis.
De Winter, J. C. F., Van Leeuwen, P. M., & Happee, R. (2012). Advantages and disadvantages of driving simulators: a discussion. In A. J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L. P. J. J. Noldus, & P. H. Zimmerman (Eds.), Proceedings of Measuring Behavior 2012, 8th International Conference on Methods and Techniques in Behavioral Research (pp. 47–50). Wageningen: Noldus Information Technology B.V..
Duncan, J., Williams, P., & Brown, I. (1991). Components of driving skill: experience does not mean expertise. Ergonomics, 34, 919–937. https://doi.org/10.1080/00140139108964835
Ebnali, M., Hulme, K., Ebnali-Heidari, A., & Mazloumi, A. (2019). How does training effect users' attitudes and skills needed for highly automated driving? Transportation Research Part F: Traffic Psychology and Behaviour, 66, 184–195. https://doi.org/10.1016/j.trf.2019.09.001
Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37, 32–64. https://doi.org/10.1518/001872095779049543
Fairclough, S. H., Tattersall, A. J., & Houston, K. (2006). Anxiety and performance in the British driving test. Transportation Research Part F: Traffic Psychology and Behaviour, 9, 43–52. https://doi.org/10.1016/j.trf.2005.08.004
Farag, W. A. (2020). Model-predictive-control complex-path tracking for self-driving cars. International Journal of Modelling, Identification and Control, 34, 265–277. https://doi.org/10.1504/IJMIC.2020.111624
Farah, H., Musicant, O., Shimshoni, Y., Toledo, T., Grimberg, E., Omer, H., & Lotan, T. (2014). Can providing feedback on driving behavior and training on parental vigilant care affect male teen drivers and their parents?

Accident Analysis & Prevention, 69, 62–70. https://doi.org/10.1016/j.aap.2013.11.005
Foxe, K. (2020). More than 50 incidents of aggression by driving test candidates reported. https://www.irishexaminer.com/news/arid-40195446.html
Fridman, L., Ding, L., Jenik, B., & Reimer, B. (2019). Arguing machines: Human supervision of black box AI systems that make life-critical decisions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, CA. https://doi.org/10.1109/CVPRW.2019.00173
Genschow, J., Sturzbecher, D., & Willmes-Lenz, G. E. (2014). Novice driver preparationan international comparison (Berichte der Bundesanstalt für Straßenwesen. Mensch und Sicherheit Heft M 234 b). Bergisch Gladbach, Germany: Federal Highway Research Institute.
Giridharan, V. (2019). How Microsoft has quietly started automating driving tests in India using AI. https://www.digit.in/news/car-tech/how-microsoft-has-quietly-started-automating-driving-tests-in-india-using-ai-50854.html
González, D., Pérez, J., Milanés, V., & Nashashibi, F. (2015). A review of motion planning techniques for automated vehicles. IEEE Transactions on Intelligent Transportation Systems, 17, 1135–1145. https://doi.org/10.1109/TITS.2015.2498841
Goodall, N. J. (2021). Comparison of automated vehicle struck-from-behind crash rates with national rates using naturalistic data. Accident Analysis & Prevention, 154, 106056. https://doi.org/10.1016/j.aap.2021.106056
Government of Dubai. (2019). RTA announces full automation of drivers testing at the smart yard [Press release]. https://www.rta.ae/wps/portal/rta/ae/home/news-and-media/all-news/NewsDetails/rta+-announces-full-automatio-of-drivers-testing-at-the-smart+yard
Government of Pakistan. (2019). National guidelines for driver licensing. Islamabad: Ministry of Communications. http://www.roadsafetypakistan.pk/download/national-guidelines-for-driver-licensing.pdf
GOV.UK. (2017). DVSA aims to stop violence against its staff. https://www.gov.uk/government/news/dvsa-aims-to-stop-violence-against-its-staff
GOV.UK. (2021). Filming or recording driving tests: what you're allowed to do. https://www.gov.uk/government/publications/filming-or-recording-driving-tests/filming-or-recording-driving-tests
Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education, 48, 612–618. https://doi.org/10.1109/TE.2005.856149
Gregersen, N. P. (1995). What should be taught? Basic vehicle control skills or higher order skills? In H. S. Simpson (Ed.), New to the road: reducing the risks for young motorists. Proceedings of the First Annual International Conference of the Youth Enhancement Service (pp. 103–114). Los Angeles, CA.
Händel, P., Ohlsson, J., Ohlsson, M., Skog, I., & Nygren, E. (2013). Smartphone-based measurement systems for road vehicle traffic monitoring and usage-based insurance. IEEE Systems Journal, 8, 1238–1248. https://doi.org/10.1109/JSYST.2013.2292721
Hatakka, M., Keskinen, E., Gregersen, N. P., Glad, A., & Hernetkoski, K. (2002). From control of the vehicle to personal self-control; broadening the perspectives to driver education. Transportation Research Part F: Traffic Psychology and Behaviour, 5, 201–215. https://doi.org/10.1016/S1369-8478(02)00018-9
Heikoop, D. D., Calvert, S. C., Mecacci, G., & Hagenzieker, M. P. (2020). A practitioner's view of driver training for automated driving from driving examiners: A focus group discussion. 2020 Forum on Integrated and Sustainable Transportation Systems (pp. 14–19). Delft, The Netherlands. https://doi.org/10.1109/FISTS46898.2020.9264869
Helman, S., Vlakveld, W., Fildes, B., Oxley, J., Fernández-Medina, K., & Weekley, J. (2017). Study on driver training, testing and medical fitness (Final report). Retrieved from https://ec.europa.eu/transport/road_safety/sites/default/files/dl_study_on_training_testing_med_fitness.pdf
Herslund, M. B., & Jørgensen, N. O. (2003). Looked-but-failed-to-see-errors in traffic. Accident Analysis & Prevention, 35, 885–891. https://doi.org/10.1016/S0001-4575(02)00095-7
Hickman, J. S., & Hanowski, R. J. (2011). Use of a video monitoring approach to reduce at-risk driving behaviors in commercial vehicle operations. Transportation Research Part F: Traffic Psychology and Behaviour, 14, 189–198. https://doi.org/10.1016/j.trf.2010.11.010
Isler, R. B., Starkey, N. J., & Sheppard, P. (2011). Effects of higher-order driving skill training on young, inexperienced drivers' on-road driving performance. Accident Analysis & Prevention, 43, 1818–1827. https://doi.org/10.1016/j.aap.2011.04.017
Kamiński, T., Wojciechowski, A., Nowacki, G., Mitraszewska, I., Kamińska, E., & Filipek, P. (2008). Video register picture evaluation to support drivers training and examination. Journal of KONES Powertrain and Transport, 15, 225–232. https://www.infona.pl//resource/bwmeta1.element.baztech-article-BUJ5-0034-0027
Kappé, B., & Van Emmerik, M. L. (2005). Mogelijkheden van rijsimulatoren in de rijopleiding en het rijexamen [The possibilities of driving simulators for initial driver training and testing] (Report No. TNO-DV3 2005 C114). Soesterberg: TNO Defense and Security.
Karvonen, H., Kujala, T., & Saariluoma, P. (2006). In-car ubiquitous computing: driver tutoring messages presented on a head-up display. Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference (pp. 560–565). Toronto, Ontario. https://doi.org/10.1109/ITSC.2006.1706800

Kashevnik, A., Lashkov, I., & Gurtov, A. (2019). Methodology and mobile application for driver behavior analysis and accident prevention. IEEE Transactions on Intelligent Transportation Systems, 21, 2427–2436. https://doi.org/10.1109/TITS.2019.2918328

Keskinen, E. (2007). What is GDE all about and what it is not. In W. Henriksson, T. Stenlund, A. Sundstrom, & M. Wiberg (Eds.), Proceedings from The GDE-Model as a Guide in driver training and testing (pp. 3–13). Umea, Sweden: Umea University.

Klauer, S. G., Ehsani, J., & Simons-Morton, B. (2016). Using naturalistic driving methods to study novice drivers. In S. G. Klauer, J. Ehsani, & B. Simons-Morton (Eds.), Handbook of teen and novice drivers (pp. 429–440). CRC Press.

Lechner, G., Fellmann, M., Festl, A., Kaiser, C., Kalayci, T. E., Spitzer, M., & Stocker, A. (2019). A lightweight framework for multi-device integration and multi-sensor fusion to explore driver distraction. In P. Giorgini & B. Weber B. (Eds.), Advanced Information Systems Engineering. CAiSE 2019 (pp. 80–95). Cham: Springer. https://doi.org/10.1007/978-3-030-21290-2_6

Lima, P. F., Pereira, G. C., Mårtensson, J., & Wahlberg, B. (2018). Experimental validation of model predictive control stability for autonomous driving. Control Engineering Practice, 81, 244–255. https://doi.org/10.1016/j.conengprac.2018.09.021

Litman, T. (2021). Autonomous vehicle implementation predictions. Implications for transport planning. Victoria, British Columbia: Victoria Transport Policy Institute.

Lum, C., Stoltz, M., Koper, C. S., & Scherer, J. A. (2019). Research on body-worn cameras: What we know, what we need to know. Criminology & Public Policy, 18, 93–118. https://doi.org/10.1111/1745-9133.12412

Magaña, V. C., & Muñoz-Organero, M. (2015). Artemisa: A personal driving assistant for fuel saving. IEEE Transactions on Mobile Computing, 15, 2437–2451. https://doi.org/10.1109/TMC.2015.2504976

Manser, M. P., Noble, A. M., Machiani, S. G., Shortz, A., Klauer, S. G., Higgins, L., & Ahmadi, A. (2019). Driver training research and guidelines for automated vehicle technology (Final Report No. 01-004). Blacksburg, VA: Virginia Tech Transportation Institute.

Marin-Plaza, P., Hussein, A., Martin, D., & Escalera, A. D. L. (2018). Global and local path planning study in a ROS-based research platform for autonomous vehicles. Journal of Advanced Transportation, 2018, 6392697. https://doi.org/10.1155/2018/6392697

Marti, E., De Miguel, M. A., Garcia, F., & Perez, J. (2019). A review of sensor technologies for perception in automated driving. IEEE Intelligent Transportation Systems Magazine, 11, 94–108. https://doi.org/10.1109/MITS.2019.2907630

Mase, J. M., Agrawal, U., Pekaslan, D., Mesgarpour, M., Chapman, P., Torres, M. T., & Figueredo, G. P. (2020). Capturing uncertainty in heavy goods vehicles driving behaviour. Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems. Rhodes, Greece. https://doi.org/10.1109/ITSC45102.2020.9294378

McGehee, D. V., Raby, M., Carney, C., Lee, J. D., & Reyes, M. L. (2007). Extending parental mentoring using an event-triggered video intervention in rural teen drivers. Journal of Safety Research, 38, 215–227. https://doi.org/10.1016/j.jsr.2007.02.009

Meijman, T. F., Mulder, G., Van Dormolen, M., & Cremer, R. (1992). Workload of driving examiners: A psychophysiological field study. In H. Kragt (Ed.), Enhancing industrial performance (pp. 245–258). London, UK: Taylor & Francis.

Merriman, S. E., Plant, K. L., Revell, K. M., & Stanton, N. A. (2021). Challenges for automated vehicle driver training: a thematic analysis from manual and automated driving. Transportation Research Part F: Traffic Psychology and Behaviour, 76, 238–268. https://doi.org/10.1016/j.trf.2020.10.011

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 851–861. https://doi.org/10.1037/0278-7393.19.4.851

Michelaraki, E., Katrakazas, C., Yannis, G., Filtness, A., Talbot, R., Hancox, G., Pilkington-Cheney, F., Brijs, K., Ross, V., Dirix, H., Neven, A., Paul, R., Brijs, T., Fortsakis, P., Frantzola, E. K., & Taveira, R. (2021). Post-trip safety interventions: state-of-the-art, challenges, and practical implications. Journal of Safety Research, 77, 67–85. https://doi.org/10.1016/j.jsr.2021.02.005

Michon, J. A. (Ed.). (1993). Generic intelligent driver support. London, UK: CRC Press. https://doi.org/10.1201/9781003208952

Microsoft. (2021). HAMS: Harnessing AutoMobiles for Safety. https://www.microsoft.com/en-us/research/project/hams

Molina, J. G., García-Ros, R., & Keskinen, E. (2014). Implementation of the driver training curriculum in Spain: An analysis based on the Goals for Driver Education (GDE) framework. Transportation Research Part F: Traffic Psychology and Behaviour, 26, 28–37. http://dx.doi.org/10.1016/j.trf.2014.06.005

Morey, R. D. (2008). Confidence intervals from normalized data: a correction to cousineau (2005). Tutorials in Quantitative Methods for Psychology, 4, 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Musicant, O., & Lotan, T. (2016). Can novice drivers be motivated to use a smartphone based app that monitors their behavior? Transportation Research Part F: Traffic Psychology and Behaviour, 42, 544–557. https://doi.org/10.1016/j.trf.2015.10.023

Mynttinen, S., Koivukoski, M., Hakuli, K., & Keskinen, E. (2011). Finnish novice drivers' competences–

Successful driving test candidates 2000–2009 evaluated by driving examiners. Transportation Research Part F: Traffic Psychology and Behaviour, 14, 66–75. https://doi.org/10.1016/j.trf.2010.10.001

Nambi, A. U., Bannur, S., Mehta, I., Kalra, H., Virmani, A., Padmanabhan, V. N., Bhandari, R., & Raman, B. (2018). HAMS: Driver and driving monitoring using a smartphone. Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (pp. 840–842). New Delhi, India. https://doi.org/10.1145/3241539.3267723

Noble, A. M., Klauer, S. G., Doerzaph, Z. R., & Manser, M. P. (2019). Driver training for automated vehicle technology–knowledge, behaviors, and perceived familiarity. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63, 2110–2114. https://doi.org/10.1177/1071181319631249

Olweus, D. (1958). Undersökning rörande bilinspektörernas bedömning av de praktiska körkortsproven [Investigation concerning the driving examiners' assessments of practical driving license tests]. Stockholm, Sweden: Institute of Psychotechnics, Stockholm University.

Oviedo-Trespalacios, O., Tichon, J., & Briant, O. (2021). Is a flick-through enough? A content analysis of Advanced Driver Assistance Systems (ADAS) user manuals. PLOS ONE, 16, e0252688. https://doi.org/10.1371/journal.pone.0252688

Parkes, K. R. (1995). The effects of objective workload on cognitive performance in a field setting: A two-period cross-over trial. Applied Cognitive Psychology, 9, S153–S171. https://doi.org/10.1002/acp.2350090710

Payre, W., Cestac, J., Dang, N. T., Vienne, F., & Delhomme, P. (2017). Impact of training and in-vehicle 'task performance on manual control recovery in an automated car. Transportation Research Part F: Traffic Psychology and Behaviour, 46, 216–227. http://dx.doi.org/10.1016/j.trf.2017.02.001

Pugnetti, C., & Elmer, S. (2020). Self-assessment of driving style and the willingness to share personal information. Journal of Risk and Financial Management, 13, 53. https://doi.org/10.3390/jrfm13030053

Ramzan, M., Khan, H. U., Awan, S. M., Ismail, A., Ilyas, M., & Mahmood, A. (2019). A survey on state-of-the-art drowsiness detection techniques. IEEE Access, 7, 61904–61919. https://doi.org/10.1109/ACCESS.2019.2914373

Ranft, B., & Stiller, C. (2016). The role of machine vision for intelligent vehicles. IEEE Transactions on Intelligent Vehicles, 1, 8–19. https://doi.org/10.1109/TIV.2016.2551553

Rangesh, A., & Trivedi, M. M. (2019). No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars. IEEE Transactions on Intelligent Vehicles, 4, 588–599. https://doi.org/10.1109/TIV.2019.2938110

Rodwell, D., Hawkins, A., Haworth, N., Larue, G. S., Bates, L., & Filtness, A. (2018). A mixed-methods study of driver education informed by the Goals for Driver Education: Do young drivers and educators agree on what was taught? Safety Science, 108, 140–148. https://doi.org/10.1016/j.ssci.2018.04.017

Rodwell, D., Hawkins, A., Haworth, N., Larue, G. S., Bates, L., & Filtness, A. (2019). What do driver educators and young drivers think about driving simulators? A qualitative draw-and-talk study. Transportation Research Part F: Traffic Psychology and Behaviour, 62, 282–293. https://doi.org/10.1016/j.trf.2019.01.008

Roemer, E. (2021). Van rijles naar rijonderwijs. Advies verbetering autorijscholenbranche [From driving lessons to driving education. Advice for improving the driving School industry]. The Hague, The Netherlands: Ministry of Infrastructure and Water Management. https://www.rijksoverheid.nl/documenten/rapporten/2021/04/14/bijlage-1-van-rijles-naar-rijonderwijs-advies-verbeteren-autorijscholenbranche

Ropelato, S., Zünd, F., Magnenat, S., Menozzi, M., & Sumner, R. (2018). Adaptive tutoring on a virtual reality driving simulator. International SERIES on Information Systems and Management in Creative EMedia, 2017, 12–17. https://doi.org/10.3929/ethz-b-000195951

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). Human motion trajectory prediction: A survey. The International Journal of Robotics Research, 39, 895–935. https://doi.org/10.1177/0278364920917446

Sætren, G. B., Pedersen, P. A., Robertsen, R., Haukeberg, P., Lindheim, C., & Rasmussen, M. (2018). Simulator training in driver education – potential gains and challenges. In S. Haugen, A. Barros, C. van Gulijk, T. Kongsvik, & J. E. Vinnem (Eds.), Safety and reliability – Safe societies in a changing world (pp. 2045–2049). London, UK: Taylor & Francis Group.

Schoettle, B. (2017). Sensor fusion: A comparison of sensing capabilities of human drivers and highly automated vehicles (Report No. SWT-2017-12). Ann Arbor, MI: University of Michigan.

Schöner, H. P., Pretto, P., Sodnik, J., Kaluza, B., Komavec, M., Varesanovic, D., Chouchane, H., & Antona-Makoshi, J. (2021). A safety score for the assessment of driving style. Traffic Injury Prevention, 22, 384–389. https://doi.org/10.1080/15389588.2021.1904508

Sécurité Routière. (2021). Résultats du permis de conduire [Driving license results]. https://www.securite-routiere.gouv.fr/resultats-du-permis-de-conduire#/step-connexion

Segers, J. (2014). Analysis techniques for racecar data acquisition. Warrendale, PA: SAE International.

Senserrick, T., McRae, D., Rome, L. D., Rees, P., & Williamson, A. (2017). Enhancing higher-order skills education and assessment in a graduated motorcycle licensing system. Safety, 3, 14. https://doi.org/10.3390/safety3020014

Shanly, C., Ieti, M., Warren, I., & Sun, J. (2018). BackPocketDriver – A mobile app to enhance safe driving for youth. Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering (pp. 246–245). Redwood City, CA. https://doi.org/10.18293/SEKE2018-011

Shaw, E., Large, D., & Burnett, G. (2020). Driver training for future automated vehicles: Introducing CHAT (CHeck, Assess, Takeover). Nottingham, UK: University of Nottingham.

Sheridan, T. B. (1992). Telerobotics, automation, and human supervisory control. Cambridge, MA: MIT Press.

Shimshoni, Y., Farah, H., Lotan, T., Grimberg, E., Dritter, O., Musicant, O., Toledo, T., & Omer, H. (2015). Effects of parental vigilant care and feedback on novice driver risk. Journal of Adolescence, 38, 69–80. https://doi.org/10.1016/j.adolescence.2014.11.002

Shinar, D., Meir, M., & Ben-Shoham, I. (1998). How automatic is manual gear shifting? Human Factors, 40, 647–654. https://doi.org/10.1518/001872098779649346

Shladover, S. E. (2016). The truth about "self-driving" cars. Scientific American, 314, 52–57. https://doi.org/10.1038/scientificamerican0616-52

Sikander, G., & Anwar, S. (2018). Driver fatigue detection systems: A review. IEEE Transactions on Intelligent Transportation Systems, 20, 2339–2352. https://doi.org/10.1109/TITS.2018.2868499

Sturzbecher, D., Mörl, S., & Kaltenbaek, J. (2015). Optimisation of the practical driving test (Berichte der Bundesanstalt für Straßenwesen. Mensch und Sicherheit Heft M 243 b). Potsdam, Germany: University of Potsdam. https://bast.opus.hbz-nrw.de/opus45-bast/frontdoor/deliver/index/docId/1360/file/M243b_en_barrierefrei.pdf

SWOV. (2012). SWOV Fact sheet: Headway times and road safety. https://www.swov.nl/sites/default/files/publicaties/gearchiveerde-factsheet/uk/fs_headway_archived.pdf

SWOV. (2016). SWOV Fact sheet: 18- to 24-year-olds: young drivers. https://www.swov.nl/en/facts-figures/factsheet/18-24-year-olds-young-drivers

Tabone, W., De Winter, J. C. F., Ackermann, C., Bärgman, J., Baumann, M., Deb, S., Emmenegger, C., Habibovic, A., Hagenzieker, M., Hancock, P. A., Happee, R., Krems, J., Lee, J. D., Martens, M., Merat, N., Norman, D. A., Sheridan, T. B., & Stanton, N. A. (2021). Vulnerable road users and the coming wave of automated vehicles: Expert perspectives. Transportation Research Interdisciplinary Perspectives, 9, 100293. https://doi.org/10.1016/j.trip.2020.100293

Terán, J., Navarro, L., Quintero M, C. G., & Pardo, M. (2020). Intelligent driving assistant based on road accident risk map analysis and vehicle telemetry. Sensors, 20, 1763. https://doi.org/10.3390/s20061763

Tsapi, A. (2015). Introducing Advanced Driver Assistance Systems (ADAS) into drivers' training and testing: The young learner drivers' perspective [MSc thesis Delft University of Technology]. http://resolver.tudelft.nl/uuid:1c8f1bb7-c68e-4596-b341-a4f3bb70cdd9

Tummala, G. K., Das, T., Sinha, P., & Ramnath, R. (2019). SmartDashCam: automatic live calibration for DashCams. Proceedings of the 18th International Conference on Information Processing in Sensor Networks (pp. 157–168). Montreal, Quebec. https://doi.org/10.1145/3302506.3310397

Van den Beukel, A. P., Van Driel, C. J., Boelhouwer, A., Veders, N., & Heffelaar, T. (2021). Assessment of driving proficiency when drivers utilize assistance systems—the case of Adaptive Cruise Control. Safety, 7, 33. https://doi.org/10.3390/safety7020033

Vavouranakis, P., Panagiotakis, S., Mastorakis, G., & Mavromoustakis, C. X. (2017). Smartphone-based telematics for usage based insurance. In C. Mavromoustakis, G. Mastorakis, & C. Dobre (Eds.), Advances in mobile cloud computing and big data in the 5G era. Studies in big data (pp. 309–339). Cham: Springer. https://doi.org/10.1007/978-3-319-45145-9_13

Vinkhuyzen, E., & Cefkin, M. (2016). Developing socially acceptable autonomous vehicles. Ethnographic Praxis in Industry Conference Proceedings, 2016, 522–534. https://doi.org/10.1111/1559-8918.2016.01108

Vissers, J., & Tsapi, A. (2020). Kwaliteitsimpuls rijscholen. Onderzoek naar een rijscholenregister en kwaliteitscontrolesysteem [Quality impulse driving schools. Research into a driving school register and quality control system] (Report No. BG9103-RHD-ZZ-XX-RP-Z-0001). The Hague, The Netherlands: Ministry of Infrastructure and Water Management.

Vlakveld, W. P., & Wesseling, S. (2018). ADAS in het rijexamen: Vragenlijstonderzoek onder rijschoolhouders en rijexaminatoren naar moderne rijtaakondersteunende systemen in de rijopleiding en het rijexamen voor rijbewijs B [ADAS in the driving test: Questionnaire study among driving school owners and driving examiners into modern driving task support systems in driver education and the driving test for driving license B] (Report No. R-2018-20). The Hague, The Netherlands: SWOV – Institute for Road Safety Research.

Wassink, I., Van Dijk, B., Zwiers, J., Nijholt, A., Kuipers, J., & Brugman, A. (2006). In the Truman Show: Generating dynamic scenarios in a driving simulator. IEEE Intelligent Systems, 21, 28–32. https://doi.org/10.1109/MIS.2006.97

Watson-Brown, N., Mills, L., Senserrick, T., Freeman, J., Davey, J., & Scott-Parker, B. (2021). A complex system of learning to drive: The instructor's perspective. Safety Science, 136, 105172. https://doi.org/10.1016/j.ssci.2021.105172

Watson-Brown, N., Scott-Parker, B., & Senserrick, T. (2019). Association between higher-order driving instruction and risky driving behaviours: Exploring the mediating effects of a self-regulated safety orientation. Accident Analysis & Prevention, 131, 275–283. https://doi.org/10.1016/j.aap.2019.07.005

Williams, A. F., & Shults, R. A. (2010). Graduated driver licensing research, 2007–present: a review and commentary. Journal of Safety Research, 41, 77–84. https://doi.org/10.1016/j.jsr.2010.03.002

Young, K. L., & Regan, M. A. (2007). Intelligent Transport Systems to support Police enforcement of road safety laws (ATSB Research and Analysis Report No. 2007-02). Canberra City, Australia: Australian Transport

Safety Bureau.
Zahabi, M., Park, J., Razak, A. M. A., & McDonald, A. D. (2020). Adaptive driving simulation-based training: framework, status, and needs. Theoretical Issues in Ergonomics Science, 21, 537–561. https://doi.org/10.1080/1463922X.2019.1698673
Ziebinski, A., Cupek, R., Grzechca, D., & Chruszczyk, L. (2017). Review of advanced driver assistance systems (ADAS). AIP Conference Proceedings, 1906, 120002. https://doi.org/10.1063/1.5012394

Appendix A. Interview guide

Part 1: Attitudes towards the existing driving exam

1. What are, according to you, the strong aspects of the exam as it is today?
2. What are the flaws in the exam as it is today? If you had the possibility to change anything, what would you modify? How would you do it?
3. Does the driving exam as it is today allow you to assess whether a candidate would drive safely later on?
4. Would you say that the examiner's intuition plays a role in the establishment of a verdict? If yes, to what extent?

Part 2: Attitudes towards the potential data-supported driving exam

- **(a) Examiners' opinion about the use of data**
    1. Do you think that the availability of driving data of any form could be of help in the exam?
    2. Do you have any examples of data, even if it sounds strange or impossible to measure?
    3. We will provide you with several examples. For each example, we will ask you to rate this idea on the following scale from 1 to 5, 1 being not useful at all and 5 being very useful. Besides the formal answer on this scale, you are encouraged to share your thoughts, or if you come up with additional ideas. Examples:
        - Video fragments of critical situations
        - Automatically detected drivers' posture and actions (mirror checking, hands on the wheel)
        - Distance to detected objects (pedestrians, cyclists, other vehicles)
        - Deceleration and acceleration (g) scores
        - Fuel consumption/economic scores
        - Recordings of speed
        - Detected failure to follow traffic rules and norms
    4. Can you think of any other data that may be useful
- **(b) Examiners' opinion about the characteristics the data should have**
    1. Now, no matter the type of data collected, when do you think it should be collected? (before, during, after the exam)
    2. When should the data be provided to you? (before, during, after the exam)
    3. How would you envision (/imagine) that the data are presented to you?
    4. With whom do you think the data should be shared? (only you, the candidate, other examiners, data scientists...)

5. What do you think candidates will think about the use of data to assess them?
6. Is there a difference to you between driving skills and driving style? Do you think that computers could be good for the two of these?

- **(c) Examiners' views about the future of the driving exam**
  1. Is there something that takes a lot of your time during the ride during the driving exam that could be automated?
  2. We talked about how data could help you in your assessment of drivers. Do you think that in the future, an artificial intelligence could assess a driver completely, partially or completely? If yes, what would it take for you to rely on this artificial intelligence?
  3. Do you think driving simulators can play a role in the examination?
  4. How do you see the future of driving examinations?

Part 3: General questions

1. How open do you think your organization is to such technological changes? And the examiners themselves?
2. Do you think the pandemic can lead to changes in the setup of the driving exam / the work at CBR?
3. Do you have experience with other training/exams (motorcycle, truck, older drivers, other special domains)? If yes, specify. How do you think that the topic of data-driven assessment applies to that work domain?
4. Do you think the current topic may be useful to driving schools, to implement data in the training?

Appendix B. Code

To group and count the tagged transcript, Microsoft Word was used in combination with Python code that aggregates tags on the participant level. The script can count hashtags and identify tails of tags that were separated by dots, allowing for counting sub-items. In the current interview this was used, for example, to count occurrences of the "lack of time" response if the tag #weak-aspect.lack-of-time is used (we used this to create Table 2.2), or it counts how often participants were positive about simulators if the tag #simulators.positive is used.

The code and a brief demonstration can be found on GitHub, see https://github.com/tomdries/content-analysis-tools

Appendix C. Correlation matrix

The means of the seven items listed in Figure 2.2 were computed per participant to get an indication of the extent to which participants were positive about the suggested concepts for data-driven assessment. The overall mean of the 37 participants was 3.51 (*SD* = 0.67) on the five-point scale from 1 (not useful at all) to 5 (very useful).

The mean ratings showed no significant correlations with the participants' age ($r = 0.02$, $p = 0.916$), years of being an examiner ($r = -0.03$, $p = 0.865$), and years of prior experience as a driving instructor ($r = -0.05$, $p = 0.751$). Also, there was no significant difference between the mean rating of males ($M = 3.49$, $SD = 0.70$, $n = 28$) and females ($M = 3.57$, $SD = 0.62$, $n = 9$), $t(35) = -0.29$, $p = 0.771$. Table 2.3 shows the full correlation matrix among the participants' characteristics and their ratings.

Table 2.3: Pearson correlation matrix for demographic variables and ratings of the concepts (n = 37)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Age (years)													
2. Gender (0: female, 1: male)	0.15												
3. Years worked as examiner	0.71	0.04											
4. Years worked as instructor	0.13	-0.12	-0.14										
5. Was instructor (0: 0 years, 1: >0 years)	0.10	-0.05	0.06	0.79									
6. No. of driving tests per week	-0.12	0.15	-0.20	0.21	0.16								
7. Distance to objects (1 to 5)	0.09	0.01	0.04	-0.20	-0.23	0.09							
8. Video fragments (1 to 5)	-0.25	0.04	-0.28	-0.03	-0.03	0.05	0.53						
9. Speed (1 to 5)	-0.23	0.08	-0.34	-0.02	-0.04	0.36	0.58	0.54					
10. De-/acceleration (1 to 5)	0.02	-0.07	0.08	-0.13	-0.01	0.19	0.51	0.29	0.52				
11. Rules/norms adherence (1 to 5)	-0.09	-0.03	-0.16	-0.05	-0.14	-0.06	0.47	0.32	0.43	0.22			
12. Posture/actions (1 to 5)	0.26	-0.16	0.15	0.16	0.20	0.01	0.16	0.23	0.22	0.53	0.05		
13. Eco driving (1 to 5)	0.30	-0.10	0.46	0.03	0.06	-0.41	0.24	-0.05	-0.05	0.14	0.05	0.14	
14. Mean rating (1 to 5)	0.02	-0.05	-0.03	-0.05	-0.04	0.08	0.79	0.67	0.75	0.75	0.57	0.56	0.29

3

Perspectives of Truck Drivers on Their Profession and Safety Technologies

This chapter was published as: De Winter, J.*, Driessen, T.*, Dodou, D., & Cannoo, A. (2024). Exploring the challenges faced by Dutch truck drivers in the era of technological advancement. Frontiers in Public Health, 12, 1352979. https://doi.org/10.3389/fpubh.2024.1352979

* joint first authors

Abstract

Despite their important role in the economy, truck drivers face several challenges, including adapting to advancing technology. The current study investigated the occupational experiences of Dutch truck drivers to detect common patterns. A questionnaire was distributed to professional drivers in order to collect data on public image, traffic safety, work pressure, transport crime, driver shortage, and sector improvements. The findings based on 3708 respondents revealed a general dissatisfaction with the image of the industry and reluctance to recommend the profession. A factor analysis of the questionnaire items identified two primary factors: 'Work Pressure', more common among national drivers, and 'Safety & Security Concerns', more common among international drivers. A ChatGPT-assisted analysis of textbox comments indicated that vehicle technology received mixed feedback, with praise for safety and fuel-efficiency improvements, but concerns about reliability and intrusiveness. In conclusion, Dutch professional truck drivers indicate a need for industry improvements. While the work pressure for truck drivers in general may not be high relative to certain other occupational groups, truck drivers appear to face a deficit of support and respect.

3.1. Introduction

Given the Netherlands' strategic position as a gateway to Europe and its port infrastructure, the truck driving profession plays a key role in the economic success of the country. As of 2021, approximately 91,000 professional truck drivers were registered in the Netherlands (Sector Institute Transport and Logistics, 2021).

Truck drivers face various challenges that can affect their well-being, such as long working hours and extended periods away from home, which may adversely impact mental health and familial relationships (Chen et al., 2015; Johnson et al., 2021; Shattell et al., 2010; Shin & Jeong, 2020; Statistics Netherlands, 2021; Williams et al., 2017). Additionally, the sedentary nature of the truck driving profession involves health risks such as obesity (Bachmann et al., 2018; Dos Reis et al., 2017; Sieber et al., 2014). Another challenge is the pressure to meet tight delivery schedules, which can result in fatigue and compromised road safety (Belzer, 2018; Chen et al., 2015; Hege et al., 2019; Useche et al., 2021). A study among truck drivers by Wijngaards et al. (2019) showed that the driving itself, as well as the rest breaks and administrative tasks, are associated with greater momentary happiness compared to logistical work and the delivery/pickup of goods.

Truck drivers also grapple with adapting to the evolving technological landscape, including the adoption of advanced driver assistance systems (ADAS), such as adaptive cruise control (ACC) and lane keeping assistance (LKA) systems, as well as digital tools that aim to improve safety and efficiency (Loske & Klumpp, 2021). While new technologies offer potential benefits, they can also generate resistance (Klumpp, 2018), cause apprehension about job displacement (Dubljević et al., 2023), and require truck drivers to acquire new skills (Jaller et al., 2022; Schuster et al., 2023; Van Fossen et al., 2023). Semeijn et al. (2019), for example, reported that the digital tachograph is a source of stress.

Various studies have been undertaken on the topic of ADAS, typically using driving simulators and focusing on passenger vehicles (Gouribhatla & Pulugurtha, 2022; Rahman et al., 2017; Rossi et al., 2020). Current literature suggests a preference among truck drivers for a silent cabin environment (Bazilinskyy et al., 2019; Fors et al., 2015). Certain systems, such as autonomous emergency braking (AEB) and warning systems, are likely beneficial from a safety perspective (Hickman et al., 2015; Teoh, 2021). However, these systems exhibit a propensity for false interventions/alarms, rendering them annoying or intrusive (Dreger et al., 2020; Fagerlönn, 2011; Morton et al., 2019; Svenson et al., 2017). Camera systems and ADAS that reduce blind spots, on the other hand, have been met with approval by truck drivers (Ostermann et al., 2016). Still, which ADAS are perceived by truck drivers as useful and which as less useful has not yet been well documented in the literature.

Research Aim

Although certain pain points in the trucking industry have been documented (e.g., tight schedules, stress, and fatigue: Anderson et al., 2017; Chen et al., 2021; Delhomme & Gheorghiu, 2021; Reiman et al., 2018), there is still limited knowledge about how truck drivers experience their daily work. This is particularly relevant in recent years, as factors such as driver shortages (Ji-Hyland & Allen, 2022) and the introduction of new technologies are playing increasingly large roles.

The aim of this study is to document the experiences of Dutch truck drivers. A large-scale questionnaire was conducted by Transporteffect BV (which is engaged in advisory services and mediation within the transportation sector) and foundation Chauffeursnieuws (a website focused on the transport industry). Although the results of the questionnaire have been published in raw form on the organization's website (Transporteffect, 2021), they have not yet been subjected to scientific evaluation. This paper analyzes the results of this questionnaire, which includes responses from over 3700 drivers, through a multivariate statistical approach and through a ChatGPT-aided text summarization approach. This analysis allowed for making informed statements about the experiences of drivers and to determine whether there are relevant patterns in their experiences, which may potentially correlate with individual differences such as gender, age, and type of work (national vs. international). By better understanding truck driver experiences, policymakers and industry stakeholders could make more informed decisions to improve the working conditions and job satisfaction of truck drivers.

3.2. Methods

3.2.1 Questionnaire design

The questionnaire header indicated that Chauffeursnieuws & Transporteffect aimed to address the long-neglected concerns of professional drivers and promote their welfare. It stated that by providing a platform for drivers to voice their opinions, the organizations were committed to creating a positive impact on the transportation sector.

The questionnaire was administered in Dutch and consisted of 68 questions divided into 9 parts. It included 51 multiple-choice questions, 1 checkbox question, and 15 open comment boxes that provided the option to the respondent to elaborate on the preceding multiple-choice questions.

Part 1: Introduction (Q2–Q6) gathered general information about the respondents. Example questions included: *"Your gender?"* (Q2) with response options *Male* and *Female*, and *"Are you a professional driver?"* (Q3) with response options *Yes* and *No*.

Part 2: Organizations (Q7–Q12) focused on the respondents' involvement and opinions on trade unions and other organizations. For example, *"Are you a member of a trade union?"* (Q7) with response options Yes and No, and *"CNV - What grade would you give?"* (Q9), with response options *1 (Very bad)* to *5 (Very good)*, and *No opinion*.

Part 3: Image (Q13–Q17) dealt with the public image of drivers and related topics. Example questions were: *"Do you think the image of the driver needs to be improved?"* (Q13) and *"Do you think a mobile toilet (DIXI) at companies is a good solution for drivers?"* (Q14), both with response options *No*, *Yes*, and *Neutral*.

Part 4: Traffic safety (Q18–Q30) explored the respondents' views on various traffic safety issues. Example questions included: *"Do you think a stand-alone air conditioner contributes to road safety in Europe?"* (Q18) with response options *Yes*, *No*, and *No opinion*, and *"Do you find text signs with information adequate for international traffic?"* (Q21) with response options *No - creates dangerous situations*, *No*, *Yes*, and *No opinion*.

Part 5: Work pressure (Q31–Q39) investigated the respondents' experiences and opinions about work pressure. Example questions were: *"Do you experience high work pressure?"* (Q31) with response options *No*, *Yes - every day*, *Yes - 1 or 2 times per week*, and *Yes - 1 time per month on average*, and *"Do you think work pressure should be addressed?"* (Q35) with response options *Yes*, *No*, and *No opinion*.

Part 6: Transport crime (Q40–Q44) focused on transport crime issues and their impact on the respondents. Example questions included: *"Have you dealt with transport crime?"* (Q40) with response options *Yes - regularly, Yes - sometimes*, and *No*, and *"Do you report all forms of crime to the authorities? Or via https://meldpunt-transport.nl/"* (Q42) with response options *No - small events not, No - never, Yes - only big events*, and *Yes - all events*.

Part 7: Driver shortage (Q45–Q56) explored the respondents' perceptions of the driver shortage and related topics. Example questions were: *"Do you feel the demand for professional drivers has increased?"* (Q45) with response options *Yes - much more demand, Yes - a little more, No - not more than in the last 10 years,* and *No opinion*, and *"Do you find the hourly wage sufficient compared to similar jobs?"* (Q47) with response options *Yes, No,* and *No opinion*.

Part 8: General questions (Q57–Q66) dealt with various topics, including paid parking and the European Mobility Package (EU regulations to improve road transport conditions; European Commission, 2022). Example questions included: *"Do you think paid parking for trucks is a solution?"* (Q57) with response options *Yes - better facilities*, *Yes - only if well organized*, *No - only take money from the sector*, *No - no rest possible*, and *No opinion*, and *"What do you think of the current European Mobility Package?"* (Q62) with response options *1 (Bad)* to *5 (Very good)*.

Part 9: The concluding section (Q67–Q68) provided space for respondents to share their opinions on the most important changes needed in the sector and any additional comments or suggestions. The two questions were: *"Open question: What is, in your opinion, the first thing that needs to change in the sector? (Please provide 1 answer)"* (Q67), and *"Comments and suggestions that you couldn't include in the questions can be written below."* (Q68).

The open comment boxes were present in each part: Part 2 (Q12), Part 3 (Q17), Part 4 (Q25, Q28, Q30), Part 5 (Q34, Q39), Part 6 (Q44), Part 7 (Q51, Q53), Part 8 (Q58, Q60, Q63), and Part 9 (Q67, Q68). For an overview of all questions, please refer to the Data availability section.

3.2.2 Questionnaire dissemination

The questionnaire was administered in September and October 2021, with invitations disseminated through the website www.transporteffect.com and the corresponding LinkedIn and Facebook pages, platforms for sharing truck-related news articles.

3.2.3 Data pre-processing

In total, 3845 respondents completed the questionnaire. Of these, 137 indicated that they were not professional truck drivers and were therefore excluded from the analysis, leaving 3708 respondents. The questionnaire contained 51 multiple-choice items, which were analyzed separately from the open comment boxes. One question (Q15, about mobile toilets) was excluded because we considered it unclear.

The 50 remaining questions were divided into three categories:

- Driver-related questions (Q2: *"Your gender?"*, Q3: *"Are you a professional driver?"*, Q4: *"Where do you primarily drive?" (1: National, 2: Benelux + Ruhr area, 3: International)*, Q5: *"How old are you?"*).
- General outcome questions (Q13: *"Do you think the image of the driver needs to be improved?"*, Q16: *"What is your general impression of the image of the professional driver?"*, Q46: *"Would you recommend the profession to family or acquaintances?"*, Q56: *"What grade would you generally give to the professional driver's profession?"*, Q64: *"How do you see the future as a Dutch professional driver?"*).
- Forty-one, more specific, questions.

The driver-related questions and general outcome questions were used as criterion variables, while the remaining questions were subjected to a multivariate statistical analysis.

Response options for questions were not always on an ordinal scale and sometimes included *Not applicable, No opinion*, or *Don't know* choices. Therefore, the response options were sorted from low to high, response options that were equivalent on an ordinal scale (for example, *No - creates dangerous situations* and *No*) were combined, and the *Not applicable/No opinion/Don't know* options were marked as missing responses, since such responses cannot be used in standard linear statistical methods. For an overview of the response frequency distributions pertaining to each question, please refer to the Data availability section.

The number of times No opinion, Not applicable, or Don't know were answered was low for some questions (e.g., 0.2% for Q50, *"Do you think the driver's profession gets the respect it deserves?"*). However, for some questions, these responses were more frequent. For example, for the question *"Do you report all forms of crime to the authorities? or via ht tps://meldpunt-transport.nl/"* (Q42), 42.4% reported *Not applicable*, presumably because these drivers had not experienced any crime.

Regarding the grading of different unions and trade organizations (Q8–Q11), there was also a high prevalence of *No opinion* responses (26.0, 44.0, 32.8, & 24.9%), likely because drivers were not members or had not dealt with every organization. Since the aim of our research was to assess the general sentiment of drivers, not specific organizations, these four questions were combined into one by averaging, reducing the percentage of missing data for this question to 10.0%.

As the overall number of missing responses was low (6.9% of the 3708 × 38 matrix of numbers), it was decided to impute these missing values, approximately preserving the means and intercorrelations between item responses. Specifically, missing data were imputed using the nearest-neighbor method, whereby the missing data in the 3708 respondents × 38 questions matrix were imputed with the value of the nearest-neighbor row according to the Euclidean distance.

3.2.4 Statistical analysis

The mean scores on the 38 questions were interpreted to describe key patterns. Following this, the data (3708 × 38 matrix of numbers) were subjected to exploratory maximum likelihood factor analysis. This statistical method aims to explain the correlations among variables by identifying latent factors that influence these variables; it is frequently used in the analysis of questionnaire data to reveal underlying psychological constructs (Fabrigar et al., 1999). The number of factors to extract was based on the scree plot, a graphical representation where eigenvalues (corresponding to the percentage of variable explained) of the correlation matrix are plotted in descending order. The plot generally begins with a steep slope before leveling off, creating an elbow-like shape. The point at which the slope starts to level off is deemed the optimal number of factors to retain (Zhu & Ghodsi, 2006).

Subsequently, the factor loadings were orthogonally rotated using the Varimax method. Although it could be expected that underlying factors would correlate positively, an orthogonal rotation was chosen. This was done because we were interested in the discriminative power of the factors and their relationship with driver characteristics (rather than a 'general positivity' that may be expressed in multiple factors). Factor scores were calculated using the weighted least-squares method. The factor scores were subsequently standardized to have a mean of 0 and a standard deviation of 1.

The scores on the extracted factors were then correlated with the aforementioned criterion variables. Note that Q3 (*"Are you a professional driver?"*) was not used in this analysis because we only included respondents who answered Yes to this question; hence, this item exhibits no variance.

3.2.5 Text analysis: summaries of open comment boxes

The questionnaire contained a number of open comment boxes. Traditional methods such as content analysis and thematic analysis involve human raters examining the text for specific themes (e.g., Braun & Clarke, 2006; Krippendorff, 2004). However, these methods come with the disadvantage of subjectivity and limited reproducibility (Kitto et al., 2023; Roberts et al., 2019).

Recently, large language models have emerged as a promising alternative. ChatGPT has been shown to perform well in reading comprehension and other linguistic tasks (Bubeck et al., 2023; De Winter, 2023; Liu et al., 2023; OpenAI, 2023). In this paper, we will use it for two purposes: summarizing open-ended responses and extracting sentiment from responses.

In summarization applications, ChatGPT's capabilities have been shown in various fields (Laban et al., 2023), including clinical texts (Van Veen et al., 2024) and news items (Pu et al., 2023). Regarding sentiment analysis, research has shown that ChatGPT can generate mean sentiment scores that correlate strongly with human sentiment ratings and with VADER sentiment analysis, an existing sentiment analysis model (Tabone & De Winter, 2023). ChatGPT has also been found to outperform humans in extracting the stance and topics of tweets (Gilardi et al., 2023; Törnberg, 2023), and surpass state-of-the-art models in analyzing various types of texts such as customer reviews, social media posts, and news items (Wang et al., 2023).

We used a custom script to upload the responses for each open comment box to OpenAI's API (GPT-4, model: gpt-4-0125-preview; date: March 2, 2024). The responses were accompanied by the following prompt: *"Please make a very very short summary of the respondents' comments shown above, IN ENGLISH; do not enumerate"*. The parameter temperature, which determines the degree of randomness of the output, was set to 0 to yield a nearly deterministic output.

Although ChatGPT can properly handle potential gibberish responses or 'empty' responses such as a single character (Tabone & De Winter, 2023), we have nonetheless applied a filter whereby only text responses of 4 or more characters were included in the input to ChatGPT. By excluding extremely short responses, we ensured our sample size was more accurately represented by respondents who offered feedback.

3.2.6 Text analysis: vehicle aids and on-board computer

A key research question of this study focuses on drivers' perceptions of technology. The responses to the open-ended question regarding vehicle aids (Q30: *"Comment: vehicle aids"*) featured numerous comments on specific assistance systems, predominantly concerning the following four types:

- Adaptive cruise control (ACC)

- Lane departure warnings (LDW) / Lane keeping assistance (LKA)
- Emergency braking / AEB
- Camera systems and smart mirrors

For the comments in Q30, we manually identified the system(s) being referred to in the comment (Appendix A). Then, for each of the systems, the corresponding quotes were fed to GPT-4, with the following prompt:

> *What do the users think about the discussed system? Give a very short summary; do not enumerate.*

The same prompt was used for the responses to the open-ended question regarding the onboard computer (Q34: *"Comment: on-board computer"*).

Finally, numerical sentiment scores were generated for the comments for each of the four ADAS in Q30, through the following prompt:

> *These text messages are obtained from a textbox in a questionnaire about technology in trucks. I need you to provide a single sentiment rating about the technology being discussed in the comments, from 1 (extremely negative) to 100 (extremely positive). Only report a single number between 1 and 100, rounded to two decimals. no text!*

A bootstrapping approach was adopted for this process (Tabone & De Winter, 2023; Tang et al., 2023), where all comments per ADAS (Q30) were sorted in random order, and the mean score over 1000 attempts was taken as an overall indicator of sentiment. The use of this method was deemed necessary because the way ChatGPT operates brings a certain randomness to the output. By averaging over a large number of repetitions under effectively identical conditions (only the order of the comments differs), a statistically reliable assessment is obtained of how ChatGPT judges the sentiment of the respondents' texts.

3.3. Results

3.3.1 Driver-related questions

A total of 3708 respondents were included in the study, with 3541 (95%) identifying as male (Q2). The age distribution of the respondents (Q5) was as follows: 270 individuals aged 18–25, 969 aged 25–40, 884 aged 40–50, 1175 aged 50–60, and 410 aged 60–75 years old. In terms of driving regions (Q4), 1483 respondents reported being national drivers, 1552 identified as international drivers, 666 specified driving in the Benelux & Ruhr area (i.e., Belgium, Netherlands, Luxembourg, and the Ruhr industrial region in Western Germany), while 7 respondents indicated that the question was not applicable to them.

3.3.2 General outcome questions

Respondents expressed some concerns about the image of their industry, hesitancy to recommend the career to others, and a neutral to slightly negative outlook on the future.

Specifically:

- 88.1% (3265) of respondents believe the image of the driver needs improvement, 7.3% (269) remain neutral, and 4.7% (174) disagree (Q13).
- The general impression of the image of the professional driver leans towards negative, with a mean score on the scale of *1 (Very negative)* to *5 (Very positive)* of 2.62 (Q16).
- 68.7% (2549 respondents) would not recommend the profession of a professional driver to family or acquaintances, while 31.3% (1159 respondents) would recommend it (Q46).
- Responding to the question, *"What grade would you generally give to the professional driver's profession?"*, the mean grade provided by respondents was 6.27 out of 10 (Q56). The most common grade was 7 (n = 1006).
- Finally, the majority of the respondents have a neutral to slightly negative outlook on the future, with a mean of 4.48 on a scale of *1 (Very negative*) to *10 (Very positive)* (Q64). The most frequently selected grade was 5 (n = 770).

3.3.3 Specific questions: mean ratings

The questionnaire used different response options for the questions, including yes/no and scales of 1–3 or 1–5. This differentiation aimed to better match the nature of each question, and may increase respondent engagement while reducing yea-saying bias (Saris et al., 2010). However, it inhibits direct comparison of items based on their mean score.

Table 3.1 shows mean scores for the 38 items, with a 'normalized mean' column ranging from 0 to 1, which allows a clearer view of the drivers' agreement with statements across items. The results are interpreted below on this 0 to 1 scale.

Regarding workplace and road safety, the use of mobile toilets at companies received a low score of 0.05 (Q14). Overtaking bans on highways scored only 0.14 (Q24). Aids in vehicles were assigned a score of 0.72, indicating a general agreement about their contribution to road safety (Q29). Furthermore, respondents found that a stand-alone air conditioner contributes to road safety (Q18, score: 0.89). A score of 0.91 was reported for the feeling that space on the roads has decreased, indicating a universal observation (Q22).

Regarding work pressure, a score of 0.45 was observed for drivers experiencing high work pressure (Q31), with a score of 0.44 regarding the feeling that work pressure affects their driving behavior (Q19). A high score of 0.81 was obtained for the belief that work pressure should be addressed (Q35).

In terms of compensation and financial aspects, a low score of 0.04 was found for the sufficiency of the hourly wage compared to similar jobs (Q47), while a score of 0.49 indicated that nearly half of the drivers find it difficult to make ends meet with one salary (Q66).

As for work-related issues, while most drivers reported that they are satisfied with their employers (0.90; Q6) and have not experienced labor exploitation (Q61) or intimidation (Q37) from their employers, a portion of respondents reported such issues (0.24 & 0.12, respectively). Additionally, a score of 0.18 was observed for having dealt with transport crime (Q40).

When considering work-life balance, a score of 0.62 was obtained for drivers who prefer to be home every evening (Q65). On the other hand, a score of 0.66 was obtained for drivers who exceed driving times out of necessity (Q38). This points to the difficulty some drivers face in maintaining a balance between work and personal life.

In the context of infrastructure, the quality of roads in the Netherlands received a high score of 0.75 (Q26). However, drivers reported a score of 0.60 for experiencing problems finding a decent parking spot in time (Q55).

Finally, regarding the perception of the profession and industry-related organizations, a score of 0.17 was reported for the belief that the truck driver's profession receives the respect it deserves (Q50). A high score of 0.83 was obtained for the importance of driver education for raising awareness (Q43), while 0.86 was reported for the increased demand for professional drivers (Q45). However, a high score of 0.89 was observed for the belief that organizations supporting transport are doing too little (Q52).

Table 3.1: Overview of the 38 items subjected to statistical analysis. This table presents the mean score and standard deviation (SD) for 3708 respondents, along with the normalized mean, which is the mean linearly scaled between the minimum and maximum possible score on the question.

No	Question	Response options	Mean	*SD*	Mean (normalized)
Q47	Do you find the hourly wage sufficient compared to similar jobs?	1 = No, 2 = Yes	1.04	0.20	0.04
Q14	Do you think a mobile toilet (DIXI) at companies is a good solution for drivers?	1 = No, 3 = Yes	1.09	0.38	0.05
Q37	Do you ever experience intimidation from your employer?	1 = No, 2 = Yes	1.12	0.32	0.12
Q36	Have you ever been asked to commit tachograph fraud?	1 = No, never, 3 = Yes, regularly	1.24	0.51	0.12
Q24	Do you find overtaking bans on highways beneficial for road safety?	1 = No, 2 = Yes	1.14	0.35	0.14
Q50	Do you think the driver's profession gets the respect it deserves?	1 = No, 3 = Yes	1.34	0.53	0.17
Q40	Have you dealt with transport crime?	1 = No, 3 = Yes, regularly	1.35	0.52	0.18

Table continued from previous page

No	Question	Response options	Mean	*SD*	Mean (normalized)
Q49	Do you think you will be able to perform the job until 70+?	1 = No, 2 = Yes	1.19	0.39	0.19
Q61	Have you ever felt that you were dealing with labor exploitation?	1 = No, 2 = Yes	1.24	0.43	0.24
Q21	Do you find text signs with information adequate for international traffic?	1 = No, 2 = Yes	1.24	0.43	0.24
Q42	Do you report all forms of crime to the authorities? Or via https://meldpunt-transport.nl/	1 = No, never, 4 = Yes, all events	1.77	0.97	0.26
Q41	Has your tarp ever been cut?	1 = No, 3 = Yes, regularly	1.53	0.61	0.27
Q7	Are you a member of a trade union?	1 = No, 2 = Yes	1.31	0.46	0.31
Q62	What do you think of the current European Mobility Package?	1 = Bad, 5 = Very good	2.45	0.88	0.36
Q33	Do you think the on-board computer contributes to high work pressure?	1 = No, 2 = Yes, definitely	1.43	0.49	0.43
Q59	What grade would you give to existing paid parking spaces ?	1 = Very bad, 5 = Very good	2.72	0.95	0.43
Q19	Do you feel work pressure that affects your driving behavior?	1 = No, 3 = Yes, regularly	1.87	0.71	0.44
Q31	Do you experience high work pressure?	1 = No, 4 = Yes, every day	2.36	1.16	0.45
Q57	Do you think paid parking for trucks is a solution?	1 = No, 2 = Yes	1.46	0.50	0.46
Q66	Can you make ends meet with one salary?	1 = No, 2 = Yes	1.49	0.50	0.49
Q20	Have you ever held the phone while driving?	1 = No, never, 3 = Yes, regularly	2.05	0.67	0.52
Q8–Q11	Organizations - What grade would you give?	1 = Very bad, 5 = Very good	3.10	0.94	0.53
Q32	Do you ever continue driving when you feel tired?	1 = No, never, 3 = Yes, regularly	2.08	0.69	0.54
Q54	Do you spend every day calculating to comply with driving and rest time regulations?	1 = No, 3 = Yes, it's difficult	2.11	0.79	0.56

Table continued from previous page

No	Question	Response options	Mean	*SD*	Mean (normalized)
Q55	Do you experience problems finding a decent parking spot in time?	1 = No, 4 = Yes, every day	2.79	0.94	0.60
Q65	As a professional driver, do you prefer to be home every evening?	1 = No, I want to be on the move as much as possible, 4 = Yes	2.87	0.83	0.62
Q48	Do you find the profession you practice demanding?	1 = No, 3 = Yes it's heavy	2.31	0.65	0.66
Q38	Do you ever exceed driving times out of necessity?	1 = No, 3 = Yes	2.32	0.82	0.66
Q23	Do you think increasing truck speed contributes to better traffic flow and safety?	1 = No, 80 kilometers is fine, 3 = Yes, 90 kilometers is ideal	2.43	0.74	0.71
Q29	Do you think the aids in vehicles contribute to road safety?	1 = No, not at all, 4 = Yes	3.15	0.63	0.72
Q26	How do you find the quality of roads in the Netherlands?	1 = Very bad, 5 = Very good	3.99	0.67	0.75
Q35	Do you think work pressure should be addressed?	1 = No, 2 = Yes	1.81	0.39	0.81
Q43	Do you think driver education is important for raising awareness?	1 = No, 2 = Yes	1.83	0.37	0.83
Q45	Do you feel the demand for professional drivers has increased?	1 = Not more than in the last 10 years, 3 = Yes, much more demand	2.72	0.58	0.86
Q52	Do you think organizations that are there for transport are doing too little?	1 = No, 2 = Yes	1.89	0.31	0.89
Q18	Do you think a stand-alone air conditioner contributes to road safety in Europe?	1 = No, 2 = Yes	1.89	0.31	0.89
Q6	Are you generally satisfied with your employer?	1 = Very negative, 4 = Positive	3.71	0.55	0.90
Q22	Do you feel that space on the roads has decreased?	1 = No, 2 = Yes	1.91	0.29	0.91

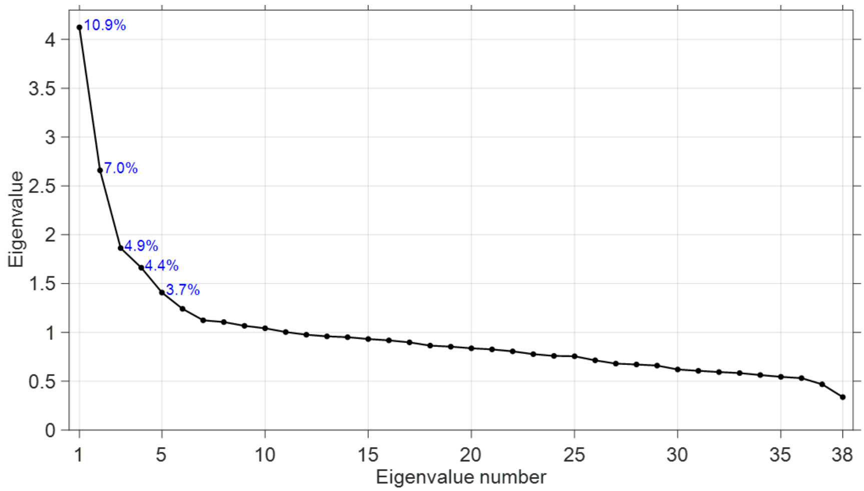


Figure 3.1: Scree plot of the 38 x 38 correlation matrix.

3.3.4 Specific questions: factor analysis

The results from the 38 questions were subjected to a factor analysis in order to extract underlying factors. The scree plot (Figure 3.1) indicated that the extraction of two factors would be appropriate, though the percentage of explained variance was not high. However, this may not impact the reliability of the constructs as long as a large number of variables correlates with the factor (De Winter et al., 2009).

The Varimax-rotated loadings (please refer to the Data availability section) allowed us to interpret the two factors as follows: (1) Work Pressure and (2) Safety & Security Concerns. More specifically:

Factor 1: Work Pressure. Items with high loadings on this factor relate to work pressure and its impact on drivers. The strongest loadings relate to experiencing high work pressure (0.75; Q31, and 0.74; Q19). Other high loadings involve ever experiencing intimidation from one's employer (0.45; Q37), experiencing the profession as demanding (0.45; Q48), being satisfied with one's employer (-0.44; Q6), continuing to drive when feeling tired (0.46; Q32), the on-board computer contributing to high work pressure (0.46; Q33), and having ever felt that one was dealing with labor exploitation (0.42; Q61).

Factor 2: Safety & Security Concerns. Items with high loadings on this factor are related to the security and working conditions of drivers. The strongest loadings are related to dealing with transport crime (0.49; Q40), having one's tarp cut (0.45; Q41), experiencing problems finding decent parking spots (0.51; Q55), and exceeding driving times out of necessity (0.45; Q38). Variables related to international driving showed strong loadings as well: preferring being home every evening (-0.49; Q65) and opinion about the European Mobility Package (-0.39; Q62).

The reported crimes (Q44) primarily involve diesel theft, alongside other offenses such as vehicle or container break-ins, and theft of personal belongings or cargo. Incidents of stowaways and intimidating encounters with migrants have also been noted.

Table 3.2: Correlation coefficients between item responses and factor scores.

No	Question	Response options	Work Pressure	Safety & Security Concerns
Q2	Your gender?	1 = Male, 2 = Female	0.05	-0.09
Q4	Where do you primarily drive?	1 = National, 2 = Benelux + Ruhr area, 3 = International	-0.21	0.51
Q5	How old are you?	1 = 18–25, 5 = 60–75	-0.03	0.00
Q13	Do you think the image of the driver needs to be improved?	1 = No, 3 = Yes	0.09	0.01
Q16	What is your general impression of the image of the professional driver?	1 = Very negative, 5 = Very positive	-0.19	-0.24
Q46	Would you recommend the profession to family or acquaintances?	1 = No, 2 = Yes	-0.22	-0.17
Q56	What grade would you generally give to the professional driver's profession?	1 = Very bad, 10 = Very good	-0.37	-0.24
Q64	How do you see the future as a Dutch professional driver?	1 = Very negative, 10 = Very positive	-0.22	-0.34

Correlation coefficients with binary variables (Q2, Q46) are also known as point-biserial correlation coefficients. Given the substantial sample size (n = 3708), minor correlations statistically deviate from zero, with $p < 0.01$ when $|r|$ is greater than or equal to 0.05

Next, factor scores were calculated and correlated with the driver-related questions and the general outcome questions. The results in Table 3.2 show that there are small gender differences, with women being slightly more burdened by work pressure and men slightly more by crime. This latter finding can be explained by the increased likelihood of men being international drivers.

The factor scores consistently correlate with the outcome measures, such as the respondents' impression of the image of the truck driver (Q16), whether they would recommend the profession to family or acquaintances (Q46), the score they attribute to the profession as a whole (Q56), and how they view the future (Q64). Work Pressure is primarily associated with the impression of the profession now (Q56), while Safety & Security Concerns is more strongly associated with whether the future is judged optimistically (Q64).

Finally, a trend emerges wherein Work Pressure is relatively high among drivers oper-

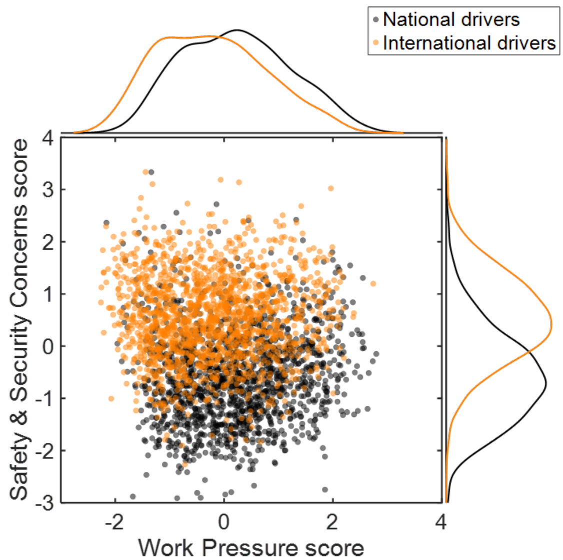


Figure 3.2: Scatter and kernel density estimate plot of factor scores, differentiating between national drivers (n = 1483) and international drivers (n = 1552).

ating nationally, while Safety & Security Concerns are relatively high among international drivers (Q4). The mean (SD) scores for Work Pressure are 0.22 (0.98) for national drivers, 0.05 (0.95) for drivers in the Benelux/Ruhr Area, and -0.23 (0.99) for international drivers.

On the other hand, the mean (SD) scores for Safety & Security Concerns are -0.57 (0.90) for national drivers, 0.00 (0.86) for Benelux/Ruhr Area drivers, and 0.55 (0.82) for international drivers. The difference in these experiences between national and international drivers is visually illustrated in Figure 3.2.

3.3.5 Text analysis: summaries of open comment boxes

Appendix B provides summaries of respondents' feedback regarding the 15 open-ended questions presented in the questionnaire. These summaries reveal themes that encompass a wide array of subjects, including the perceived commitment of organizations, concerns surrounding driving times, availability of parking spaces, and the contentious issue of overtaking bans for trucks. Given that the summaries of the 15 open comment boxes still contain a lot of information, and sometimes in a redundant form, we asked ChatGPT to generate a single meta-summary using the following prompt.

> *Formulate a very concise summary from the following summaries without mentioning specific persons or organizations.*

The meta-summary generated by ChatGPT is as follows:

Truck drivers express widespread dissatisfaction with their profession, citing undervaluation, mistreatment, and poor working conditions. They highlight issues such as unsanitary facilities, lack of respect from employers, clients, and the public, and the negative impact of foreign drivers. Concerns also include dangerous road conditions due to overtaking bans, inadequate education on blind spots, and mixed opinions on technological aids, which some say decrease attentiveness. Many criticize trade unions and employer organizations for not effectively representing their interests, particularly regarding wages and working conditions. The use of onboard computers is seen as increasing work pressure and invading privacy. Criminal activities like diesel theft and insufficient secure parking are major concerns, alongside the high costs and poor quality of paid parking facilities. Drivers call for improved treatment, compensation, and better enforcement of regulations to ensure fair competition and working conditions, emphasizing the need for higher wages, more flexible working hours, and better parking and rest facilities.

3.3.6 Text analysis: vehicle aids

In addition to the general summary of the open comment boxes provided above, we conducted an analysis specifically focusing on the evaluation of technologies in the vehicle. The results of this analysis are shown in Table 3.3.

ACC is valued for its safety, fuel efficiency, and calming effect on driving, though concerns about incorrect reactions and potential distractions exist. LDW and LKA systems can be irritating due to frequent alerts, leading some drivers to disable them. AEB systems can prevent accidents but raise safety concerns due to false activations, causing potential hazards for following traffic and heavy loads. Opinions on camera systems and smart mirrors are mixed; they increase visibility and reduce blind spots but can malfunction and disconnect drivers from the environment. Finally, on-board computers improve logistical efficiency and communication but may increase work pressure, create additional tasks, and infringe on driver autonomy due to real-time tracking.

Table 3.3: Number of comments, ChatGPT summaries, and manually selected comments for systems mentioned in Q30 and Q34.

Category	Count	GPT-4 response and manually selected quotes
Adaptive cruise control (ACC)	74	The opinions on the Adaptive Cruise Control (ACC) system among users are mixed. Some praise it for its convenience, safety enhancements, and fuel-saving potential, appreciating its ability to maintain distance and assist in driving. Others criticize it for causing unnecessary braking, reacting to non-hazardous objects, and potentially leading to decreased driver attentiveness and reliance on technology. Concerns are also raised about its effectiveness in heavy traffic and interactions with other drivers' unpredictable behaviors. Overall, while many see the benefits of ACC, there are significant concerns about its current implementation and impact on driving habits.
		• *"Wouldn't want to miss the ACC … anymore!"* • *"Some aids also react to other situations like adaptive cruise control that also reacts to viaducts traffic signs and then it makes an emergency stop out of nowhere also extremely dangerous"* • *"Certain aids are fantastic, but some also promote inattentiveness such as ACC."*
Lane departure warnings (LDW) / Lane keeping assistance (LKA)	25	The users generally find the lane assist or line detection system to be annoying, irritating, and often unnecessary. They express frustration with false alarms and the system's intrusiveness, with some preferring traditional driving methods without such interventions.
		• *"Lane assist is mega annoying and I turn it off when I'm fit. Later in the day, it comes on."* • *"You get insanely annoyed by all those alarms. Especially from that line detection. 9 out of 10 times it goes off for no reason."* • *"Line detection … encourages you to pick up your mobile. And all that touchscreen stuff only takes the eyes off the road. Just give me press and twist buttons. The more stuff on a car the less alert people are. If something suddenly happens, people no longer know how to intervene."*

Emergency braking / AEB	58	The users express significant concerns and dissatisfaction with the automatic emergency braking systems in vehicles, particularly trucks. They report that these systems often activate inappropriately, responding to non-hazards like traffic signs, reflections, or even shadows, leading to dangerous and unexpected braking situations. While a few see the potential safety benefits, the majority criticize the systems for creating more risks than they mitigate, especially in situations where following drivers are caught off guard by sudden stops. There's a general sentiment that these systems need improvement to truly enhance road safety.
		• *"Yes absolutely. My truck automatically brakes in an emergency situation. If it ever comes to the point where every truck has this technology (mandatory), then at least no truck will ever run into the back of a stationary traffic jam."* • *"I turn them off. Almost had an accident because the truck went full on the brakes in a slight curve at 80 km/h. The automatic braking system was triggered because my own light (headlights) reflected on a traffic sign."* • *"Some systems are downright life-threatening. For example, the emergency braking system, when you are cut off by a motorist, the system goes into action causing a great chance that your follower will shoot under your trailer."*
Camera systems and smart mirrors	58	The users express mixed opinions about the use of camera systems in vehicles. Some appreciate the enhanced visibility and safety features cameras provide, such as reducing blind spots and aiding in maneuvers like reversing. They find cameras, including blind spot and reversing cameras, to be helpful tools that can prevent accidents. However, others raise concerns about reliability issues, such as cameras being affected by weather conditions or failing to accurately reflect depth. There's also a sentiment that reliance on cameras can lead to decreased attention to traditional driving practices, like using mirrors and making eye contact with other drivers, potentially reducing interaction with other traffic and increasing distraction. Overall, while many see the benefits of camera systems for safety and visibility, there are significant reservations about their effectiveness and impact on driving habits.
		• *"I have a camera system etc. for London on my car, this camera greatly reduces my blind spot and I now see much more on the highway but also on roundabouts and through cities."* • *"... Some camera systems can help. Cameras instead of mirrors, not so much, because you lose visual contact with other road users."*

		• *"Camera mirrors do not reflect depth and when it rains you see nothing and they break quickly."*
On-board computer	625	The users have mixed feelings about the system, with some seeing it as a helpful tool that can make work more efficient and reduce the need for constant communication with the planning department. Others feel it increases work pressure by allowing for constant monitoring and adding more tasks, leading to a sense of being constantly watched and reducing personal freedom. Some users also mention the system can be distracting and contribute to stress, especially when it leads to additional administrative tasks or when planning uses it to push for more work to be done in less time.
		• *"It depends on how the on-board computer is used. You as a driver and on the other side the planning that provides you with work. If there is good consultation with the planning, then the on-board computer is also an addition that could bring peace."* • *"You are continuously monitored, if you are ahead of schedule then extra loading addresses are added."* • *"... The on-board computer does take away the so-called "sense of freedom" although I have complete understanding for the need to account for hours."*

The above findings are corroborated by numerical sentiment scores computed using ChatGPT. More specifically, the mean (*SD*) sentiment scores across the bootstrapped batches were 58.0 (5.07) for ACC, 26.7 (3.83) for LDW/LKA, 32.6 (4.92) for emergency braking, and 67.5 (5.14) for camera systems and smart mirrors, on a scale from 1 (*Extremely negative*) to 100 (*Extremely positive*). The reported means are shown in Figure 3.3.

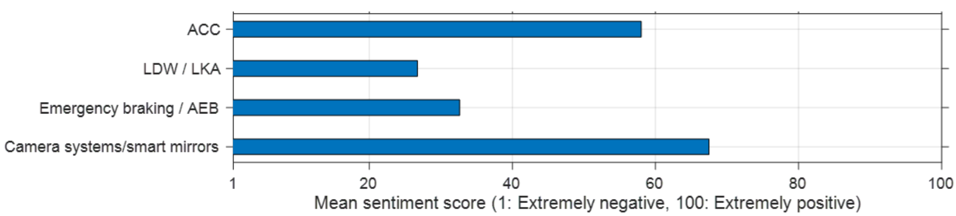


Figure 3.3: Sentiment scores for four categories of vehicle aids, as assessed by ChatGPT based on textbox comments. ACC: adaptive cruise control; LDW: lane departure warnings; LKA: lane keeping assistance; AEB: autonomous emergency braking.

3.4. Discussion

Truck drivers play a vital role in national distribution and international trade, yet face considerable challenges, with the rapid adoption of new technologies adding to these challenges (Gittleman & Monaco, 2020; Loske & Klumpp, 2021). However, comprehension of truck drivers' daily experiences has been limited. The current study aimed to fill this knowledge gap through a large-scale questionnaire from 3708 Dutch professional truck drivers. The responses to multiple-choice questions were statistically analyzed, while a large language model was used to analyze the responses to the open comment boxes.

The results revealed that Dutch professional drivers view the image of their profession as needing improvement, are hesitant to recommend it, and possess a neutral to slightly negative outlook. There was evident concern about decreased space on roads. When considering work pressure, compensation, and work-life balance, scores indicated moderate work pressure, high dissatisfaction with wages, challenges in maintaining a balance between work and personal life, and lack of support from transport organizations.

Factor analysis revealed two primary types of concerns among drivers: Work Pressure and Safety & Security Concerns. Work Pressure, characterized by high loadings on items like the impact of pressure on driving behavior and intimidation from employers, was more commonly reported by national drivers. Safety & Security Concerns, marked by high loadings on items like dealing with transport crime and finding decent parking, were more prevalent among international drivers. These results can be explained as crime primarily pertains to fuel or cargo theft when the vehicle is parked, in addition to instances of unauthorized migrants clandestinely boarding the truck (De Leeuw van Weenen et al., 2019; García & Insa, 2017; UK Parliament, 2015). Moreover, long-distance drivers more frequently work during night hours, which may contribute to a feeling of unsafety. Work pressure was more of an issue for the national (short-distance) drivers, which may be explained by the larger number of trips they have to complete, the busier daytime traffic conditions, or the more urban traffic environments they are exposed to, in typically less comfortable vehicles (Friswell & Williamson, 2013).

In addition, our research addressed the perception of technological systems, namely ADAS and the on-board computer. ACC was appreciated for its safety features and fuel-saving properties, but concerns exist due to incorrect interventions. LKA systems were often perceived as irritating due to frequent false alerts, leading some drivers to turn them off. Some respondents saw emergency braking technology as useful in preventing accidents, but various safety concerns were raised regarding false activations (see also Dreger et al., 2020; Grove et al., 2020). Camera systems and mirror technology received mixed reviews; while many respondents appreciated increased visibility and reduced blind spots (see also An et al., 2023; Ostermann et al., 2016), others pointed out that the substitution of conventional mirrors with digital cameras disrupts the reciprocal visual communication between the driver and other road users, and may lead to a disconnection from the surrounding environment. Finally, on-board computers were found to improve logistical efficiency and communication but also increased perceived work pressure due to real-time tracking, potential for additional work, and a sense of surveillance. Similar concerns apply to data-driven driver coaching. Although data recorded by onboard computers has been shown to be predictive of traffic incidents (e.g., Driessen et al., 2024), drivers may not readily accept driver monitoring systems. This reluctance could arise from drivers being unaware of the benefits or their discomfort with sharing their data with external parties (Picco et al., 2023).

These findings can be broadly interpreted in the context of automation disuse (Nordhoff & De Winter, 2023; Parasuraman & Riley, 1997): in general, drivers appeared to value systems that tangibly contribute to accident prevention and workload reduction, while demonstrating resistance towards less reliable systems, false alarms, and perceived intrusions of autonomy. The findings of our research can also be interpreted through the lens of Ivan Illich's concept of 'Tools for Conviviality', which advocates for technology that promotes autonomy and fruitful interaction (Hancock, 2019; Illich, 1973). While features such as ACC, AEB, and camera systems can increase driver autonomy and safety when working optimally, concerns about false activations, reliability, and a sense of intrusive surveillance represent a departure from conviviality.

The sentiment ratings revealed that LDW and emergency braking yielded the lowest scores. However, these results should be interpreted with caution, as there is a possibility that drivers might have confused AEB with ACC. In recent years, ADAS have progressed substantially, typically integrating a variety of subsystems (Rahman & Mekker, 2022; Souman et al., 2021), and their functionality may not always be clear to drivers (McDonald et al., 2018; Trimble et al., 2020). Also for the authors of the current study, it was occasionally challenging to accurately classify specific comments. For example, drivers frequently referred to the term 'distance sensor'. Technically, this is not an ADAS, but measurement equipment that is used in both ACC and AEB. This confusion may partially account for the low sentiment score for AEB, where false-positive braking interventions are typically ascribed to AEB, rather than ACC. Furthermore, for AEB, it is predominantly these false positives that drivers perceive, while the number of instances in which AEB averts accidents is logically low (Grove et al., 2017), since (near-)accidents are infrequent events. However, from a cost-benefit perspective, the AEB system might still be beneficial despite the low sentiment score, considering the substantial costs of accidents.

The acceptance of technology by drivers is essential, particularly in the context of the increasing mandating of technological systems in trucks. As of November 2015, EU regulations have made it compulsory for all new trucks to be equipped with AEB and LDW systems (Regulation 661/2009). From July 2022, new trucks are required to have additional systems, such as a blind spot information system, pedestrian/cyclist collision prevention, reversing detection, a driver availability monitoring system, and tire pressure monitoring. The mandate extends further in January 2026, when systems such as direct vision for vulnerable user protection, event data recorders, and advanced driver distraction warning systems will become obligatory (Regulation 2019/2144). As more technologies become mandatory, the need for such systems to be reliable and conducive to the driver is reinforced.

Several limitations must be considered with this study. One is that the questionnaire was conducted at the end of 2021. During the Covid-19 pandemic, truck drivers dealt with less social contact as amenities closed down, while social media sentiment analysis revealed that public appreciation for their work actually grew (Sperry et al., 2022).

Furthermore, it should be considered that ADAS sensors and algorithms continue to improve. While these improvements likely result in fewer false positives, there also exists the issue of human variability: false positive warnings in AEB and LDW may be inevitable considering that a threshold needs to be set for a critical time-to-collision or lateral deviation. According to the principles of signal-detection theory, this will involve a trade-off between timely warnings and false positives, as interpreted by the driver (e.g., Berge et al., 2024; Brookhuis & De Waard, 2003). This inescapable threshold could potentially explain why, despite many years of development, AEB and LDW systems are still perceived as irritating by drivers (e.g., Ayoub et al., 2022; Kidd et al., 2017). Arguably, a more fundamental consideration needs to be given to the usefulness of warning systems compared to systems that automatically maintain the lane or exert torque feedback on the steering wheel (De Winter et al., 2023; Roozendaal et al., 2021).

In this study, a large number of drivers were surveyed, which implies that the results are statistically precise. However, the results are not necessarily free of bias: it is possible that the mean values as shown in Table 3.1 are negatively skewed if primarily drivers who wished to complain completed the questionnaire, or if drivers exaggerated certain points in the hope that their responses would prompt a shift in national politics and business practices. In this context, it is useful to compare our results with questionnaires said to be nationally representative, specifically the National Employment Survey conducted by the Netherlands Organisation for Applied Scientific Research (TNO), Statistics Netherlands, and the Ministry of Social Affairs and Employment (Statistics Netherlands, 2023). In our questionnaire, there were two questions that were highly similar to questions in this nationally representative survey. Specifically, to the question *"Do you ever experience intimidation from your employer?"* (Q37), 11.9% of our respondents answered *Yes*, compared to 10.9% in the national survey who answered *Yes (occasionally, often, or very often)* to the question *"Can you indicate to what extent you have personally experienced intimidation by superiors or colleagues in the past 12 months?"* Another comparable question was Q31: *"Do you experience high work pressure?"*, to which 19.2% of our respondents answered *Yes - every day* and 34.3% *Yes - 1 or 2 times per week* (a total of 53.5%). In the nationally representative survey, 37.1% answered *Often* or *Always* to the question "Do you have to do a lot of work?". In summary, our results are in line with results from a representative sample of truck drivers in the Netherlands, suggesting no substantial bias in our questionnaire. However, it is worth noting that our open comment boxes were often left empty, with response rates ranging widely between questions (see Appendix B). It may be that drivers who wanted to suggest improvements in particular took the opportunity to fill in the open comment boxes, still introducing a form of bias.

Besides representativeness for the Dutch population, it is necessary to consider how our results relate to those of other countries. There are large national differences in road network density, road quality, accident risk, and the quality of organizations and operations. Despite this, certain factors concerning the well-being of drivers, such as stress, fatigue, and physical and mental health, recur both within Europe (Delhomme & Gheorghiu, 2021; Reiman et al., 2018; Useche et al., 2021) and on other continents (Hege et al., 2019; Jiang et al., 2017; Koul & Singh, 2022; Pritchard et al., 2023; Sabir et al., 2018; Wadley et al., 2020).

The impression that drivers left in our questionnaire was quite negative. They appeared pessimistic about the profession as a whole and found their salary to be mediocre. At the same time, respondents were satisfied with their own employer, and the majority did not experience high work pressure, with 36.7% of respondents reporting no high work pressure and 19.2% indicating high work pressure on a daily basis. This is also evident from the aforementioned national survey, where other professional groups such as elementary school teachers, managers, cooks, lawyers, doctors, directors, social workers, and caregivers reported much higher work pressure than truck drivers (Statistics Netherlands, 2023). Possible explanations are that, even though truck drivers have many grievances about their field, 'being on the road' is a job that offers a certain level of satisfaction (Kishore Bhoopalam et al., 2021; Ruiner & Klumpp, 2022). It is also possible that truck drivers experience pressure, but do not perceive or express it as such due to their hardship and stoicism (Johnson et al., 2021). Additionally, while truck drivers may not have to work hard in physical terms, their work scheduling is highly dictated as compared to some other professions like directors, scientists, and advisors. The literature concurs that flexibility and autonomy over work hours can influence job satisfaction; a meta-analysis by Shifrin and Michel (2022) highlights the positive impact of flexible work arrangements on overall job stress levels. Work-related pressures, often tied to truck driving accidents, can stem from various factors such as supervisor pressure, inadequate training, and unsupportive management (Anderson et al., 2017; Delhomme & Gheorghiu, 2021; Reiman et al., 2018). Further, loading/off-loading site culture (Friswell & Williamson, 2019; Grytnes et al., 2016; Reiman et al., 2018), as well as other road users' behavior (Gray, 2019; Häkkänen & Summala, 2001; Huang et al., 2005; Semeijn et al., 2019; Williams et al., 2017), can be a source of stress.

Beyond the issue of representativeness, it is important to also monitor the quality of the responses, that is, whether the questionnaire appears to have been completed sincerely. Our impression is that the quality of the responses was high compared to other questionnaires that seem to be plagued by acquiescence bias (for discussions, see De Winter & Nordhoff, 2022; Krosnick, 1999; Podsakoff et al., 2003). An illustration of the high quality of responses is that only 3 of the 3708 respondents (0.08%) rated the quality of roads in the Netherlands (Q26) as very bad. If there were mindless responses, the distribution of responses would be more uniform.

A noteworthy aspect of our study is that the text analysis was done automatically. Our observation is that the summaries and sentiment scores correspond to how we ourselves would summarize and rate the truck drivers' comments. This statement is supported by a growing body of literature demonstrating that ChatGPT performs well in linguistic tasks, such as answering exam questions, labeling tweets and reviews, and analysis of sentiment (De Winter, 2023; Gilardi et al., 2023; Nori et al., 2023; OpenAI, 2023; Törnberg, 2023; Zhang et al., 2023). The fact that texts were submitted to ChatGPT in Dutch rather than English is not necessarily a problem, as shown in several studies (Lai et al., 2023; Tan et al., 2023). We agree with Mellon et al. (2024) that the availability of large language models makes the use of open-ended questions in future questionnaires more attractive.

Nevertheless, there are some limitations to using ChatGPT. While ChatGPT is proficient in summarization and sentiment analysis (e.g., Bubeck et al., 2023; Tabone & De Winter, 2023; Zhao et al., 2023), it may lack domain-specific expertise (Li et al., 2023). Moreover, its output can be sensitive to the specific wording of the prompt (Bubeck et al., 2023). For these reasons, we undertook a manual classification of individual comments into the four ADAS categories (Q30). This approach ensured the sentiment scores were directly relevant to the specific ADAS under evaluation.

3.5. Conclusion

This study provided new insights into the experiences and perceptions of Dutch professional truck drivers. The findings illustrate the need for improved working conditions and support from transport organizations, as well as greater attention to safety and security concerns, especially among international drivers.

What policy recommendations arise from this research? Truck drivers often indicate that they should receive better financial compensation. However, when we consider the entirety of this work, including Appendix B, it becomes clear that the drivers are not just concerned with monetary incentives but also with recognition and respect for their profession. The current study offers various starting points that can help improve the welfare and status of drivers, including better sanitary and parking facilities. Additionally, it is recommended to act at an international level against fuel theft, break-ins, and other forms of transport crime. In the development of new technology, the minimization of perceived intrusiveness should be a key design criterion, both in a direct sense (unnecessary automated braking interventions and alarms) and in an indirect sense (perceived intrusions in work flexibility and autonomy). Although truck drivers appreciate technologies that improve safety and efficiency, the feeling of autonomy being compromised indicates a need for less meddlesome technology.

Data availability statement

The questionnaire items, response options, response frequency distributions, and factor analysis results can be found online at: https://doi.org/10.4121/577c120a-b5bb-4ba5-93b8-6143759d0249.

Ethics statement

The study was conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants in accordance with the local legislation and institutional requirements. Approval for analysis of the questionnaire data was provided by the TU Delft Human Research Ethics Committee (approval number 3013).

Conflicts of Interest

Aschwin Cannoo is CEO of Transporteffect BV and Chauffeurnieuws. He was not involved in the current analysis and interpretation of the data. The other authors declare no conflict of interest.

References

An, S., Lee, S., Park, G., Lee, H., Son, M., & Beck, D. (2023). User perception and ergonomic display layout design of truck camera monitor system. SSRN. https://doi.org/10.2139/ssrn.4334589

Anderson, N. J., Smith, C. K., & Byrd, J. L. (2017). Work-related injury factors and safety climate perception in truck drivers. American Journal of Industrial Medicine, 60, 711–723. https://doi.org/10.1002/ajim.22737

Ayoub, J., Wang, Z., Li, M., Guo, H., Sherony, R., Bao, S., & Zhou, F. (2022). Cause-and-effect analysis of ADAS: A comparison study between literature review and complaint data. Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 139–149, Seoul, Republic of Korea. https://doi.org/10.1145/3543174.3547117

Bachmann, L. H., Lichtenstein, B., St. Lawrence, J. S., Murray, M., Russell, G. B., & Hook, E. W., III (2018). Health risks of American long distance truckers: Results from a multi-site assessment. Journal of Occupational and Environmental Medicine, 60, e349. https://doi.org/10.1097/JOM.0000000000001319

Bazilinskyy, P., Larsson, P., Johansson, E., & De Winter, J. C. F. (2019). Continuous auditory feedback on the status of adaptive cruise control, lane deviation, and time headway: An acceptable support for truck drivers? Acoustical Science and Technology, 40, 382–390. https://doi.org/10.1250/ast.40.382

Belzer, M. H. (2018). Work-stress factors associated with truck crashes: An exploratory analysis. The Economic and Labour Relations Review, 29, 289–307. https://doi.org/10.1177/1035304618781654

Berge, S. H., De Winter, J., Feng, Y., & Hagenzieker, M. (2024). Phantom braking in automated vehicles: A theoretical outline and cycling simulator demonstration. 15th International Conference on Applied Human Factors and Ergonomics, Nice, France.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3, 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhuis, K., & De Waard, D. (2003). On the assessment of criteria for driver impairment; In search of the golden yardstick for driving performance. Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, University of Iowa. https://doi.org/10.17077/drivingassessment.1081

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Chen, G. X., Sieber, W. K., Collins, J. W., Hitchcock, E. M., Lincoln, J. E., Pratt, S. G., & Sweeney, M. H. (2021). Truck driver reported unrealistically tight delivery schedules linked to their opinions of maximum speed limits and hours-of-service rules and their compliance with these safety laws and regulations. Safety Science, 133, 105003. https://doi.org/10.1016/j.ssci.2020.105003

Chen, G. X., Sieber, W. K., Lincoln, J. E., Birdsey, J., Hitchcock, E. M., Nakata, A., Robinson, C. F., Collins, J. W., & Sweeney, M. H. (2015). NIOSH national survey of long-haul truck drivers: injury and safety. Accident Analysis & Prevention, 85, 66–72. https://doi.org/10.1016/j.aap.2015.09.001

De Leeuw van Weenen, R., Newton, S., Menist, M., Maas, F., Penasse, D., Nielsen, M., Halatsis, A., Männistö, T., Stamos, I., & Ruschin, P. P. (2019). Study on safe and secure parking places for trucks (Technical Report). European Commission. https://doi.org/10.2832/067535

Delhomme, P., & Gheorghiu, A. (2021). Perceived stress, mental health, organizational factors, and self-reported risky driving behaviors among truck drivers circulating in France. Journal of Safety Research, 79, 341–351. https://doi.org/10.1016/j.jsr.2021.10.001

De Winter, J., & Nordhoff, S. (2022). Acceptance of conditionally automated cars: Just one factor? Transportation Research Interdisciplinary Perspectives, 15, 100645. https://doi.org/10.1016/j.trip.2022.100645

De Winter, J. C. F. (2023). Can ChatGPT pass high school exams on English language comprehension? International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-023-00372-z

De Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. Multivariate Behavioral Research, 44, 147–181. https://doi.org/10.1080/00273170902794206

De Winter, J. C. F., Petermeijer, S. M., & Abbink, D. A. (2023). Shared control versus traded control in driving: A debate around automation pitfalls. Ergonomics, 66, 1494–1520. https://doi.org/10.1080/00140139.2022.2153175

Dos Reis, L. A. P., Costa, C. D. D., Rodrigues, D. S., & De Alcântara, K. C. (2017). Obesity, hypertension and diabetes among truck drivers in the middle-west, Brazil. Bioscience Journal, 33. https://doi.org/10.14393/BJ-v33n2-34285

Dreger, F. A., De Winter, J. C. F., & Happee, R. (2020). How do drivers merge heavy goods vehicles onto freeways? A semi-structured interview unveiling needs for communication and support. Cognition, Technology & Work, 22, 825–842. https://doi.org/10.1007/s10111-019-00601-3

Driessen, T., Dodou, D., De Waard, D., & De Winter, J. C. F. (2024). Predicting damage incidents, fines, and fuel consumption from truck driver data: A study from the Netherlands. Transportation Research Record. https://doi.org/10.1177/03611981231211897

Dubljević, V., Douglas, S., Milojevich, J., Ajmeri, N., Bauer, W. A., List, G., & Singh, M. P. (2023). Moral and social ramifications of autonomous vehicles: A qualitative study of the perceptions of professional drivers. Behaviour & Information Technology, 42, 1271–1278. https://doi.org/10.1080/0144929X.2022.2070078

European Commission. (2022). Mobility Package I. https://transport.ec.europa.eu/transport-modes/road/mobility-package-i_en

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4, 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fagerlönn, J. (2011). Urgent alarms in trucks: Effects on annoyance and subsequent driving performance. IET Intelligent Transport Systems, 5, 252–258. https://doi.org/10.1049/iet-its.2010.0165

Fors, C., Kircher, K., & Ahlström, C. (2015). Interface design of eco-driving support systems – Truck drivers' preferences and behavioural compliance. Transportation Research Part C: Emerging Technologies, 58, 706–720. https://doi.org/10.1016/j.trc.2015.03.035

Friswell, R., & Williamson, A. (2013). Comparison of the fatigue experiences of short haul light and long distance heavy vehicle drivers. Safety Science, 57, 203–213. https://doi.org/10.1016/j.ssci.2013.02.014

Friswell, R., & Williamson, A. (2019). Management of heavy truck driver queuing and waiting for loading and unloading at road transport customers' depots. Safety Science, 120, 194–205. https://doi.org/10.1016/j.ssci.2019.06.039

García, B. G., & Insa, E. M. (2017). MindFuel; Product-Service System implementation for truck fueling optimization business case (MSc thesis). Chalmers University of Technology. https://odr.chalmers.se/items/eedb09df-2372-4372-8b15-0627a6e891b7

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120, e2305016120. https://doi.org/10.1073/pnas.2305016120

Gittleman, M., & Monaco, K. (2020). Truck-driving jobs: Are they headed for rapid elimination? ILR Review, 73, 3–24. https://doi.org/10.1177/0019793919858079

Gouribhatla, R., & Pulugurtha, S. S. (2022). Drivers' behavior when driving vehicles with or without advanced driver assistance systems: A driver simulator-based study. Transportation Research Interdisciplinary Perspectives, 13, 100545. https://doi.org/10.1016/j.trip.2022.100545

Gray, G. (2019). A bird's eye view of driving safety culture: Truck drivers' perceptions of unsafe driving behaviors near their trucks. Work, 64, 187–194. https://doi.org/10.3233/WOR-192985

Grove, K., Atwood, J., Blanco, M., Krum, A., & Hanowski, R. (2017). Field study of heavy vehicle crash avoidance system performance. SAE International Journal of Transportation Safety, 5, 1–12. https://doi.org/10.4271/2016-01-8011

Grove, K., Camden, M. C., Krum, A., & Hanowski, R. J. (2020). Research and testing to accelerate voluntary adoption of automatic emergency braking (AEB) on commercial vehicles (Report No. FMCSA-RRT-18-013). Washington DC: Federal Motor Carrier Safety Administration, U.S. Department of Transportation. https://rosap.ntl.bts.gov/view/dot/49335

Grytnes, R., Shibuya, H., Dyreborg, J., Grøn, S., & Cleal, B. (2016). Too individualistic for safety culture? Non-traffic related work safety among heavy goods vehicle drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 40, 145–155. https://doi.org/10.1016/j.trf.2016.04.012

Häkkänen, H., & Summala, H. (2001). Fatal traffic accidents among trailer truck drivers and accident causes as viewed by other truck drivers. Accident Analysis & Prevention, 33, 187–196. https://doi.org/10.1016/S0001-4575(00)00030-0

Hancock, P. A. (2019). Some pitfalls in the promises of automated and autonomous vehicles. Ergonomics, 62, 479–495. https://doi.org/10.1080/00140139.2018.1498136

Hege, A., Lemke, M. K., Apostolopoulos, Y., & Sönmez, S. (2019). The impact of work organization, job stress, and sleep on the health behaviors and outcomes of U.S. long-haul truck drivers. Health Education & Behavior, 46, 626–636. https://doi.org/10.1177/1090198119826232

Hickman, J. S., Guo, F., Camden, M. C., Hanowski, R. J., Medina, A., & Mabry, J. E. (2015). Efficacy of roll stability control and lane departure warning systems using carrier-collected data. Journal of Safety Research, 52, 59–63. https://doi.org/10.1016/j.jsr.2014.12.004

Huang, Y.-H., Roetting, M., McDevitt, J. R., Melton, D., & Smith, G. S. (2005). Feedback by technology: Attitudes and opinions of truck drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 8, 277–297. https://doi.org/10.1016/j.trf.2005.04.005

Illich, I. (1973). Tools for conviviality. New York: Harper & Row.

Jaller, M., Otero-Palencia, C., & D'Agostino, M. (2022). Jobs and automated freight transportation: How automation affects the freight industry and what to do about it (Report No. NCST-UCD-WP-22-42). National Center for Sustainable Transportation. https://doi.org/10.7922/G2SX6BHW

Jiang, C., Lu, L., & Lu, J. J. (2017). Socioeconomic factors affecting the job satisfaction levels of self-employed container truck drivers: a case study from Shanghai Port. Maritime Policy & Management, 44, 641–656. https://doi.org/10.1080/03088839.2017.1326182

Ji-Hyland, C., & Allen, D. (2022). What do professional drivers think about their profession? An examination of factors contributing to the driver shortage. International Journal of Logistics Research and Applications, 25, 231–246. https://doi.org/10.1080/13675567.2020.1821623

Johnson, J. K., Vingilis, E., & Terry, A. L. (2021). Qualitative study of long-haul truck drivers' health and healthcare experiences. Journal of Occupational and Environmental Medicine, 63, 230–237. https://doi.org/10.1097/JOM.0000000000002107

Kidd, D. G., Cicchino, J. B., Reagan, I. J., & Kerfoot, L. B. (2017). Driver trust in five driver assistance technologies following real-world use in four production vehicles. Traffic Injury Prevention, 18, S44–S50. https://doi.org/10.1080/15389588.2017.1297532

Kishore Bhoopalam, A., Van den Berg, R., Agatz, N., & Chorus, C. (2021). The long road to automated trucking: Insights from driver focus groups. SSRN. https://doi.org/10.2139/ssrn.3779469

Kitto, K., Manly, C. A., Ferguson, R., & Poquet, O. (2023). Towards more replicable content analysis for learning analytics. Proceedings of Learning Analytics and Knowledge 2023, Arlington, TX. https://doi.org/10.1145/3576050.3576096

Klumpp, M. (2018). Automation and artificial intelligence in business logistics systems: Human reactions and collaboration requirements. International Journal of Logistics Research and Applications, 21, 224–242. https://doi.org/10.1080/13675567.2017.1384451

Koul, S., & Singh, A. (2022). Mapping vulnerabilities of Indian long-haul truck drivers. Proceedings of the 2022 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), Delhi, India. https://doi.org/10.1109/SOLI57430.2022.10294623

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage Publications.

Krosnick, J. A. (1999). Survey research. Annual Review of Psychology, 50, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Laban, P., Kryściński, W., Agarwal, D., Fabbri, A. R., Xiong, C., Joty, S., & Wu, C.-S. (2023). SUMMEDITS: Measuring LLM ability at factual reasoning through the lens of summarization. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9662–9676). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.600

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 13171–13189). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, 408–422, Singapore. https://doi.org/10.18653/v1/2023.emnlp-industry.39

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv. https://doi.org/10.48550/arXiv.2304.03439

Loske, D., & Klumpp, M. (2021). Intelligent and efficient? An empirical analysis of human–AI collaboration for truck drivers in retail logistics. The International Journal of Logistics Management, 32, 1356–1383. https://doi.org/10.1108/IJLM-03-2020-0149

McDonald, A., Carney, C., & McGehee, D. V. (2018). Vehicle owners' experiences with and reactions to advanced driver assistance systems. AAA Foundation for Traffic Safety. http://aaafoundation.org/wp-content/uploads/2018/09/VehicleOwnersExperiencesWithADAS_TechnicalReport.pdf

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. Research & Politics, 11. https://doi.org/10.1177/20531680241231468

Morton, R., Richards, D., Dunn, N., & Coulton, P. (2019). Questioning the social and ethical implications of autonomous vehicle technologies on professional drivers. The Design Journal, 22, 2061–2071. https://doi.org/10.1080/14606925.2019.1594930

Nordhoff, S., & De Winter, J. C. F. (2023). Why do drivers and automation disengage the automation? Results from a study among Tesla users. arXiv. https://doi.org/10.48550/arXiv.2309.10440

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv. https://doi.org/10.48550/arXiv.2303.13375

OpenAI. (2023). GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Ostermann, B., Koppenborg, M., Staffel, M., Paridon, H., Hofmann, S., Ahrends-Krossner, G., Fischer, M., Schemel, A., Wilbig, E., & Heider, G. (2016). Kamera-monitor-systeme (KMS) zur Vermeidung von Abbiegeunfällen. Kriterien für die Eignung von Kamera-Monitor Systemen in LKW zur Vermeidung von Rechtsabbiegeunfällen [Camera Monitor Systems (CMS) to prevent turning accidents. Criteria for the suitability of Camera Monitor Systems in trucks to prevent right-turn accidents]. BG Verkehr. https://www.bg-verkehr.de/redaktion/medien-und-downloads/broschueren/branchen/gueterkraftverkehr/bgverkehr_kms_a4_studie_komplett.pdf

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39, 230–253. https://doi.org/10.1518/001872097778543886

Picco, A., Stuiver, A., De Winter, J. C. F., & De Waard, D. (2023). The use of monitoring and feedback devices in driving: An assessment of acceptability and its key determinants. Transportation Research Part F: Traffic Psychology and Behaviour, 92, 1–14. https://doi.org/10.1016/j.trf.2022.10.021

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88, 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pritchard, E., Van Vreden, C., Xia, T., Newnam, S., Collie, A., Lubman, D. I., De Almeida Neto, A., & Iles, R. (2023). Impact of work and coping factors on mental health: Australian truck drivers' perspective. BMC Public Health, 23, 1090. https://doi.org/10.1186/s12889-023-15877-4

Pu, X., Gao, M., & Wan, X. (2023). Summarization is (almost) dead. arXiv. https://doi.org/10.48550/arXiv.2309.09558

Rahman, M. A., & Mekker, M. (2022). Development of educational materials for the public and first responders on the limitations of Advanced Driving Assistance Systems (Report No. UT-22.11). Salt Lake City, UT: Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/64860

Rahman, M. M., Lesch, M. F., Horrey, W. J., & Strawderman, L. (2017). Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. Accident Analysis & Prevention, 108, 361–373. https://doi.org/10.1016/j.aap.2017.09.011

Regulation 661/2009. Regulation (EC) No 661/2009 of the European Parliament and of the Council of 13 July 2009 concerning type-approval requirements for the general safety of motor vehicles, their trailers and systems, components and separate technical units intended therefor. https://eur-lex.europa.eu/eli/reg/2009/661/oj

Regulation 2019/2144. Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users. https://eur-lex.europa.eu/eli/reg/2019/2144/oj

Reiman, A., Forsman, M., Målqvist, I., Parmsund, M., & Lindahl Norberg, A. (2018). Risk factors contributing to truck drivers' non-driving occupational accidents. International Journal of Physical Distribution & Logistics Management, 48, 183–199. https://doi.org/10.1108/IJPDLM-06-2017-0216

Roberts, K., Dowell, A., & Nie, J.-B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. BMC Medical Research Methodology, 19, 66. https://doi.org/10.1186/s12874-019-0707-y

Roozendaal, J., Johansson, E., De Winter, J., Abbink, D., & Petermeijer, S. (2021). Haptic lane-keeping assistance for truck driving: A test track study. Human Factors, 63, 1380–1395. https://doi.org/10.1177/0018720820928622

Rossi, R., Gastaldi, M., Biondi, F., Orsini, F., De Cet, G., & Mulatti, C. (2020). A driving simulator study exploring the effect of different mental models on ADAS system effectiveness. In L. de Paolis & P. Bourdot (Eds.), Augmented reality, virtual reality, and computer graphics. AVR 2020 (pp. 102–113). Cham: Springer. https://doi.org/10.1007/978-3-030-58465-8_7

Ruiner, C., & Klumpp, M. (2022). Autonomy and new modes of control in digital work contexts – a mixed-methods study of driving professions in food logistics. Employee Relations: The International Journal, 44, 890–912. https://doi.org/10.1108/ER-04-2021-0139

Sabir, A. A., Shahru, A., Isha, N. B., Langove, N., & Javaid, M. U. (2018). An integrative approach to study drivers perceived stress on aberrant driving behaviour: the mediation role of driver fatigue. SHS Web of Conferences, 56, 02009. https://doi.org/10.1051/shsconf/20185602009

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with construct-specific response options. Survey Research Methods, 4, 61–79. https://doi.org/10.18148/srm/2010.v4i1.2682

Schuster, A. M., Agrawal, S., Britt, N., Sperry, D., Van Fossen, J. A., Wang, S., Mack, E. A., Liberman, J., & Cotten, S. R. (2023). Will automated vehicles solve the truck driver shortages? Perspectives from the trucking industry. Technology in Society, 74, 102313. https://doi.org/10.1016/j.techsoc.2023.102313

Sector Institute Transport and Logistics. (2021). Arbeidsmarktrapportage beroepsgoederenvervoer over de weg en logistiek: Instroom en inzetbaarheid [Labour market report for professional goods transport by road and logistics: Inflow and employability]. https://www.stl.nl/STL/media/STLMedia/Trends%20en%20marktontwikkelingen/Arbeidsmarktrapportage-2021_lowress.pdf

Semeijn, J., De Waard, B., Lambrechts, W., & Semeijn, J. (2019). Burning rubber or burning out? The influence of role stressors on burnout among truck drivers. Logistics, 3, 6. https://doi.org/10.3390/logistics3010006

Shattell, M., Apostolopoulos, Y., Sönmez, S., & Griffin, M. (2010). Occupational stressors and the mental health of truckers. Issues in Mental Health Nursing, 31, 561–568. https://doi.org/10.3109/01612840.2010.488783

Shifrin, N. V., & Michel, J. S. (2022). Flexible work arrangements and employee health: A meta-analytic review. Work & Stress, 36, 60–85. https://doi.org/10.1080/02678373.2021.1936287

Shin, D. S., & Jeong, B. Y. (2020). Relationship between negative work situation, work-family conflict, sleep-related problems, and job dissatisfaction in the truck drivers. Sustainability, 12, 8114. https://doi.org/10.3390/su12198114

Sieber, W. K., Robinson, C. F., Birdsey, J., Chen, G. X., Hitchcock, E. M., Lincoln, J. E., Nakata, A., & Sweeney, M. H. (2014). Obesity and other risk factors: The national survey of U.S. long-haul truck driver health and injury. American Journal of Industrial Medicine, 57, 615–626. https://doi.org/10.1002/ajim.22293

Souman, J., Van Weperen, M., Hogema, J., Hoedemaeker, M., Westerhuis, F., Stuiver, A., & De Waard, D. (2021). Human factors guidelines report 3: Use and mental models (Report No. R12165). TNO Human Factors Research Institute. https://hdl.handle.net/11370/9b42ef9c-d063-4252-9b00-f483e84f78a3

Sperry, D., Schuster, A. M., Cotten, S. R., Agrawal, S., Mack, E., Britt, N., & Liberman, J. (2022). Trucking in the era of COVID-19. American Behavioral Scientist. https://doi.org/10.1177/00027642211066039

Statistics Netherlands. (2021). The Netherlands in numbers: Which jobs often cause a poor work-life balance? (p. 29). https://www.cbs.nl/-/media/_pdf/2021/48/nederland-in-cijfers_2021_boekje.pdf

Statistics Netherlands. (2023). Psycho-sociale arbeidsbelasting werknemers; beroep (data for 2022). https://www.cbs.nl/nl-nl/cijfers/detail/84436NED

Svenson, A. L., Grove, K., Atwood, J., & Blanco, M. (2017). Using naturalistic data to evaluate heavy vehicle crash avoidance systems performance. Proceedings of the 25th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration, Detroit, MI. https://www-esv.nhtsa.dot.gov/Proceedings/25/25ESV-000231.pdf

Tabone, W., & De Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. Royal Society Open Science, 10, 231053. https://doi.org/10.1098/rsos.231053

Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., & Qi, G. (2023). Evaluation of ChatGPT as a question answering system for answering complex questions. arXiv. https://doi.org/10.48550/arXiv.2303.07992

Tang, R., Zhang, X., Ma, X., Lin, J., & Ture, F. (2023). Found in the middle: Permutation self-consistency improves listwise ranking in large language models. arXiv. https://doi.org/10.48550/arXiv.2310.07712

Teoh, E. R. (2021). Effectiveness of front crash prevention systems in reducing large truck real-world crash rates. Traffic Injury Prevention, 22, 284–289. https://doi.org/10.1080/15389588.2021.1893700

Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. arXiv. https://doi.org/10.48550/arXiv.2304.06588

Transporteffect. (2021). De vuist. Grootste enquête ooit onder de beroepschauffeurs. Resultaten enquête [The Fist. The largest survey ever among professional drivers. Survey results]. https://transporteffect.com/wp-content/uploads/2021/12/Enquete-Transporteffect-Chauffeursnieuws-1.pdf

Trimble, T. E., Baker, S. A., Russell, S. M., & Tidwell, S. (2020). Consumer driving automation system education: A learning and retention assessment (Report No. 20-UO-089). Blacksburg, VI: Virginia Tech Transportation Institute. http://hdl.handle.net/10919/99302

UK Parliament. (2015). Migrants in Calais. https://publications.parliament.uk/pa/cm201415/cmselect/cmhaff/902/90203.htm

Useche, S. A., Alonso, F., Cendales, B., & Llamazares, J. (2021). More than just "stressful"? Testing the mediating role of fatigue on the relationship between job stress and occupational crashes of long-haul truck drivers. Psychology Research and Behavior Management, 14, 1211–1221. https://doi.org/10.2147/PRBM.S305687

Van Fossen, J. A., Chang, C.-H., Ford, J. K., Mack, E. A., & Cotten, S. R. (2023). Identifying alternative occupations for truck drivers displaced due to autonomous vehicles by leveraging the O*NET database. American Behavioral Scientist, 67, 1693–1715. https://doi.org/10.1177/00027642221127239

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Pontes Reis, E., Seehofnerova, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. S. (2024). Clinical text summarization: Adapting large language models can outperform human experts. Nature Medicine. https://doi.org/10.1038/s41591-024-02855-5

Wadley, A. L., Iacovides, S., Roche, J., Scheuermaier, K., Venter, W. D. F., Vos, A. G., & Lalla-Edward, S. T. (2020). Working nights and lower leisure-time physical activity associate with chronic pain in southern African long-distance truck drivers: A cross-sectional study. PLOS ONE, 15, e0243366. https://doi.org/10.1371/journal.pone.0243366

Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv. https://doi.org/10.48550/arXiv.2304.04339

Wijngaards, I., Hendriks, M., & Burger, M. J. (2019). Steering towards happiness: An experience sampling study on the determinants of happiness of truck drivers. Transportation Research Part A: Policy and Practice, 128, 131–148. https://doi.org/10.1016/j.tra.2019.07.017

Williams, D. F., Jr., Thomas, S. P., & Liao-Troth, S. (2017). The truck driver experience: Identifying psychological stressors from the voice of the driver. Transportation Journal, 56, 54–76. https://doi.org/10.5325/transportationj.56.1.0054

Zhang, J., Chen, Y., Liu, C., Niu, N., & Wang, Y. (2023). Empirical evaluation of ChatGPT on requirements information retrieval under zero-shot setting. Proceedings of the 2023 International Conference on Intelligent Computing and Next Generation Networks (ICNGN), Hangzhou, China. https://doi.org/10.1109/ICNGN59831.2023.10396810

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A survey of large language models. arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. Computational Statistics & Data Analysis, 51, 918–930. https://doi.org/10.1016/j.csda.2005.09.010

Appendix A: Vehicle Aids Comments by Category

Table 3.4 contains the responses to Q30 (vehicle aids) which were identified to contain comments about assistance systems. The responses can appear in more than one category.

Table 3.4: All responses by assigned assistance system. The responses have been translated from Dutch using GPT-4 and manually inspected for accuracy.

	Adaptive cruise control (ACC)
1.	"Not all tools are suitable. During cutting, short overtaking, or red LEDs (infrared) from the matrix, the ACC sometimes brakes hard with all the consequences. Here too, everyone should follow the rules with keeping distance and merging/exiting in the right way/on time (accelerate), stay behind the truck and not pass it at the last moment; signs are already at 1200/600/300 meters."
2.	"I often drive on ACC, a great invention! …"
3.	"ACC works reasonably, but I'm now driving a DAF that reacts to viaducts and the portals. Then it's a disastrous system."
4.	"You quickly become comfortable with ACC, thinking the vehicle will handle it."
5.	"It will only become cheaper when every truck is equipped with this; currently still dangerous, especially with adaptive cv …"
6.	"Technology is very advanced these days.. On the truck really sanctified especially acc"
7.	"ACC is not workable"
8.	"ACC is dangerous"
9.	"Adaptive control sometimes intervenes unnecessarily"
10.	"I don't reject everything, the old-fashioned cruise control is a blessing. But nowadays you are more of a 'transport supervisor' than an independent thinking driver. No wonder you then start doing 'wrong' things, purely out of boredom. This time needs to be bridged until no drivers are needed at all. But for now, I don't find it becoming any more fun."
11.	"ACC works perfectly …"
12.	"Active cruise control is a good tool …"
13.	"Distance holder works well, luxury cars keep creeping in between and truck brakes automatically creating a yo-yo effect"
14.	"ACC is a great tool."
15.	"Although I think an eye for keeping distance that doesn't work in the fog is a missed opportunity. That's exactly a moment when you need an extra tool."
16.	"I am in favor of aids like ACC … but there really should not be too many of them or it will be distracting."
17.	"Maybe automatic distance keeping but for everyone and blocking the phone. Really can't!!!!"
18.	"Adaptive cruise control …"

19. "… ACC encourages you to pick up your mobile. And all that touchscreen stuff only takes your eyes off the road. Just give me press and twist buttons. The more stuff on a car the less alert people are. If something suddenly happens, people no longer know how to intervene."
20. "ACC mandatory …"
21. "Adaptive Cruise … ideal indeed!"
22. "… That automatic distance keeping is an improvement."
23. "Definitely adaptive cruise control"
24. "Some aids also react to other situations like adaptive cruise control that also reacts to viaducts traffic signs and then it makes an emergency stop out of nowhere also extremely dangerous"
25. "Some aids also respond to other situations such as adaptive cruise control, which also reacts to traffic signs on viaducts and then makes an emergency stop out of nowhere, also life-threatening."
26. "Wouldn't want to miss the ACC … anymore!"
27. "… adjustable speed limiter is something I use a lot,"
28. "Adaptive cruise control does not belong in a truck. There are too many car drivers who cut in front of a truck at the last moment to take an exit or brake unnecessarily, causing such a system to react too abruptly. A start-stop system also does not belong in a truck."
29. "Distance keeper is good …"
30. "… enforce adaptive control at 75 meters …"
31. "active cruise control sufficient distance not a few meters"
32. "Adaptive cruise control has pros and cons. When a passer-by suddenly flashes in front of you to take the exit and your car makes an emergency stop, it's not always funny."
33. "You're either a driver or you're not. I don't like these new safety systems at all like … ACC. … I love the old-fashioned work, shifting gears myself and keeping the vehicle under control myself! And not a computer or something!"
34. "The ACC is a good example, but sometimes it reacts too well, causing the truck to brake unnecessarily from time to time. And with other vehicles cutting across at too short a distance, it's a matter of being careful."
35. "The ACC is great … "
36. "ACC is not safer than CC, ACC causes people to be less alert."
37. "Distance meter on Volvo is top."
38. "… Active cruise control is a good invention."
39. "I have the automatic distance and brake system, but it's more annoying than functional. Partly because of people who fly braking from the 3rd lane to the right to take the exit."
40. "… The Actros mp5 can largely drive itself on the highway, in a sense this certainly promotes more inattentiveness. As you gain more trust in the system with each kilometer and driving also starts to get more boring. Then people are quicker to pick up the phone."
41. "ACC works well when all trucks are equipped with it."

42. “Adaptive cruise is great …”
43. “The ACC is a very good tool for road safety and in my view also saves fuel.”
44. “Ban/abolish cruise control …”
45. “ACC in combination with cruise control is a great tool.”
46. “I drive with ADR and we have many aids, and especially keeping distance works well.”
47. “… following systems like automatic distance keeping also make lazy. It is of course safe for rear-end collisions.”
48. “ACC is handy but in rush hour you are almost standing still as traffic keeps coming in …”
49. “Is safer, especially the ACC, I think it’s absolutely great.”
50. “… Distance cruise control also not conducive to alertness, when you are allowed to overtake, people don’t drive behind each other like elephants and you don’t need it.”
51. “The distance sensor … Yes, GREAT!”
52. “Had a truck with a distance keeper, etc. Sold it and bought an older one without all that stuff on it. I notice that I’m much more alert and involved in my work.”
53. “Certain aids are fantastic, but some also promote inattentiveness such as ACC.”
54. “Only the distance keeper of Volvo works well.”
55. “Adapted cruise control, respect for each other.”
56. “Distance control systems … work well, but unfortunately not every truck has it yet, so you still get them crawling in front of you and your car starts braking hard.”
57. “As an aid, I actually find all options quite nice, I just find that in most cases too much is relied on, I sometimes catch myself letting it run with the flow on the ACC, which makes you a little less attentive I think, the distance is more than sufficient, but still.”
58. “The ACC also reacts to the wrong things like traffic signs.”
59. “Something like adaptive cruise control is a great aid.”
60. “Automatic distance keeping is good.”
61. “ACC; is still far from perfect, own truck performs unnecessary emergency braking at a bridge. Sensors to stay in the lane work fine.”
62. “ACC for example works beautifully as long as those cars don’t drive at 80 between the trucks and don’t maintain a steady speed, very irritating.”
63. “That adaptive cruise control is a lousy system, it brakes when it really doesn’t need to. Consumes a lot of fuel and causes dangerous situations behind you because you don’t notice the truck braking.”
64. “Recently got … adaptive cruise control, I find it ideal, you can focus more on the actual driving itself.”
65. “Certainly, that system which brakes very quickly if someone crosses in front of you or merges.”
66. “I find that the distance keeper is in principle quite nice to keep distance. But I do find that we become lazy and pay less attention. On the road.”
67. “ACC, my employer … has taken it off again. Not for safety but as they say for fuel consumption.”

68. “I am not satisfied with … ACC, because car drivers pest test brake against trucks.”
69. “You have to get used to aids like ACC. How the car then reacts when someone comes in between. There is also a difference in truck brand.”
70. “Distance keeper is nothing on busy roads and in traffic jams.”
71. “Have had ACC for 3 years. I find it a delight.”
72. “The ”assist systems” are the first thing I turn off before I drive. Very dangerous (think of the distance meter that just makes a mini emergency stop for a matrix sign or viaduct on the highway) and highly annoying and disturbing to have to listen to beeping for 10 hours.”
73. “If there is little more to do, such as automatic cruise control, there is a high chance that drivers will pay much less attention to the road.”
74. “Adaptive cruise control allows me to do my work much more calmly. I have never noticed that this makes me more inattentive … It is a very safe aid.”

Lane departure warnings (LDW) / Lane keeping assistance (LKA)

1. “Lane assist is mega annoying and I turn it off when I’m fit. Later in the day, it comes on”
2. “… lane assist always turns off.”
3. “Lane detection is extremely irritating! It is disabled by many drivers! But safety for the driver is not necessary! Driver’s airbag is not standard!”
4. “Lane assist is hopeless.”
5. “Just look if you drive over a continuous line, it makes noise right away, you stay alert.”
6. “You get insanely annoyed by all those alarms. Especially from that line detection. 9 out of 10 times it goes off for no reason.”
7. “… Keep within the lines.”
8. “Line detection … encourages you to pick up your mobile. And all that touchscreen stuff only takes your eyes off the road. Just give me press and twist buttons. The more stuff on a car the less alert people are. If something suddenly happens, people no longer know how to intervene.”
9. “Remove lane warning …”
10. “You’re either a driver or you’re not. I don’t like these new safety systems at all like line guard … I love the old-fashioned work, shifting gears myself and keeping the vehicle under control myself! And not a computer or something!”
11. “… Lane assist, nice, nice sound too, radio turns off from it. My driving instructor always said those white lines are as flat as possible so you can drive over them.”
12. “Some roads are too narrow to keep your car strictly between the lines, the thing just keeps nagging.”
13. “No opinion, bullshit lane control.”
14. “Lane warning system is irritating …”

15. "… staying in your lane is useless because it often goes off due to peak hour lane driving, road works and even when the lines on the road shift. That's why most people turn off that sensor when starting. Given how narrow the peak hour lanes are, you can get a fright and hit a guardrail or the vehicle to your left. But about the phone in the car, since there was no response box. We are now obliged by the employer to call the customers half an hour before arrival. + the planning team messages us all day and personally I don't think that's okay. Driving and mobile phone don't go together."
16. "… Lane control is useful but annoying in road works due to stripes."
17. "Line control is not handy if you already drive defensively towards the right line."
18. "… Lane warning not really necessary, emergency lane lines make noise so you know you're not in the right place …"
19. "… lane detection … Yes, GREAT!"
20. "Lane … assist is also called the Facebook button. That says enough, I think."
21. "… Lane assist and other bells and whistles are distracting."
22. "Line detection hinders emergency brake system excellent but often at least 10 times a day utterly useless because it recognizes too many things as a collision distance keeper sensor works excellently well but only if you want to drive in a train all day."
23. "… Sensors to stay in the lane work fine."
24. "You get completely crazy from line protection."
25. "Line detection useless …"

Emergency braking / AEB

1. "It often creates dangerous situations itself, for example when the emergency brake is activated. When nothing happens in front of you."
2. "It will only become cheaper when every truck is equipped with it, now it is still dangerous especially with … emergency stop."
3. "It also creates dangerous situations, especially the brake assist."
4. "It occasionally brakes automatically because of traffic signs above the highway, extremely dangerous if you are driving behind it and do not have those systems. Always looking out of the window works better."
5. "Those systems that brake automatically seem safe in terms of head-to-tail collisions, but my experience teaches me that these systems create very unexpected situations that an experienced driver would never create!"
6. "I don't always find the braking system safe."
7. "The emergency brake system sometimes activates for a sign, unfortunately."
8. "Brake assist often jumps in when I have the situation completely under control, while the (dumb) system has a different opinion."
9. "The automatic braking system on new trucks IS LIFE-THREATENING. It is often wrongly adjusted."
10. "I am in favor of aids like … automatic braking in danger but there really should not be too many of them or it will be distracting."
11. "If it works well yes personally I have often sat on an empty highway with my face on the windshield because the emergency braking system saw ghosts."

12. “Emergency brake that goes off or responds while there is nothing wrong.”
13. “I turn them off. Almost had an accident because the truck went full on the brakes in a slight curve at 80 km/h. The automatic braking system was triggered because my own light (headlights) reflected on a traffic sign.”
14. “The distance radar at DAF works fantastically but I still had an accident with it at the Breda ramp I was driving on the A27 at Breda/navel we were only driving 60, merging traffic was holding up the works and at the next ramp a massive amount of mergers but on the main road it went a bit faster again so I moved to the left to make room for the mergers suddenly a black Audi in front of me was cut off by a colleague Audi in front of him the collision radar intervenes but they keep driving I couldn’t cancel the stop action by giving more gas and the combination kept braking until stopped by 3 cars behind so it doesn’t work flawlessly.”
15. “The emergency brake that activates as soon as you get cut off or it sees a cyclist on the bike path as an oncoming vehicle, works counterproductively at such a moment.”
16. “… self-braking vehicle ideal indeed!”
17. “Distance sensor is a great thing if it works properly, but if it reacts at random to things that aren’t there and then suddenly brakes the car, you are behind the wheel with a heart sinking feeling, let alone the person driving behind you. And why doesn’t that thing work in fog or bad weather?”
18. “The emergency brake intervenes so often unnecessarily that it would actually be better to remove it …”
19. “Emergency brake system sometimes overly sensitive.”
20. “I wouldn’t want to miss the … emergency brake anymore!”
21. “I experienced it once when a motorist came to drive in front of me and hit his brake. Very briefly but long enough to ensure that I was nose against the windshield because the truck made an emergency stop on its own. Life-threatening, there might be snow on the ground. And the truck also does this sometimes when I drive on cruise control and there comes a portal or viaduct which it thinks is too low and bam full brake on the highway really dangerous.”
22. “Auto brake assist on inner city roads and in the city very bad system.”
23. “Direction warning systems are horrible, it sees a sign or a car that needs to take the exit and has to brake hard due to a sharp turn and the system intervenes by braking fully, another behind you never expects this and is then helpless for a collision which is then inevitable, the truck is only 5 months old.”
24. “Near accident, self-braking system or whatever it’s called, often brakes by itself when there’s nothing wrong. If at that moment another vehicle is too close to me, it can end badly …”
25. “It’s all nonsense oh the technical gadgets that brake for me.”
26. “AEBS needs to be improved.”
27. “… The automatic brake sometimes has problems with an airplane or something, as it sometimes reacts when it’s not necessary.”
28. “My emergency brake system intervenes at the strangest moments. At a tree viaduct or parked car.”

29. “If it works well, not like with many cars from DAF that the emergency braking system intervenes in places where it is not needed.”
30. “Car braking system does not always work as desired”
31. “My Actros mp4 has an emergency brake, fortunately, I haven’t needed it yet, however, it is sometimes activated by matrix signs or bridges in rainy weather. Also at highway exits that lie in a curve, then it seems to the camera as if the braking car is still in front of you …”
32. “The braking system when someone suddenly crosses in front.”
33. “When motorists cross too short in front of you, the emergency brake system reacts and the truck slows down with full force so that the driver behind (overtaking ban) almost runs into it …”
34. “Some systems are downright life-threatening. For example, the emergency braking system, when you are cut off by a motorist, the system goes into action causing a great chance that your follower will shoot under your trailer.”
35. “… brake assist is also called the Facebook button. That says enough, I think.”
36. “The anti-collision system is very irritating during twilight or at night with viaducts.”
37. “I regularly encounter interventions by the automatic braking system due to reflections and shadows from a 2019 DAF.”
38. “I often find them dangerous because they see things that aren’t there and then intervene, like the distance/braking system that intervenes while there’s nothing wrong.”
39. “… the AEBS system work well, but unfortunately, not every truck has them yet, so you still get people cutting in front of you and your car starts to brake hard.”
40. “I find the emergency braking system handy, but if the person behind you doesn’t have it, they’ll crash into you, so I have mixed feelings.”
41. “I once turned off the assistance systems. Because my tractor slammed on the brakes in a curve. Even though I had plenty of space. If I had steel plates loaded at that moment. I’m sure they would have come out.”
42. “That automatic braking system can be useful, but not when you’re in a bend and it mistakes a traffic sign for a car and thinks you’re going to have a collision and so suddenly goes full on the brakes!”
43. “I have so many beeps now, I turn them off nowadays. Imagine what it does to me when I get a warning at every viaduct at night that I’m driving towards a traffic jam and the truck goes into emergency braking. If that really happens, I instinctively step on the gas.”
44. “That the vehicle itself intervenes when a car cuts you off is terrible.”
45. “The automatic braking system can activate if a car suddenly shoots in front of you, and that can lead to dangerous situations.”
46. “Emergency brake … super”
47. “Emergency brake can also be extremely dangerous since the car stops almost instantly.”
48. “Emergency brakes that respond to matrix signs are not really beneficial.”

49. “Yes absolutely. My truck automatically brakes in an emergency situation. If it ever comes to the point where every truck has this technology (mandatory), then at least no truck will ever run into the back of a stationary traffic jam.”
50. “Automatic braking not always because you are also dependent on other traffic and freight.”
51. “AEBS can be dangerous if someone, while braking, wants to merge in front of you quickly.”
52. “Recently got emergency brake system … I find it ideal, you can focus more on the actual driving itself.”
53. “Emergency brake system sometimes sees the strangest objects as a danger.”
54. “I am not satisfied with emergency brake … because car drivers pest test brake against trucks.”
55. “Emergency brake system is abused by merging traffic, this is because people suddenly forget that as merging traffic they need to adjust their speed to the traffic on the main lane, and not serve as an obstacle causing the emergency brake system to regularly intervene resulting in rear-end collisions on the main lane and the culprit can suddenly speed off.”
56. “… AEBS can be dangerous when it brakes for a turning car or matrix signs, which happens quite regularly with my DAF. In that respect, paying attention is much more accurate …”
57. “Some of these aids assume an ideal situation where all road users behave impeccably. But if, for example, a motorist wants to take the exit at the last minute and shoots across your grill, you have to be lucky that the emergency stop doesn’t kick in and you’re hanging with your seat belt.”
58. “The ’assistance systems’ are the first thing I turn off before I drive. Very dangerous (think of the distance meter that just makes a mini emergency stop for a matrix sign or viaduct on the highway) and highly irritating and disturbing to have to listen to beeping for 10 hours.”

Cameras and smart mirrors

1. “Those crazy cameras are also sensitive to interference and when it’s dark they’re an annoying light source, which just creates another blind spot.”
2. “Very much so, a backup camera and a front camera.”
3. “I have worked with a backup camera and side camera that automatically turn on, or can be turned on when you choose to do so. Works great!”
4. “Camera system and signals for when someone is on your left or right …”
5. “Standard equipment for all brands, cameras included, nothing more expensive.”
6. “As soon as possible, cameras with sound signals, and when you look in the mirrors, there should be a warning light if there’s something next to you.”
7. “only cameras and no mirrors is not good, the outside world then has no idea whether the driver can see you or not, the camera must serve as an expansion of the field of view”
8. “… Some camera systems can help. Cameras instead of mirrors, not so much, because you lose visual contact with other road users.”

9. “You start to rely on it, you think there’s a buzzer when I cross the line. Almost don’t look around anymore.”
10. “In the past, there have been various cameras, blind spot mirrors, alarms for swinging out on the cars, and it doesn’t work ideally with swinging out, an alarm goes off and when you look there are bushes, after a number of times this causes irritation, you get distracted by it and therefore you don’t pay attention to the important things.”
11. “Cameras help.”
12. “Because of my cameras, I look less in my mirrors.”
13. “Camera mirrors do not reflect depth and when it rains you see nothing and they break quickly.”
14. “I have driven for a while with a front and rearview camera … when reversing I didn’t have a front mirror or I had to keep pressing buttons during manoeuvring in a narrow busy street to switch between front and back.”
15. “Camera or sensors work well, in city distribution a window in the door.”
16. “I have a camera system etc. for London on my car, this camera greatly reduces my blind spot and I now see much more on the highway but also on roundabouts and through cities.”
17. “Cameras all around really help.”
18. “Camera works quite well but it’s still the driver who drives.”
19. “… This is also distracting: … 360-degree camera, reversing camera, blind spot camera.”
20. “Cameras.”
21. “For example, a camera behind the car is top,”
22. “Camera makes a difference.”
23. “Mandate cameras. Both downwards and backwards.”
24. “Cameras can certainly help.”
25. “Camera.”
26. “I once had a car in my blind spot that I didn’t see, but the system in the Volvo started beeping when I turned on the right turn signal, that prevented an accident. What I see in passenger cars can also be in trucks, that there’s a light in the mirror if something is driving on your right or left …”
27. “More cameras all around.”
28. “You’re either a driver or you’re not … Camera instead of mirrors also not safe! See me, see you is then not applicable. I love the old-fashioned work, shifting gears myself and keeping the vehicle under control myself! And not a computer or something!”
29. “Nowadays everything has to be done with a camera, I find a blind spot mirror for the front and the passenger side more than sufficient.”
30. “Light in the mirror when someone is driving on the right is an important thing.”
31. “More cameras for all-round visibility.”
32. “Blind spot camera standard!”
33. “I have a camera for the blind spot and reverse, ideal.”
34. “Camera is a good tool.”

35. “Camera system instead of mirror has increased the blind spot along the cab, still need a solution for this, otherwise satisfied with the camera system instead of mirrors due to a wider field of view.”
36. “Cameras for blind spot are perfect.”
37. “Camera is good …”
38. “I am strongly against camera instead of mirrors. I teach my children ”if you see the driver, the driver sees you” With the cameras on for example the Mercedes it’s not possible to make eye contact via mirror or camera.”
39. “… Yes, GREAT! Blind spot sensors and cameras too!”
40. “A camera would be handy, but I don’t have it myself.”
41. “Reversing camera.”
42. “Car kit or camera for blind spot.”
43. “As mentioned above the camera on the back for reversing.”
44. “5 mirrors + camera should be sufficient, right?”
45. “Only camera for blind spot …”
46. “You now get cameras outside with screen inside so you don’t have to look outside which neglects interaction with other traffic.”
47. “By aids, I mean backup cameras, and for example, for the blind spot …”
48. “Reversing camera. Dashcam because of annoying car drivers.”
49. “Cameras.”
50. “Also causes more distraction constantly looking in cameras.”
51. “Camera systems, reversing sounds, certainly help.”
52. “Cameras 360 and back separately.”
53. “Camera for blind spot works.”
54. “When it’s busy on the road, in the built-up area, a camera is really a must.”
55. “Reversing camera mandatory on every vehicle and trailer.”
56. “No cameras, however easy; because it happens outside..”
57. “… Cameras are handy in large vehicles because you have more visibility and can act on that yourself instead of a sensor that detects something and then just starts braking.”
58. “I have already driven in a truck with monitor mirrors in the cabin, these distort and when it rains the camera outside closes and you see very little. (dangerous for safety, abolish immediately)”

Appendix B: Summaries of the 15 Open Comment Boxes

The questionnaire contained 15 open comment boxes that provided the opportunity to give comments. The comments were automatically summarized with ChatGPT’s API:

Q12. Organizations (*n* = 506, 13.6%): The respondents express a range of negative sentiments towards trade unions and the TLN (Transport and Logistics Netherlands), criticizing them for not effectively representing or supporting truck drivers. Many feel that these organizations prioritize their own interests or those of employers over the needs of the workers. There's a sense of dissatisfaction with the lack of action or improvement in working conditions, wages, and respect for drivers. Some mention specific grievances such as inadequate representation during negotiations, failure to address the influx of cheaper foreign labor, and a general disconnect from the realities of the trucking profession. A few respondents also mention personal experiences of receiving little to no assistance when facing workplace issues. Overall, there's a strong sentiment that these organizations have lost touch with the drivers they are supposed to represent and have not done enough to advocate for better conditions in the transport sector.

Q17. Image (*n* = 490, 13.2%): The respondents express a strong sentiment of being undervalued and mistreated in their profession as truck drivers. They highlight issues such as poor sanitation facilities, particularly the use of mobile toilets (Dixis), and a lack of respect from employers, clients, and the general public. Many drivers feel they are treated as inferior, facing harsh working conditions and inadequate appreciation for their role in the supply chain. The influx of foreign drivers, particularly from Eastern Europe, is frequently mentioned as contributing to a negative image and working conditions for drivers. There is a call for better treatment, respect, and working conditions for truck drivers.

Q25. Overtaking bans for trucks (*n* = 1334, 36.0%): The respondents express strong opinions against the current overtaking bans for trucks on highways, arguing that such restrictions lead to dangerous situations and increased traffic congestion. Many feel that the bans create long lines of trucks ("treintjes") that make merging and exiting more difficult and dangerous for all road users. Some suggest that overtaking should be allowed, especially outside of peak hours or on roads with three or more lanes, to improve traffic flow and safety. There's also frustration about the lack of enforcement on those who ignore the bans, particularly foreign truck drivers. Additionally, some respondents believe that the speed difference between trucks and passenger cars is now minimal due to speed limit changes, making the bans unnecessary. Overall, there's a call for reevaluation of overtaking bans to enhance road safety and efficiency.

Q28. Blind spot (*n* = 761, 20.5%): The respondents emphasize the importance of increased awareness and education regarding the blind spot issue around large vehicles, particularly for schools and during driving lessons for all types of licenses. They suggest that more attention should be given to teaching both children and adults about the dangers of blind spots. Many advocate for the use of technology such as cameras and warning signals to help mitigate blind spot accidents, while others believe that better mirror adjustment and driver vigilance are key. There's a consensus that stickers, like those mandated in France, are ineffective and that a combination of education, technology, and possibly regulatory changes (such as making certain driving behaviors around trucks illegal) could help reduce blind spot-related accidents.

Q30. Vehicle aids (*n* = 441, 11.9%): The respondents' comments reflect a mix of opinions on the use of technological aids and safety systems in trucks. While some find features like adaptive cruise control and rear-view cameras beneficial for safety and convenience, others express concerns that an over-reliance on these systems can lead to decreased attentiveness and potential danger, especially when systems malfunction or react unexpectedly. There's a sentiment that too many aids can make drivers complacent, relying too much on technology rather than their own skills and awareness. Additionally, some suggest that more education for all road users, not just truck drivers, would improve safety. Overall, there's a call for a balanced approach that leverages technology without undermining the role of the driver's judgment and attentiveness.

Q34. On-board computer (*n* = 625, 16.9%): The respondents' comments reflect a range of opinions on the use of onboard computers (BC) in the transportation industry. Many express concerns about increased work pressure and surveillance, feeling constantly monitored by their employers. Some mention the BC as a mere tool for hour registration, lacking added value, while others see it as a source of stress, especially when it leads to additional tasks or constant communication from planners. A few respondents appreciate the BC for its efficiency in handling administrative tasks and providing clear job instructions. However, the overarching sentiment is that BCs contribute to an impersonal work environment and can exacerbate work pressure, with the quality of the planning and company culture playing significant roles in how the technology impacts the driver's experience.

Q39. Driving times (*n* = 722, 19.5%): The respondents express significant concerns about the difficulty in finding parking spaces, with many mentioning that parking areas are often full early, leading to stress and sometimes forcing them to park in unauthorized areas. There's a consensus that there are too few parking spaces available, especially after 17:00, and that this issue is exacerbated by the presence of trucks from Eastern Europe occupying these spaces. Some respondents suggest that minor infractions of the driving hours regulations should not result in high fines, especially when the alternative could pose a greater danger. The complexity and strictness of the European driving hours regulations are also highlighted, with many calling for more flexibility to accommodate the realities of traffic and delays. Stress from trying to adhere to these regulations, along with the digital tachograph, is a common theme. There's a feeling that the current system increases pressure on drivers, with some admitting to taking more risks when they are close to exceeding their driving hours. The lack of parking spaces is seen as a significant contributor to this stress, with many advocating for an increase in safe and accessible parking areas for truck drivers.

Q44. Transport crime (*n* = 316, 8.5%): The respondents express frustration and concern over transport-related criminal activities, such as diesel theft and vehicle break-ins, highlighting a perceived lack of police presence and response, especially at night and on parking lots. They mention the scarcity of secure parking spaces and the need for better surveillance and lighting. Some suggest that reporting incidents feels futile as there's a belief that authorities do not take adequate action. There's also a call for more awareness and preventive measures, including secure parking and better education on avoiding theft. A few respondents note their specific strategies for minimizing risk, like choosing safer parking locations or using physical deterrents. Overall, there's a sentiment of dissatisfaction with the current state of security and support for transport workers.

Q51. Respect (*n* = 623, 16.8%): The respondents express a strong sentiment of feeling undervalued and disrespected in their profession as truck drivers. They highlight issues such as being treated poorly by companies and the public, facing long wait times without explanation, and being compared unfavorably to drivers from Eastern Europe. Many feel that their work is not appreciated despite the essential role they play in the economy, noting that without their services, many goods would not reach their destinations. There is a call for better treatment, higher wages, and more respect for the profession, emphasizing that the current conditions are not sustainable or fair.

Q53. Organizations' commitment (*n* = 497, 13.4%): The respondents express dissatisfaction with organizations representing truck drivers, feeling that their needs and concerns are not adequately addressed. They mention issues such as inadequate support during times of personal crisis, insufficient action to make the profession more attractive to younger generations, and a perceived bias towards employers over employees. There's a call for more human treatment, better representation during negotiations, and more visible actions like strikes to demand better conditions. Some feel that organizations are more focused on their own interests rather than genuinely supporting drivers, with specific frustrations around low wages, lack of parking and sanitary facilities, and the need for better enforcement of existing regulations. There's a sense of abandonment and a desire for more effective advocacy and tangible improvements in their working conditions.

Q58. Paid parking (*n* = 735, 19.8%): The respondents express strong dissatisfaction with the current state of paid parking for trucks, highlighting issues such as high costs, lack of services, poor sanitation, and inadequate security. Many emphasize the need for better-regulated parking areas with essential facilities like restaurants and clean restrooms. There's a common sentiment against the financial burden falling on drivers, suggesting that employers or clients should cover parking costs. The lack of sufficient parking spaces, especially in the Netherlands, is also a concern. Some suggest that parking fees should be refundable through meal purchases at associated restaurants. Overall, there's a call for improved parking infrastructure and services, with costs not being passed onto drivers.

Q60. Would you like to see changes in existing paid parking spaces? (*n* = 3080, 83.1%*): The respondents' comments focus on the need for improvements in parking facilities for truck drivers in the Netherlands. Key points include the desire for lower or no parking fees, better and cleaner sanitary facilities, the availability of decent eating options, and the provision of more parking spaces, including specific areas for ADR (hazardous materials) parking. There is also a call for better security measures, including surveillance and fencing, and the suggestion that parking fees should be covered by employers or compensated through meal purchases. Additionally, respondents express a need for more amenities such as WiFi, healthy food options, and leisure facilities. Overall, there is a strong demand for higher quality, more affordable, and better-maintained parking facilities that cater to the needs of truck drivers.

Q63. Mobility Package (*n* = 1510, 40.7%*): The comments express a wide range of frustrations and concerns from drivers regarding the current state of the transportation industry in Europe. Key issues include the lack of enforcement on existing regulations, particularly regarding cabotage and the exploitation of drivers, especially those from Eastern Europe. Many respondents feel that there is an unfair competition and that the regulations coming from Brussels are either ineffective or not properly enforced. There is a call for more parking spaces with better facilities, and a significant concern about the working conditions and pay of drivers. Some comments also mention the need for more uniform rules across European countries and better control of fraud and exploitation within the industry. Overall, there is a sense of dissatisfaction with the current state of affairs, with many feeling that the needs and well-being of drivers are being overlooked.

Q67. What is, in your opinion, the first thing that needs to change in the sector? (Please provide 1 answer) (*n* = 3234, 87.2%): The respondents' comments primarily focus on the need for improved treatment and respect for truck drivers, higher wages, and better working conditions. They express a desire for a higher basic salary, reduced taxation on overtime, and more flexible working hours to reduce work pressure. There is a strong call for equal pay and conditions for all drivers, regardless of nationality, to address unfair competition from lower-paid drivers from Eastern Europe. Additionally, respondents highlight the need for better parking facilities and rest areas, as well as a reduction in excessive regulation and fines. Overall, there is a clear demand for greater appreciation of the truck driving profession, with financial and working condition improvements to make the sector more attractive to new entrants.

Q68. Comments and suggestions that you couldn't include in the questions can be written below (*n* = 559, 15.1%): The respondents' comments reflect a range of concerns and suggestions from individuals likely involved in the transportation and trucking industry. Key themes include the need for better pay and working conditions, frustration with strict regulations and excessive fines, and a desire for more respect and appreciation for the profession. Many express concerns about the impact of foreign drivers on the market, suggesting that tolls or fees should be implemented for foreign trucks to level the playing field. There's also a call for better parking facilities and rest areas for drivers, as well as suggestions for improving the overall image of the profession to attract new drivers. Some respondents also mention the need for more consistent and fair enforcement of rules across Europe, and a few suggest changes to retirement age and pension arrangements to better reflect the demands of the job.

**Q60 and Q63 were inadvertently set as mandatory for either the entire survey period or a portion of it, which likely explains their relatively high response rates. Note that Q67 also has a high response rate, but this question was not mandatory. A likely explanation for this is that Q67 is a generic closing question.*

Part II

Measuring Driving Behavior: Practical Approaches

4

Detecting Lane Change Maneuvers from GPS data

This chapter was published as: Driessen, T., Prasad, L., Bazilinskyy, P., & De Winter, J. (2022). Identifying lane changes automatically using the GPS sensors of portable devices. 13th International Conference on Applied Human Factors and Ergonomics, New York. https://doi.org/10.54941/ahfe1002433

Abstract

Mobile applications that provide GPS-based route navigation advice or driver diagnostics are gaining popularity. However, these applications currently do not have knowledge of whether the driver is performing a lane change. Having such information may prove valuable to individual drivers (e.g., to provide more specific navigation instructions) or road authorities (e.g., knowledge of lane change hotspots may inform road design). The present study aimed to assess the accuracy of lane change recognition algorithms that rely solely on mobile GPS sensor input. Three trips on Dutch highways, totaling 158 km of driving, were performed while carrying two smartphones (Huawei P20, Samsung Galaxy S9), a GPS-equipped GoPro Max, and a USB GPS receiver (GlobalSat BU343-s4). The timestamps of all 215 lane changes were manually extracted from the forward-facing GoPro camera footage, and used as ground truth. After connecting the GPS trajectories to the road using Mapbox Map Matching API (2022), lane changes were identified based on the exceedance of a lateral translation threshold in set time windows. Different thresholds and window sizes were tested for their ability to discriminate between a pool of lane change segments and an equally-sized pool of no-lane-change segments. The overall accuracy of the lane-change classification was found to be 90%. The method appears promising for highway engineering and traffic behavior research that use floating car data, but there may be limited applicability to real-time advisory systems due to the occasional occurrence of false positives.

4.1. Introduction

Systems capable of detecting lane changes, such as lane departure warning systems, have become common in new cars. These systems usually rely on cameras to detect lane boundaries (e.g., Toyota, 2022; Volkswagen, 2021). Less common are methods that identify lane changes without using cameras. Such methods could be relevant for three reasons.

The first reason is that, even though modern cars are equipped with cameras, it may take many years before this technology is commonplace. Young drivers, for example, often buy their vehicles second-hand and thus have to rely on safety systems in old models, yet it can be argued that this is the group most in need of modern safety systems (Lee, 2007). The widespread availability of smartphones may provide such an opportunity. If lane departure warning systems became available on smartphones, they could provide safety alerts and lane-level navigation assistance to virtually all drivers. A second motivation for developing cameraless methods of lane change detection lies in their potential for traffic behavior research and road design. The increasing availability of floating car data allows for studying traffic with more detail than traditional, hardware-intensive methods of data collection such as induction loops and traffic cameras (e.g., Arman & Tampère, 2021). Floating car data can reveal how groups of drivers perform maneuvers on specific sections, which may inform the design of highways (Vos et al., 2021).

Thirdly, knowledge on where, how often, or how aggressively drivers change lanes can serve as input for driving style recognition algorithms, which are used in smartphone applications that give drivers feedback and coaching about their driving style (for reviews, see Michelaraki et al., 2021; Singh & Kathuria, 2021). Such applications are increasingly used by vehicle insurance companies to offer discounted premiums to drivers that adopt non-risky driving styles (Baecke and Bocca, 2017; Tselentis et al., 2017).

Related work

The accuracy of consumer-grade and smartphone-based GPS receivers is in the range of 3–13 meters (Merry & Bettinger, 2019; Izet-Ünsalan and Ünsalan, 2020; Wing et al., 2005), which is too low to estimate the receiver's location with lane-level resolution. However, as the error in the measurements is largely caused by atmospheric disturbances or signal reflections on surrounding structures, it is expected to remain relatively constant on open highways (Sanz Subirana, 2011; Izet-Ünsalan & Ünsalan, 2020). This means that relative changes in the GPS trajectory may be indicative of certain highway maneuvers. When combined with information about the road trajectory, changes in the lateral distance between the road and the vehicle's trajectory may be used to identify lane changes. Sekimoto et al. (2012) demonstrated this by plotting the lateral distance to the road centerline of six lane changes. Their results showed that lane changes were visually discriminable from straight driving, but a formal assessment was lacking. A further evaluation of this concept was performed by Faizan et al. (2019). By calculating the difference in heading angle between the vehicle's trajectory and the road trajectory and multiplying its sine by the traveled distance since the last observation, they obtained the "instantaneous lateral distance," which is, in fact, a measure of lateral velocity. They then integrated this variable by summing up subsequent values, obtaining the "accumulative lateral distance." When this lateral drift exceeded a threshold of 1.5, it would present an alarm. They reported high detection accuracies, but it should be noted they relied on a GPS device that sampled at 10 Hz, whereas most smartphones typically operate at sampling rates of 1 Hz.

Aim

The literature to date suggests that it is feasible to detect lane changes based solely on GPS signals. However, in the existing analyses that we found, detail was missing on how such algorithms perform on highway sections that contain many irregularities such as curves and on- and off-ramps, and how performance varies between devices. The current paper describes the design and evaluation of a lateral-distance-based algorithm on 'easy' roads (a straight highway section leading from Delft to Rotterdam) and on a more difficult highway (Rotterdam's ring road). Furthermore, we investigated if performance varies between four portable devices.

4.2. Method

Data collection

Data were collected during three trips from the city of Delft (exit "Zuid"), via the A13 to the Rotterdam Ring Road, making a full lap on the Ring, and back to Delft-Zuid over the A13 (Figure 4.1). The total distance traveled during the three trips, excluding the Beneluxtunnel and an accidental detour in Trip 1, was 158 km. The first trip was on June 4, 2021, in a 2018 Peugeot 108 (915 kg), and the second and third trips were on October 21, 2021, in a 2021 KIA Picanto (974 kg), both small city cars. The first author drove the car and changed lanes whenever it was judged safe and unobtrusive to other traffic. This resulted in 215 lane changes (110 right, 105 left). The lane width on the route was 3.5 m. The speed limit was 100 km/h, which was also the target speed of the driver. The speed varied somewhat due to occasional busy segments on the Rotterdam Ring road.

GPS data were recorded at a frequency of 1 Hz on a Samsung Galaxy S9 and a Huawei P20 Lite (using the Android app "GPS Logger" by BasicAirData, 2022), on a GlobalSat BU343-s4 USB GPS receiver, and on a GoPro Max. The GoPro recorded GPS at 18 Hz, which was downsampled to 1 Hz for comparability with the other signals by taking the last entry of every 18 instances. The smartphones were mounted to the dashboard using standard car phone holders, whereas the GlobalSat's antenna was magnetically attached to the top of the car. The GoPro was mounted facing forward behind the windshield in the middle of the dashboard. Besides recording GPS, the GoPro made video recordings which were later used to manually annotate the moments the car changed lanes. Lane change timestamps were annotated when the GoPro's view was visually centered with a lane boundary marking (Figure 4.2). Double lane changes were annotated when the car drove on the middle of the center lane and were treated the same as single-lane change events in the analysis.

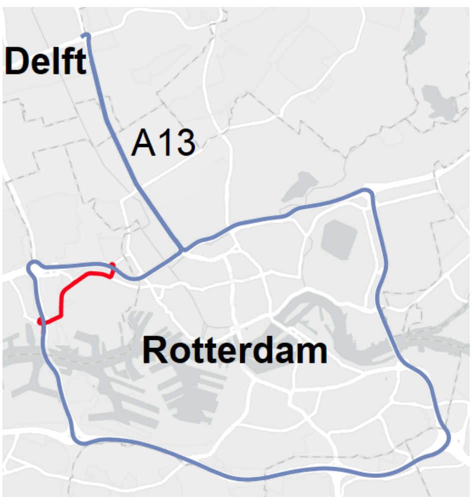


Figure 4.1: Route (blue) with excluded section from Trip 1 after a wrong exit (red).

Figure 4.2: GoPro's view at the moment a lane change was annotated.

Data processing

The road geometry was obtained by snapping the GoPro's GPS recordings to OpenStreetMap's road network using Mapbox Map Matching API v5 (2022). For each GPS coordinate, the lateral distance to the road's trajectory was calculated. The distance was given a positive sign when the GPS coordinate was on the right side of the road (when facing in the direction of travel) and a negative sign when it was on the left side of the road. This resulted in a signal representing the lateral position of each GPS coordinate

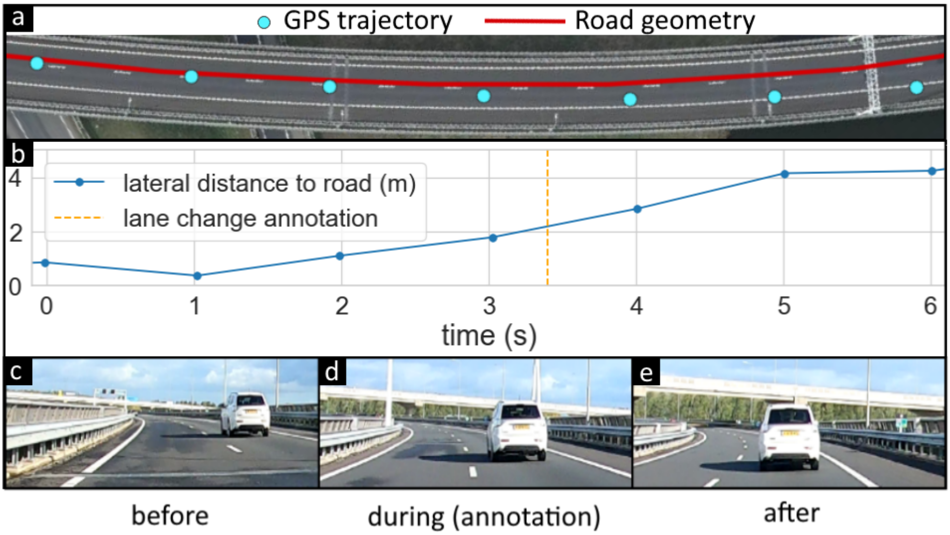


Figure 4.3: a: Map view of road section with a lane change. b: Lateral distance of the GPS points relative to the road during the lane change (in meters). c, d, e: Video segments of the lane change from the perspective of the ego vehicle, where panel (d) shows the moment the lane change was manually annotated (the camera view is visually centered with the lane boundary).

with respect to the road. Figure 4.3 shows an example of a recorded lane change and its lateral position signal during the lane change.

Analysis

We developed an algorithm that discriminates segments with a lane change from segments without a lane change. Therefore, we first created two classes consisting of isolated segments with a window size W of data points. The positive class contained segments during which a lane change occurred. A segment was created for each lane change annotation timestamp by extracting the W data points with timestamps nearest to the lane change timestamp. As the sampling frequency of each device was 1 Hz, the duration of each extracted segment (the time between the first and last element) was $W - 1$ seconds. The lateral position values of all left lane changes were multiplied by -1. This way, lane changes in both directions can be recognized by the same algorithm.

The negative class consisted of segments during which no lane change was performed. These segments were obtained by splitting up segments during which no lane change annotation was present into non-overlapping intervals of W – 1 seconds. Only segments were included that were at least 5 seconds removed from any lane change annotation. This procedure resulted in a larger pool of negative segments than positive intervals (more time is spent driving without changing lanes). To create classes of equal size, samples were randomly drawn from this pool without replacement.

The difference between the first and the last element was computed for each segment.

This value represented the accumulated lateral translation during a segment. If this value exceeded a threshold T, the segment was classified as a lane change; if this value did not exceed T, the segment was classified as no lane change. The first step of the evaluation was to find the threshold values T and window size W that gave the best classification accuracy for all devices.

For the remainder of the analysis, we fixed the values of T and W to those that gave the highest classification accuracy. Next, using the same procedure as above, we evaluated the classification accuracy between the devices and between the A13 section and the Rotterdam Ring section.

4.3. Results

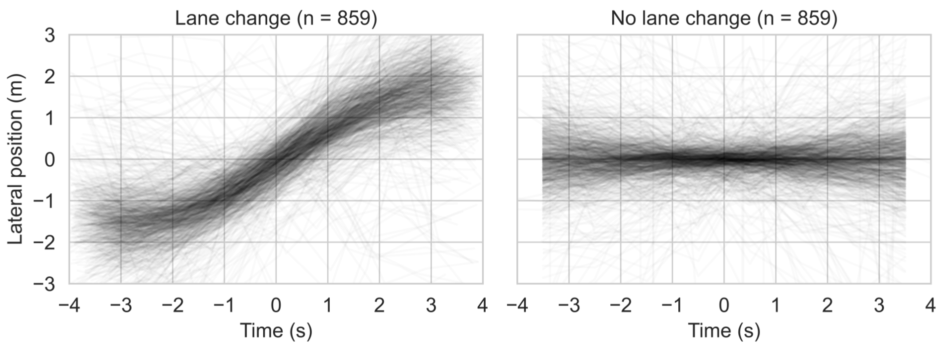


Figure 4.4: Left: Lateral position relative to the segment's mean for all lane changes and all four devices (n = 215 lane changes × 4 devices − 1 missing segment = 859). The lane change took place at Time = 0 s. Right: Lateral position relative to the segment's mean for segments in which no lane change took place.

Figure 4.4 shows lateral position data for all segments and devices combined when using a window size W of 8 data points. It can be seen that, on a group level, lane-change segments are distinguishable from no-lane-change segments.

Table 4.1 shows the effect of the lateral translation threshold T for a fixed window size of W = 8 seconds. It can be seen that, as the threshold increases, the number of true positives (TP) and false positives (FP) monotonically decrease, whereas the number of true negatives (TN) and false negatives (FN, i.e., misses) monotonically increase. Note that the true positive rate (TPR) is defined as TP / (TP + FN), whereas the false positive rate (FPR) is defined as FP / (FP + TN).

The combination of W and T that yielded maximal classification accuracy was found by varying window size W from 2 to 8 data points and varying threshold T with increments of 0.1 m for each window size W. Table 4.2 shows that the highest accuracy (0.905) was achieved using a window size of 6 and a lateral translation threshold of 1.5 m, with accuracy defined as (TP + TN)/(TP + FP + TN + FN). Of note, classification accuracy was still high (0.868) for a window size of 2, i.e., when the lateral position difference between only two data points was used.

Table 4.1: Classification accuracy when varying the lateral translation threshold T for a fixed window size W = 8.

Threshold T (m)	Accuracy	TP	FP	TN	FN	TPR	FPR
0.0	0.724	822	438	421	37	0.957	0.510
0.2	0.771	815	349	510	44	0.949	0.406
0.4	0.799	805	292	567	54	0.937	0.340
0.6	0.825	801	242	617	58	0.932	0.282
0.8	0.847	798	202	657	61	0.929	0.235
1.0	0.865	790	163	696	69	0.902	0.190
1.2	0.881	779	125	734	80	0.907	0.146
1.4	0.889	769	100	759	90	0.895	0.116
1.6	**0.894**	757	80	779	102	0.881	0.093
1.8	0.892	745	71	788	114	0.867	0.083
2.0	0.888	722	55	804	137	0.841	0.064
2.2	0.875	691	47	812	168	0.804	0.055
2.4	0.857	657	43	816	202	0.765	0.050
2.6	0.836	608	31	828	251	0.708	0.036
2.8	0.814	566	27	832	293	0.659	0.031
3.0	0.786	517	25	834	342	0.602	0.029

Table 4.2: Optimal lateral translation thresholds for varying window size W.

Window size W	Threshold T (m)	Accuracy	TP	FP	TN	FN	TPR	FPR
8	1.5	0.895	767	89	770	92	0.893	0.104
7	1.6	0.888	752	86	773	107	0.875	0.100
6	**1.5**	0.905	765	69	790	94	0.891	0.080
5	1.3	0.902	767	77	782	92	0.893	0.090
4	1.2	0.899	747	62	797	112	0.870	0.072
3	0.8	0.880	745	93	766	114	0.867	0.108
2	0.4	0.868	740	107	752	119	0.861	0.125

The above parameter values (W = 6, T = 1.5 m) were used to compare the devices (Table 4.3) and the two highway segments (Table 4.4). It can be seen that the four GPS devices yielded similar accuracies, with the GoPro Max coming out slightly better than the other three devices. Accuracy was considerably worse on the Ring Rotterdam than the relatively straight and uncluttered A13 highway.

Table 4.3: Classification accuracy per device (W = 6, T = 1.5 m).

Device	Accuracy	TP	FP	TN	FN	TPR	FPR
GlobalSat	0.893	188	20	194	26	0.879	0.093
GoPro Max	0.923	197	15	200	18	0.916	0.070
Huawei P20 Lite	0.886	188	22	193	27	0.874	0.102
Samsung Galaxy S9	0.895	190	20	195	25	0.884	0.093

Table 4.4: Classification accuracy for the A13 (Delft–Rotterdam, Rotterdam–Delft) and Ring Rotterdam sections, all devices combined (W = 6, T = 1.5 m).

Road	Accuracy	TP	FP	TN	FN	TPR	FPR
A13	0.939	215	8	228	21	0.911	0.034
Ring Rotterdam	0.890	550	64	559	73	0.883	0.103

A visual inspection of falsely classified segments revealed that incorrect predictions tended to be caused by road geometry definitions. We found this to be the case occasionally on curved sections, under overpasses, and at sections with lane splits. Figure 4.5 shows two relatively straight sections where all four recording devices gave an incorrect prediction. The first segment shows a scenario where a lane change was made, but it was not identified (false negative) as the road geometry moves in the lane change direction. In the bottom example, no lane change was performed, but as the road definition jumped sideways, it decreased the lateral threshold, resulting in an incorrectly flagged fragment (false positive).

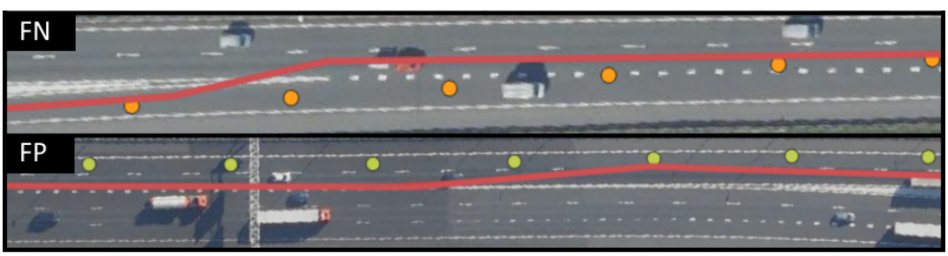


Figure 4.5: Examples of a false negative (top) and a false positive (bottom). In these figures, the vehicles travel from left to right.

4.4. Discussion

This study examined whether lane changes can be identified using GPS position data for different off-the-shelf devices and different highway sections. The results showed an overall true positive rate of 89% at a false positive rate of 8%. It was found that the GPS devices yielded similar classification performance. Furthermore, it was shown that the false positive rate was 3% on straight highway roads, while it was 10% on the more cluttered ring road consisting of curves, exits, lane splits, etc.

The development of a real-time warning system, such as a drowsiness detection system that measures if the vehicle drifts out of the lane, would require further investigation. The findings of the current study suggest that this may be hard to achieve using only GPS. In the current evaluation, balanced classes were created. In reality, there will be proportionally more segments without lane change, meaning that the number of false positives per minute of driving will be high (8% for every 5 seconds totals approximately 1 false alarm per minute). Systems that frequently provide false warnings tend to be turned off by drivers (Reagan et al., 2018). Although the false alarm rate can be decreased by increasing the detection threshold, this would go at the cost of the ability to correctly predict lane changes. Another factor is that lane changes in the positive class were centered around the lane changes. For real-time applications, a rolling window approach will have to be used. In our study, lane changes were detected after allowing some time for the lateral position to accumulate, which might not occur fast enough for lane drift warnings. Also, it is noted that our current method detects lane changes, not the lane on which the car was driving.

On the other hand, the obtained accuracy may be high enough for driving style recognition algorithms which aim to establish whether the driver is a frequent lane changer or not (if only road segments free of irregularities such as exits or curves are considered). Such driving style recognition algorithms could benefit from further information such as traffic density or from geo-specific approaches that compare the driver with other drivers driving on the same road at the same time of day. Our study focused on detecting the occurrence of a lane change. Future research could try to infer the aggressiveness of the lane change, for example, by incorporating lateral velocity information. It is also expected that our method is useful for road design applications, for example by determining 'lane change hotspots' based on floating car data.

The collected data were limited to driving by the first author, who changed lanes whenever this was deemed safe. The classification results reported in this paper should be validated on naturalistic driving data from a more diverse pool of drivers, vehicles, devices, roads, and weather conditions. The present study used GPS signals to detect lane changes. Future research could use gyroscopes and accelerometers, which are available in smartphones as well. This approach has been tried by Ramah et al. (2021), who observed that lane-change detection using these sensors alone is difficult if a lane change is gentle. Future research could use sensor fusion of smartphone GPS and IMU data (and see Islam and Abdel-Aty, 2021). In conclusion, this study established the feasibility of detecting lane changes using portable GPS devices. Lane change information based on floating car data may be useful for road design and traffic flow management.

Supplementary material

Code and data can be found on https://github.com/tomdries/gps-lane-changes and https://doi.org/10.4121/19170302.

References

Arman, M.A. & Tampère, C.M.J. (2021) 'Lane-level routable digital map reconstruction for motorway networks using low-precision GPS data', Transportation Research Part C: Emerging Technologies, 129, 103234. https://doi.org/10.1016/j.trc.2021.103234

Baecke, P. & Bocca, L. (2017) 'The value of vehicle telematics data in insurance risk selection processes', Decision Support Systems, 98, pp. 69–79. https://doi.org/10.1016/j.dss.2017.04.009

BasicAirData (2022) 'BasicAirData GPS Logger'. Available at: https://www.basicairdata.eu/projects/android/android-gps-logger (Accessed: February 14, 2022).

Faizan, M., Hussain, S. & Hayee, M.I. (2019) 'Design and development of in-vehicle lane departure warning system using standard global positioning system receiver', Transportation Research Record: Journal of the Transportation Research Board, 2673, pp. 648–656. https://doi.org/10.1177/0361198119844751

Islam, Z. & Abdel-Aty, M. (2021) 'Real-time vehicle trajectory estimation based on lane change detection using smartphone sensors', Transportation Research Record, 2675, pp. 137–150. https://doi.org/10.1177%2F0361198121990681

Izet-Ünsalan, K. & Ünsalan, M. (2020). 'Classroom study of GNSS position accuracy using smartphones' Scientific Bulletin of Naval Academy, XXIII, pp. 83–89.

Lee, J.D. (2007) 'Technology and teen drivers', Journal of Safety Research, 38, pp. 203–213. https://doi.org/10.1016/j.jsr.2007.02.008

Mapbox Map Matching API v5 (2022). https://docs.mapbox.com/help/glossary/map-matching-api (Accessed: January 28, 2022)

Merry, K. & Bettinger, P. (2019) 'Smartphone GPS accuracy study in an urban environment', PLOS ONE, 14, e0219890. https://doi.org/10.1371/journal.pone.0219890

Michelaraki, E. et al. (2021) 'Post-trip safety interventions: State-of-the-art, challenges, and practical implications', Journal of Safety Research, 77, pp. 67–85. https://doi.org/10.1016/j.jsr.2021.02.005

Reagan, I.J., Cicchino, J.B., Kerfoot, L.B. & Weast, R.A. (2018) 'Crash avoidance and driver assistance technologies–Are they used?' Transportation Research Part F, 52, pp. 176–190. https://doi.org/10.1016/j.trf.2017.11.015

Ramah, S.-E., Bouhoute, A., Boubouh, K. & Berrada, I. (2021) 'One step further towards real-time driving maneuver recognition using phone sensors', IEEE Transactions on Intelligent Transportation Systems, pp. 6599–6611. https://doi.org/10.1109/TITS.2021.3065900

Sanz Subirana, J., Juan Zornoza, J.M. & Hernández-Pajares, M. (2011) Multipath - Navipedia, European Space Agency. Available at: https://gssc.esa.int/navipedia//index.php/Multipath (Accessed: January 28, 2022).

Sekimoto, Y., Matsubayashi, Y., Yamada, H., Imai, R., Usui, T. & Kanasugi, H. (2012) 'Lightweight lane positioning of vehicles using a smartphone GPS by monitoring the distance from the center line', 15th International IEEE Conference on Intelligent Transportation Systems (ITSC 2012), Anchorage, AK, USA, pp. 1561–1565. https://doi.org/10.1109/ITSC.2012.6338737

Singh, H. & Kathuria, A. (2021) 'Analyzing driver behavior under naturalistic driving conditions: A review'. Accident Analysis and Prevention, 150, 105908. https://doi.org/10.1016/j.aap.2020.105908

Toyota (2022) 'Toyota 2022 Corolla Owner's Manual'. Available at: https://www.toyota.com/owners/resources/warranty-owners-manuals.corolla.2022 (Accessed: February 15, 2022).

Tselentis, D.I., Yannis, G. & Vlahogianni, E.I. (2017) 'Innovative motor insurance schemes: A review of current practices and emerging challenges', Accident Analysis and Prevention, 98, pp. 139–148. https://doi.org/10.1016/j.aap.2016.10.006

Volkswagen (2021) 'Owner's Manual Golf, Golf GTI, U.S. Edition, Model Year 2021'.

Vos, J., Farah, H. & Hagenzieker, M. (2021) 'Speed behaviour upon approaching freeway curves', Accident Analysis and Prevention, 159, 106276. https://doi.org/10.1016/j.aap.2021.106276

Wing, M.G., Eklund, A. & Kellogg, L.D. (2005) 'Consumer-grade global positioning system (GPS) accuracy and reliability', Journal of Forestry, 103, pp. 169–173. https://doi.org/10.1093/jof/103.4.169

5

Detecting Driving Style from Accelerometer and GPS data

This chapter was published as: Driessen, T., Stefan, D., Heikoop, D., Dodou, D., & De Winter, J. (2024). Using mobile devices for driving test assessment: A study of acceleration and GPS data. Transportation Letters. https://doi.org/10.1080/19427867.2024.2352198

Abstract

There is a need to improve the validity of the driving test as a measure of an individual's ability to drive safely. This paper explores the use of algorithms to analyze acceleration and GPS data from a smartphone and a GoPro to distinguish between different driving styles, as performed by experienced examiners who portray stereotypical driving test candidates. Measures from nine driving tests were analyzed, including (harsh) accelerations, jerk, mean speed, and speeding. Results showed that the type of car, instructed driving style, and driving route impacted the dependent measures. It is concluded that GPS and accelerometer data can effectively distinguish between cautious, normal, and aggressive driving. However, it is important to consider additional sensors, such as cameras, to allow for more context-aware assessments of driving behavior. Furthermore, we demonstrate methods to quantify variations in road conditions and we provide suggestions for presenting the data to driving examiners.

5.1. Introduction

Young drivers face a high risk of road accidents due to a combination of factors such as inexperience and limited skills as well as immaturity and risk-taking behavior (Lajunen et al., 2022; Rolison & Moutari, 2020; Weast & Monfort, 2021). To address this issue, several countermeasures have been implemented. One such measure is graduated driver licensing (GDL), which restricts the driving privileges of new drivers in stages as they gain experience (Curry et al., 2017; Fell et al., 2011; Poirier et al., 2018; Williams, 2017). Another countermeasure is the introduction of new vehicle technology, such as front crash prevention and blind spot monitoring, which can help reduce the accident risk of young drivers in particular (Mueller & Cicchino, 2022). Effective enforcement of traffic laws (Bates et al., 2020; De Waard & Rooijers, 1994) and anti-speeding and anti-drink-drive education campaigns can further reduce the number of accidents involving young drivers (Tay, 2005). Finally, the driving test is considered an important screening mechanism that helps ensure that only drivers who are deemed skilled receive their driver's license.

The driving test is often the only formal evaluation of a person's driving skills before they are granted a driver's license (Helman et al., 2017). However, the driving test may not provide a veridical assessment of a person's driving abilities, as it only provides a snapshot of the candidate's skills. The likelihood of making mistakes during the test can be influenced by external factors, such as weather conditions and the occurrence of specific situations on the road. Furthermore, even though driving examiners are trained and qualified, there is still room for subjectivity and human error or inconsistency in their verdict (Baughan et al., 2005). Another challenge in driver testing is that some candidates are disagreeable or may even become aggressive when they hear that they failed the exam (Alsharef et al., 2022; Foxe, 2020). Hence, there is a need for a more data-driven presentation of the test verdict.

In a previous study with driving examiners in the Netherlands, it was found that examiners would like to have access to data-based evidence to support their decisions to pass or fail a candidate (Driessen et al., 2021). Examples mentioned by the examiners include dashcam footage, recordings of the candidate's viewing behavior, and data on speed, headway, and braking behavior. The examiners indicated they wanted to be able to access such data in a raw (e.g., graphs, footage) or semi-processed (e.g., good/bad evaluations) form so that they could provide more detailed explanations for their verdict. However, the examiners also believed that current technology is not advanced enough to fully replace human judgment with an automated pass-fail system, indicating that while technology can assist but not replace the human evaluator (Driessen et al., 2021).

In the area of usage-based insurance, devices like mobile phones and dongles are widely used to monitor driving behavior. These devices can record driving measures such as speed and acceleration and offer the advantage of not requiring modifications to the vehicle or special hardware installation. Studies have shown that hard braking is a reliable predictor of accident risk for car drivers (Hunter et al., 2021; Ma et al., 2018; Stipancic et al., 2018) and truck drivers (Cai et al., 2021; Driessen et al., 2024). Additionally, studies have explored the use of mobile phones to identify different driving styles, such as dangerous and aggressive driving (Carlos et al., 2020; Chan et al., 2020; Johnson & Trivedi, 2011; Othman et al., 2022). Research has also explored the use of smartphone apps for providing personalized feedback to drivers (Marafie et al., 2021) and stimulating their receptivity for feedback by means of gamification techniques, such as leaderboards, rewards, and group forming (Musicant & Lotan, 2016; Shanly et al., 2018). Nambi et al. (2019) demonstrated several techniques for measuring maneuvers during Indian driving tests and claimed success using driver gaze monitoring to detect mirror scanning before lane changes. They further used a combination of camera, inertial, and GPS data for trajectory tracking.

Despite the widespread use of sensor measurements for driver assessment, there is limited research examining the validity of these methods from an algorithmic viewpoint and in such a way that it can be applied to the on-road driving test. Our study aims to fill this gap by presenting a series of algorithms for evaluating driving performance in these tests. These driving tests were carried out by experienced driving examiners, who emulated typical driving styles encountered during exams. The algorithms are explained in a step-by-step manner, allowing others to use them, and the code for this work is provided as supplementary material.

5.2. Methods

The data were collected in cooperation with the driving examiner training center of the Dutch Central Office of Driving Certification (CBR) in Leusden, Netherlands. At this training facility, driving examiner trainees are trained to become licensed driving examiners. A part of their training consists of on-road sessions, in which qualified instructors (active or former driving examiners) emulate driver behaviors commonly encountered at the driving test. The examiner trainee in the passenger seat receives no information about the role the driver takes and is expected to take the role of a real driving examiner and form a pass or fail verdict based on the acted driving style.

Data collection

The observations took place during 21 training sessions between 30 March 2022 and 13 April 2022. All drivers involved were asked for consent before the start of the experiment. The research was approved by the Human Research Ethics Committee of the Delft University of Technology, approval number 2302.

Acceleration data were recorded using the smartphone app Matlab Mobile version 9.1.2 (Mathworks, 2021) at a frequency of 10 Hz on an iPhone X (model A1865) and stored on the smartphone's local drive. The phone was placed on the backseat, with its back part fixed between the backrest and the seating area of the seat, and with its longitudinal axis and the car's longitudinal axis aligned. The screen faced upward, and the charging port pointed to the back of the car. Additionally, a GoPro Max was used to record video of the road ahead (1920×1080 pixels at 30 Hz). The video files contained embedded accelerometer recordings (at about 200 Hz) and GPS data (at about 17 Hz). These data logs were extracted from the video files using goprotelemetryextractor.com (Telemetry Overlay S.L., 2022). The appendix shows several example rows of data for both devices.

Driving tests

The drives all emulated a standard driving exam conducted by the CBR, having a duration of approximately 30 minutes. The drives started and ended at the same CBR location and sometimes involved driving on the same road segments. However, the drivers (i.e., 'test candidates') drove different routes, as the routes in Dutch driving tests are not set in advance but rather are determined by the examiner (in our study: the examiner trainee), based on factors such as traffic conditions and road closures.

The 21 driving tests emulated various driving styles, including 'a good driving candidate', 'a candidate who was close to passing or failing due to certain mistakes', 'a good candidate with poor viewing behavior or timing of actions', 'a slow candidate', 'a nonchalant candidate', 'a fast candidate', etc. Before each drive, the driver received a sheet containing the role description for the current ride and the intended result (pass/fail). The examiner in the passenger seat was blind to the instructed driving style.

From the 21 driving tests, we selected a total of nine driving tests (3 per car) because they allowed for systematic comparison. The other driving tests were of limited validity for further analysis because of various issues (e.g., inconsistent phone placement between drives in the same car, or an interruption of a drive). An overview of the nine selected driving tests is provided in Table 5.1. Each car was driven by a different driver, so there were a total of three drivers.

Table 5.1: Nine driving tests used in the analysis

No	Car	Weather	Date & Time	Emulated driving style of 'candidate'	Emulated test result
1	Peugeot 308 SW 2014	Rainy	4 Apr, 10:13	Difficulty with position on road	Fail
2	Peugeot 308 SW 2014	Rainy	4 Apr, 13:11	Difficulty with vehicle control and steady steering	Fail
3	Peugeot 308 SW 2014	Rainy	4 Apr, 13:58	Inappropriate timing; acting too early/late	Fail
4	Volkswagen T-Roc 2015	Sunny	11 Apr, 10:50	Inappropriate viewing behavior, engine stalling, position on road	Fail
5	Volkswagen T-Roc 2015	Sunny	11 Apr, 13:06	Aggressive/dangerous driving	Fail
6	Volkswagen T-Roc 2015	Sunny	11 Apr, 14:01	Desirable driving, but one large error (merging without looking)	Pass
7	Seat Ateca 2016	Cloudy	13 Apr, 09:21	Cautious/slow driving	Fail
8	Seat Ateca 2016	Cloudy	13 Apr, 10:06	Negligent viewing behavior	Fail
9	Seat Ateca 2016	Cloudy	13 Apr, 11:00	Desirable driving style, but occasional inappropriate looking	Pass

Data processing

The first step in processing data was to rotate the accelerometer data. Though care was taken to ensure that the GoPro and the phone's x- and y-axes were aligned with the car's frame, the devices still had non-negligible pitch and roll angles relative to the earth, which had to be corrected for.

We computed the orientation of the device (phone or GoPro) from the acceleration measurements in the three perpendicular directions (Figure 5.1). First, we computed the mean acceleration values in each direction over the entire drive. Next, the orientation of the device was computed using equations for determining orientation relative to the earth's gravitational field (Pedley, 2013). Our assumption here is that, although the car is moving and hence continuously experiencing accelerations, the accelerations due to vehicle motion can be expected to average out across the entire drive, leaving just the acceleration component caused by gravity. Next, we computed the roll and pitch angles of the device using arctangent functions (see Pedley, 2013, Eqs. 25 & 26). Then, a rotation

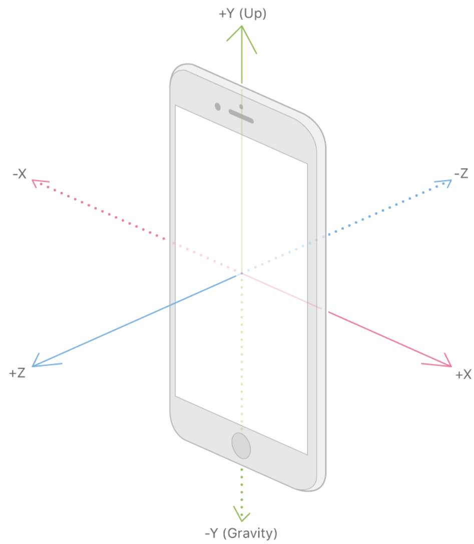


Figure 5.1: Phone coordinate system (Apple Inc., 2024).

matrix was computed, which was used to rotate the original acceleration measurements to their new orientation, aligned with the earth's downward gravitational field.

It is noted that the yaw angle is undetermined since it cannot be inferred based on the gravitation vector. In our calculations, the yaw angle with respect to the car was assumed to be 0 deg, which is a valid assumption since the GoPro and phone were positioned in this manner. Note that selecting another yaw angle will not change the accelerations in the rotated vertical direction (z), but will affect how the accelerations are distributed along the rotated x and y directions (while not changing their combined magnitude).

The measurement of acceleration in a moving vehicle is complicated by high-frequency vibrations caused by uneven roads and engine vibrations. How this noise protrudes in the signal depends on device placement (e.g., hard or soft surface) and the vehicle's damping properties.

A second-order Butterworth zero-phase filter with a cutoff frequency of 0.5 Hz was used to remove these vibrations, resulting in a smoother and more accurate representation of the car's acceleration (Figure 5.2). Figure 5.2 illustrates the effect of the filter on the rotated acceleration data in the y-direction, which represents the longitudinal direction of the car. The GPS speed of the GoPro is also shown in the figure.

Speed limit extraction

Speed limits were obtained using the Map Matching API provided by Mapbox (2023). This service takes a driven GPS path and returns the coordinates of the route that was most likely driven, including the speed limits on these roads. To obtain a robust response, the data were first downsampled to a sample rate of 5 s between points, as advised in the

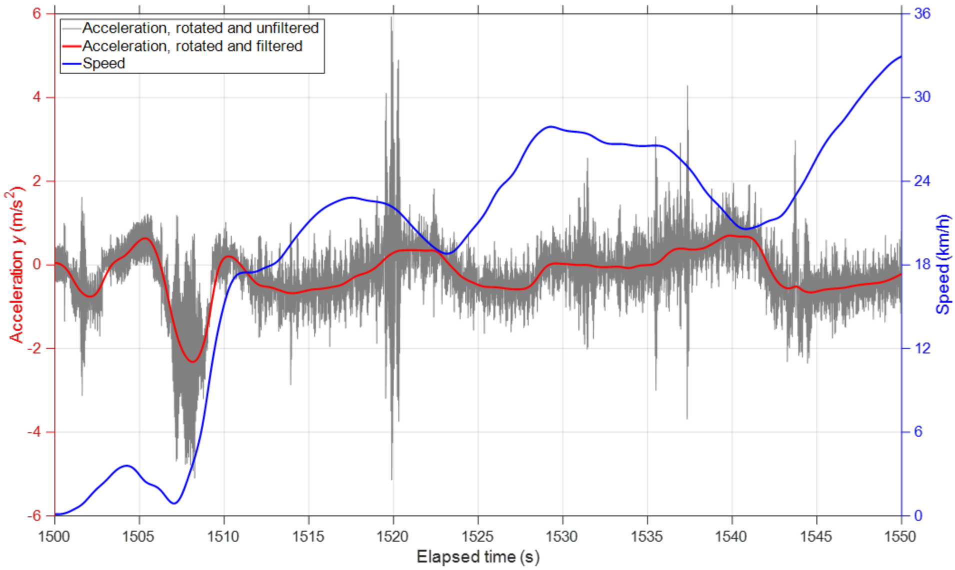


Figure 5.2: Illustration of the effect of low-pass filtering of the accelerometer data of the GoPro for a portion of Driving Test 1. The figure also shows the low-pass filtered vehicle speed recorded using the GPS of the GoPro. A negative acceleration means that the car is accelerating

API documentation. The paths were split into sections of 100 points each (the maximum number of coordinates allowed per request) and then merged after receiving the speed limits from the API. Then, the coordinates that were left out due to the resampling received the same speed limit of the nearest neighbor from the downsampled set. Upon visual inspection, it was found that the speed zones were correctly assigned, including short exceptions in residential districts, such as school areas. The API requests and processing were done using a Python script that is provided in the supplementary material.

Measures

After the above data pre-processing, five measures were calculated for each of the nine driving tests:

1. Macc: Mean absolute acceleration in the combined x and y directions (m/s^2). The rotated and filtered longitudinal (y) and lateral (x) accelerations were combined using the Pythagorean theorem.
2. Mjerk: Mean absolute jerk in the combined x and y directions (m/s^3). The rotated and filtered longitudinal (y) and lateral (x) accelerations were combined using the Pythagorean theorem. Next, the derivative was computed (i.e., jerk in m/s^3), and the mean absolute value was taken. The jerk can be seen as a measure of the abruptness of changing acceleration, and has been previously used in driver assessment (De Groot et al., 2011; Feng et al., 2017; Itkonen et al., 2017). It has been found to be associated with tailgating and traffic violations, self-reported accident involve-

ment (Bagdadi & Várhelyi, 2011), and recorded culpable crashes (Khorram et al., 2020). Figure 5.3 illustrates the meaning of jerk, where between 1505 and 1510 s, the driver accelerates; the onset and offset of the acceleration are accompanied by peaks in jerk.

3. Mspeed: Mean speed (km/h)
4. MspeedE: Speed limit exceedance (proportion of driving time)
5. HarshA: Harsh acceleration events, defined as the mean number of combined x and y acceleration threshold exceedances per hour of driving (# per hour). This measure was obtained by identifying all peaks in the combined acceleration and counting the number of peaks that exceeded a threshold value of 3 m/s^2. In the literature, there is no consensus about which threshold to choose (e.g., Khorram et al., 2020; Stipancic et al., 2016). Depending on the application and sample size, different threshold values may need to be adopted. Selecting a low threshold will yield a large number of threshold exceedances which may reflect driving style but may also involve false positives, such as accelerations due to road unevenness. Selecting a high threshold, on the other hand, risks missing important events and will reduce statistical power. After inspection of the acceleration signal, we opted for a threshold of 3 m/s^2. Indicatively, longitudinal decelerations of up to 3 m/s^2 are perceived as “reasonably comfortable” (Harwood, 2015, p. 41).

The accelerometer-based measures (Macc, Mjerk, HarshA) were computed for both the phone and GoPro, while the GPS-derived measures (Mspeed, MspeedE) were computed only for the GoPro. The reason for relying on the GoPro’s GPS measurement was that it was more accurate. During several trips, the phone’s receiver lost connectivity to the GPS satellites.

Data samples with a GPS GoPro speed below 3 km/h, indicative of the car being stationary or near-stationary, were excluded from the above driver assessment score computations. This was done as such instances, which may include special maneuvers or waiting at a traffic light, do not provide a valid representation of driving abilities.

5.3. Results

Table 5.2 displays the nine driving tests and the corresponding dependent measures. Firstly, it seems the type of car used in the test has an impact on the results. Specifically, Driving Tests 4 to 6, conducted in a Volkswagen, are distinct from the tests performed in a Peugeot (Driving Tests 1 to 3) or a Seat (Driving Tests 7 to 9), where the results of Driving Tests 4 to 6 show relatively high values for the mean absolute acceleration (Macc), mean absolute jerk (Mjerk), and harsh acceleration rate (HarshA).

Secondly, the instructed driving style seems to have an impact on the driving measures. Specifically, Driving Test 7, which was performed with a cautious driving style, was characterized by low scores on all measures compared to Driving Tests 8 and 9, which were conducted in the same car. Driving Test 7 had minimal speeding and was characterized by very few harsh acceleration events. Moreover, Driving Test 5, which was performed

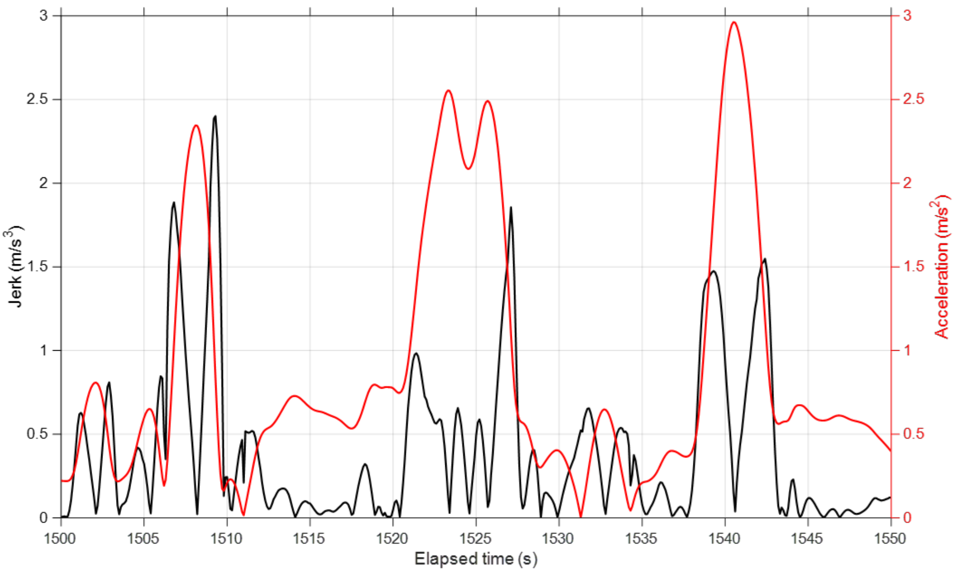


Figure 5.3: Jerk based on combined acceleration in the *xy*-plane for a portion of Driving Test 1. The selected time interval is the same as shown in Figure 5.2

Table 5.2: Dependent measure scores for the nine driving tests.

	Phone			GoPro				
Emulated driving style of 'candidate'	Macc	Mjerk	HarshA	Macc	Mjerk	Mspeed	MspeedE	HarshA
1 Difficulty with position on road	0.86	0.35	35.4	0.81	0.34	36.12	0.08	22.9
2 Difficulty with vehicle control and steady steering	0.78	0.32	34.7	0.73	0.32	36.89	0.10	23.1
3 Inappropriate timing; acting too early/late	0.77	0.33	30.1	0.69	0.31	36.12	0.13	23.2
4 Inappropriate looking, engine stalling, position on road	0.82	0.38	43.2	0.81	0.38	37.31	0.12	40.4
5 Aggressive/dangerous driving	0.81	0.41	54.4	0.81	0.41	47.72	0.43	53.9
6 Desirable driving, but one large error (merging without looking)	0.86	0.37	43.8	0.87	0.37	38.32	0.09	48.1
7 Cautious/slow driving	0.65	0.22	10.5	0.59	0.21	32.61	0.04	8.4
8 Negligent viewing behavior	0.77	0.27	33.3	0.72	0.26	37.03	0.11	16.7
9 Desirable driving style, but occasional inappropriate looking	0.82	0.30	41.0	0.73	0.29	38.35	0.08	28.9

Note. Color coding is applied per column from blue (lowest value) to white (median) to red (highest value).

with an instructed aggressive/dangerous driving style, had high scores on most of the measures compared to Driving Tests 4 and 6. An exception was the mean absolute acceleration (Macc), which was relatively low, at 0.81 m/s^2 for both the phone and GoPro.

We suspect that the driving route also impacted the measures observed. This is illustrated in Figure 5.4, which presents the absolute jerk in the *xy*-plane during a portion of Driving Test 5. While all the other driving tests were conducted in environments consisting mostly of roads with speed limits of 30 km/h, 50 km/h, and 100 km/h highways, the driver in Driving Test 5 chose a route through rural areas, primarily consisting of 60 km/h roads. Although the driver was tasked to drive with an aggressive driving style, there were often limited opportunities for aggressive driving other than exceeding the speed limit. That is, the driver drove most of the time on 60 km/h roads consisting of smooth asphalt, sometimes following behind another vehicle (between 1100 and 1400 s). There were some exceptions, such as the 800–1100 s interval, where the driver entered a small village. During these moments, the instructed driving style of the driver became more manifest, as shown by the spikes in the jerk.

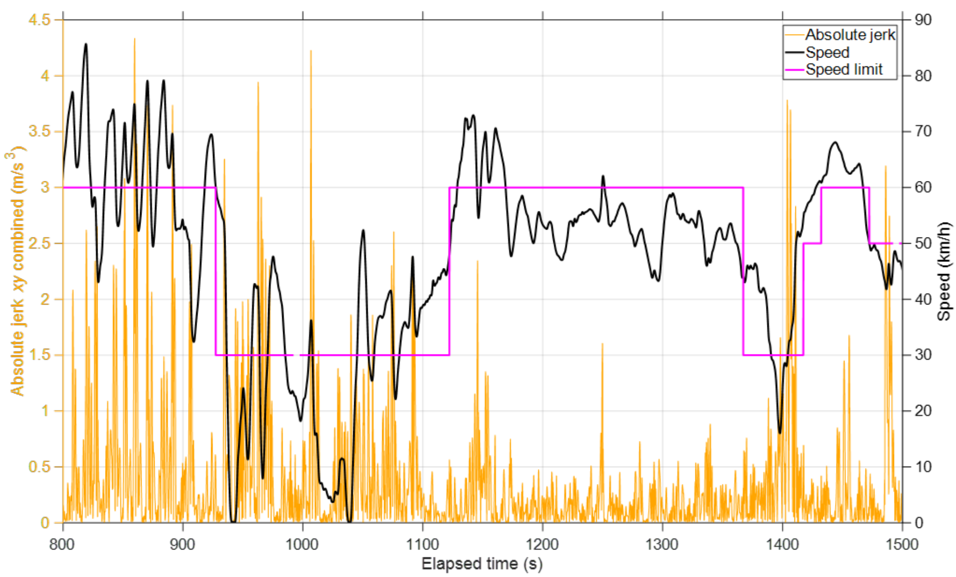


Figure 5.4: Mean absolute jerk in the *xy*-plane, vehicle speed recorded using GPS, and the speed limit for a portion of Driving Test 5.

Three measures, namely mean combined acceleration, mean absolute jerk, and mean number of harsh accelerations per hour of driving, were calculated using both the accelerometers in the phone and the accelerometers in the GoPro (Table 5.2). These measures were computed for the nine driving tests, and the results showed a high correlation between the two devices for all three measures (r = 0.929, r = 0.996, and r = 0.891, respectively). The correlations for mean combined acceleration and mean absolute jerk are illustrated in Figure 5.5.

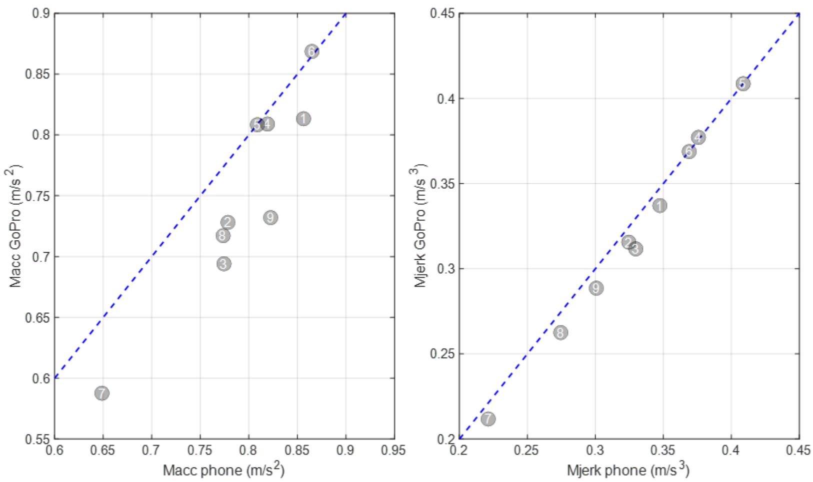


Figure 5.5: Scatter plot of mean absolute acceleration in the *xy*-plane (left) and mean absolute jerk in the *xy*-plane (right) for the GoPro versus the phone. The dashed line represents the line of unity. Each marker represents a driving test number.

Characterizing the route driven

As evidenced above, the evaluation of driving proficiency using accelerometers is influenced not only by the driving style of the candidate but also by the opportunities for high acceleration that are contingent upon the type of road being driven. Different road conditions require different driving behaviors: Driving on a straight road is different from driving on a road with multiple curves. This raises the issue of how to account for such variations in road conditions. Here, we draw upon prior research that used instrumented vehicles (such as Melman et al., 2021), which indicates that the assessment of driving behavior should be specific to a location, rather than relying on measures from the entire drive.

The number and curvature of curves can indicate the complexity of the road conditions (and the driver's ability to handle these conditions). The curves were extracted using the GPS measurements. First, the bearing of the car was computed from all subsequent GPS coordinates, assuming the earth is a sphere with a radius of 6371 km. The bearing was computed only if the vehicle speed exceeded 5 km/h (at low speeds, the distance between GPS points became too small to determine the bearing reliably). The bearing angle was filtered with a median filter (time window: 2 s), the gaps in the data caused by GPS speeds below 5 km/h were linearly interpolated, and a low-pass Butterworth filter (cutoff frequency of 0.5 Hz) was applied. The effect of the filtering is shown in Figure 5.6, which compares the bearing before and after filtering. The bearing was differentiated to obtain bearing rate. To prevent abrupt jumps in the angle due to its limited range between 0 and 360 degrees, the unwrap function was used before differentiating, which replaces jumps greater than 180 degrees by their 360-degree complement, resulting in a continuous line. As differentiating the data amplifies noise, the bearing rate was filtered using a median

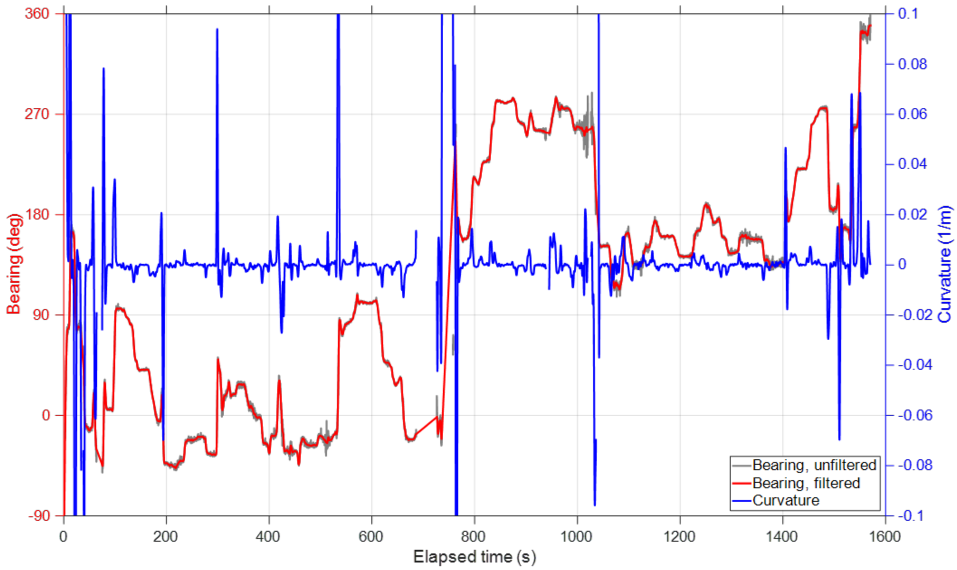


Figure 5.6: Calculated bearing before and after filtering, and path curvature, for Driving Test 5. A bearing angle of 90 deg corresponds to driving northbound, an angle of 180 corresponds to driving westbound, etc. The y-axis for the curvature was constrained to -0.1 and 0.1, corresponding to a turn radius of 10 m. Note that high or low curvature values occurred when the vehicle was driving slowly (see Figure 5.4).

filter and a Butterworth filter with the same parameters as mentioned above. Finally, the curvature of the car's path was computed by dividing the bearing rate by the momentary GPS speed. To further reduce any noise in the curvature data, a Butterworth filter with a cutoff frequency of 0.5 Hz was applied.

We extracted the peaks in the curvature data to count the number of curved paths of the car as well as the moments of those curves. Then, several route statistics were computed, which were tabulated in Table 5.3:

- Proportion of time driven under each speed limit (30, 50, 60, 80, or 100 km/h). These five values were normalized so that the total is equal to 100%.
- MildC#: The number of mild curves (absolute curvature between 0.005 and 0.05) per hour of driving. This corresponds to a turn radius between 20 and 200 m.
- SharpC#: The number of sharp curves (absolute curvature of 0.05 or greater) per hour of driving. This corresponds to a turn radius of 20 m or less, and can be seen as turning at an intersection, turning around, etc.
- MildC Macc: Mean absolute acceleration in the combined x and y directions (m/s^2), averaged across the mild curves.
- SharpC Macc: Mean absolute acceleration in the combined x and y directions (m/s^2), averaged across the sharp curves.

Table 5.3: Route statistics, computed from the GoPro GPS data.

Emulated driving style of 'candidate'	Speed limit (km/h) 30	50	60	80	100	MildC#	SharpC#	MildC Macc	SharpC Macc
1 Difficulty with position on road	0.28	0.53	0.04	0.06	0.10	114.3	43.7	1.18	2.26
2 Difficulty with vehicle control and steady steering	0.27	0.50	0.05	0.06	0.12	101.6	48.5	0.94	1.99
3 Inappropriate timing; acting too early/late	0.33	0.47	0.05	0.07	0.08	108.8	39.4	0.93	1.90
4 Inappropriate looking, engine stalling, position on road	0.30	0.47	0.04	0.07	0.12	130.8	45.2	1.07	2.87
5 Aggressive/dangerous driving	0.20	0.11	0.69	0.00	0.00	90.6	17.1	1.65	2.68
6 Desirable driving, but one large error (merging without looking)	0.32	0.38	0.04	0.01	0.26	101.6	61.5	1.30	2.35
7 Cautious/slow driving	0.29	0.49	0.05	0.06	0.11	125.9	35.7	0.74	1.81
8 Negligent viewing behavior	0.26	0.42	0.06	0.11	0.15	130.5	50.0	1.05	1.77
9 Desirable driving style, but occasional inappropriate looking	0.30	0.25	0.22	0.06	0.18	118.0	48.2	1.26	2.01

Note. Color coding is applied per column from blue (lowest value) to white (median) to red (highest value).

The statistics for Driving Test 5, referred to as the 'aggressive/dangerous' drive, show that 77% of the drive took place in a 60 km/h zone. Although the number of curves was low, the acceleration in these curves was relatively high compared to other driving tests. This highlights the importance of presenting driving examiners with both objective performance measures (as shown in Table 5.2) and route statistics (as shown in Table 5.3) in order to provide a more complete understanding of the driver's behavior. The combination of these two tables makes it clear that the driver in Driving Test 5 was driving aggressively in relatively easy road conditions.

5.4. Discussion

We presented algorithms that could help distinguish between overcautious, normal, and aggressive driving during the driving test. We solely relied on accelerometer and GPS data and found that these sensors were enough to identify the overcautious and aggressive driving styles. The percentage of driving time exceeding the speed limit, mean jerk, and mean harsh acceleration rate were effective measures in this discrimination. However, mean absolute acceleration across the entire drive was not a clear indicator, as it can vary greatly depending on the eventfulness of a drive, such as the presence of curves. To overcome this issue, we proposed additional measures, namely the speed limit distribution, mild and sharp curve rate, and mean absolute acceleration in curves, to assess the route driven.

The current study provides several insights into the use of accelerometers and GPS. One of our observations was that the combination of x and y acceleration proved robust. In particular, the mean absolute jerk measurement demonstrated a particularly high consistency between a smartphone and a GoPro ($r = 0.996$; see Figure 5.5, right), even though they employed a different measurement unit and were positioned differently in the car (flat on the back seat vs. upright on the dashboard). The robustness of the jerk measure could be attributed to it reflecting changes in acceleration and thus being less susceptible to possible offsets in the acceleration measurement.

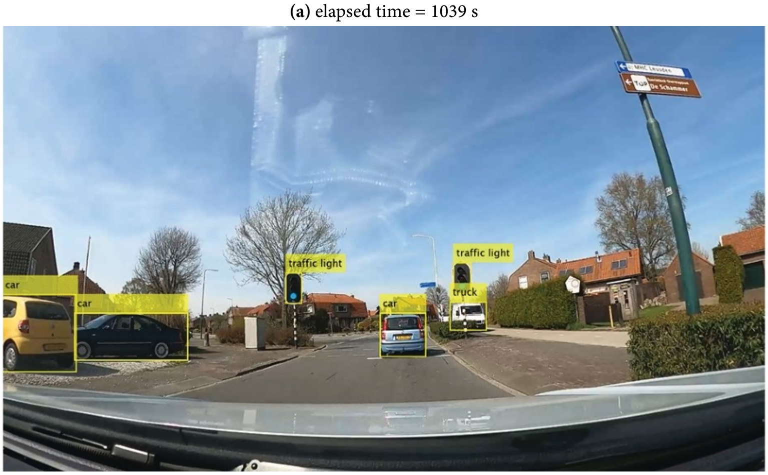
Previous interviews revealed that driving examiners could benefit from data-driven support, particularly in communicating their evaluation to test candidates (Driessen et al., 2021). The current study demonstrates that it is possible to generate numerical scores that reflect the driving style and the dynamic nature of the route. These scores could be presented in graph form or in a map format, for example, by visualizing the GPS-recorded route along with color-coded speed, deviation from the speed limit, harsh accelerations, or moving averages of accelerations or jerk values. If dashcam images are used, it may also be feasible to automatically identify and replay segments in the video where driving behavior was particularly noteworthy (e.g., around the moment of the highest recorded acceleration or jerk). Such techniques have the potential to aid the examiner in explaining their evaluation.

The current study highlights the effectiveness of using GPS and accelerometers to distinguish between slow and fast driving styles. Some drivers were instructed to exhibit poor viewing behavior but otherwise normal driving behavior. Indeed, their driving behavior, as measured by acceleration and speed, appeared normal (Table 5.2: Driving Tests 4, 8, 9). Incorrect viewing behavior is a common cause of failing the driving exam (De Winter et al., 2008; UK Government, 2022), and while it is possible that poor anticipation skills may manifest as harsh accelerations and high jerk (Fisher et al., 2002; Parmet et al., 2015), this relationship is only indirect. Other types of sensors may have to be explored to support the assessment of a candidate's viewing behavior. For example, eye-tracking technology is feasible: eye-tracking systems that detect visual distraction are becoming available in modern cars (e.g., DS Automobiles, 2023), and several recent research studies have used eye-tracking in combination with object detection to establish at which object the driver was looking (Kim et al., 2020; Qin et al., 2022).

As previously noted, accelerometer and GPS data alone offer a limited perspective on a driver's performance as they fail to capture the driver's interaction with other road users. To gain a better understanding of driving behavior, object detection based on camera images, similar to those employed by automated vehicles, may be necessary (see Figure 5.7 for an illustration). An online experiment has revealed that the number of identified individuals and the bounding boxes surrounding other road users can predict perceived risk (De Winter et al., 2023). Figure 5.8 provides examples of how computer vision techniques could be used, namely by counting the number of persons and estimating headway to the car in front using the width of the bounding box (for more, see Rezaei et al., 2021). Automated identification of high-risk scenarios, such as passing another road user too closely, may help examiners form a more objective assessment of driving behavior. However, this type of approach towards the driving test would require further research and validation.

A potential issue in the driver assessment process is that the examiner in the passenger seat occasionally applied the secondary brake pedal to intervene in dangerous situations, which complicates the analysis of the accelerometer data. Future investigations may need to record brake pedal inputs of both the driver (i.e., candidate) and the passenger (i.e., examiner), to isolate their respective contributions. This could also aid in the debriefing session following the driving test, where the examiner could replay the moments of intervention that occurred during the drive, as identified automatically based on the examiner's brake pedal inputs.

(a) elapsed time = 1039 s



(b) elapsed time = 1400 s

Figure 5.7: Bounding boxes generated using the YOLOv4 algorithm for Driving Test 5. A YOLOv4 model (Bochkovskiy et al., 2020) pretrained on the COCO dataset was used (Lin et al., 2014; obtained from sbairagy-MW, 2021)

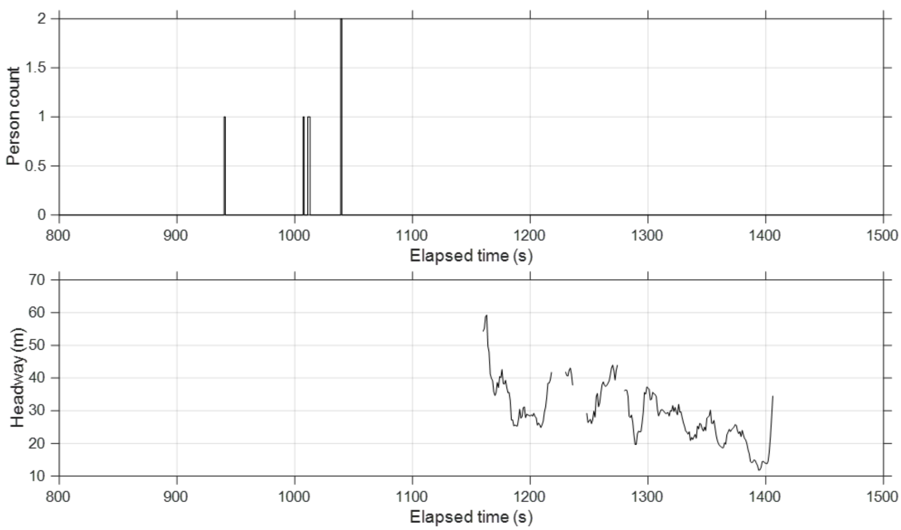


Figure 5.8: Number of persons (top) and headway to the vehicle in front (bottom) measured at a frequency of 1 Hz for a portion of Driving Test 5 (same portion as shown in Figure 5.4). In this 700-s interval, the driver encountered a number of persons (see Figure 5.7a), proceeded to a 60 km/h road, and followed another car for approximately 250 s (see Figure 5.7b). Only objects straight in front of the ego-vehicle were considered (top figure: a 400-pixel horizontal range, bottom figure: a 50-pixel horizontal range. Gaps in the headway of 5 s or less were spline-interpolated.

A limitation of the current study is its small sample size, comprising nine driving tests and three drivers. In studies analyzing naturalistic driving data, larger sample sizes are typically recommended to ensure that external factors such as weather impact and traffic variability over time are adequately represented across the sample. It is advisable to augment the current analysis with larger sample sizes in future research. Despite the small sample size, this study can serve as an initial framework for designing algorithms aimed at detecting anomalous driving styles typically associated with novice or trainee drivers.

A second limitation is that the study assumed that experienced driving instructors are able to realistically imitate the driving styles of test candidates. Additionally, the driving examiners may have focused on specific, extreme scenarios, which may not be representative of typical driving tests. For example, it has been argued that the driving test is a test of driving skill rather than driving style, and that individuals attempting to obtain their driving license are unlikely to engage in e.g., excessive speeding (Alsharef et al., 2021; Senserrick & Haworth, 2005). Another limitation is that traffic density in the test region was relatively low compared to dense city environments.

Apart from driver testing, we see further use in driving data collection during driver training, either for student drivers prior to obtaining their driver's license (Driessen et al., 2021) or for more experienced drivers who engage in self-coaching (Takeda et al., 2012). Similarly, a government report on reforming the Dutch driver education system by Roemer (2021) pointed out the value of using data to help students reflect on their learning progress. Information gathered prior to the final exam may assist driving schools in determining the candidate's readiness for the driving test, thereby reducing the number of unsuccessful test takers (Alsharef et al., 2022). Additionally, at present, there is no motivation or requirement for individuals to maintain their driving skills after obtaining a license. The use of data could potentially address this issue.

To determine if recorded data can predict driving test results, collecting data from more driving tests is recommended. For example, sensor data, dash-camera images, and map data could be fed to a machine learning algorithm that predicts pass or fail outcomes. However, it is noted that predicting the test outcome may be challenging due to the fact that test candidates tend to apply for the exam when they have just that amount of driving experience where they have a moderate probability of passing (Baughan et al., 2005).

5.5. Conclusions and outlook

The study used accelerometer and GPS data to distinguish between slow, normal, and aggressive driving during driving tests. The findings show that these sensors are sufficient to identify different driving styles, and that the percentage of driving time exceeding the speed limit, mean jerk, and mean harsh acceleration rate are effective measures in this discrimination. However, the study also highlights the limitations of using these sensors alone, as they fail to provide insight into the driver's viewing behavior and interaction with other road users. Future investigations may address this issue by incorporating computer vision methods.

The study concludes that the use of GPS and accelerometers has the potential to aid driving examiners in their assessments and communication with test candidates. However, more research is needed, as the number of driving tests was small and there are limitations associated with experienced driving instructors imitating the driving styles of test candidates. Instead of using ex-examiners, future studies could record data from real candidates in driving exams or lessons, provided proper precautions regarding consent and data protection are taken.

It is also acknowledged that the current data proved to be specific to the vehicle used, as different vehicles have varying spring-damper characteristics, engine power, and therefore different acceleration capabilities. This can influence the accelerometer readings and should be considered when interpreting the results.

The use of sensors may contribute to increasing the efficiency of the driving test, and potentially provide valuable data for improving driver training programs. Data-based driving assessments may also prove useful to pre-license driver training and post-license driver monitoring.

Acknowledgments

This research is supported by Transitions and Behaviour grant 403.19.243 ("Towards Safe Mobility for All: A Data-Driven Approach"), provided by the Netherlands Organization for Scientific Research (NWO). We would like to thank the Dutch Central Office of Driving Certification (CBR) for making this research possible.

Data availability

Raw data and analysis scripts are available at https://doi.org/10.4121/3bb2f535-59ec-426c-b69a-e113810543b2

References

Alsharef, A., Albert, A., & Bhandari, S. (2022). Safety challenges experienced by driver license examiners and related safety measures. Safety Science, 154, 105808. https://doi.org/10.1016/j.ssci.2022.105808

Alsharef, A., Albert, A., Uddin, S. M. J., Kittur, N. B., Chavan, S., & Jaselskis, E. (2021). Work-related safety incidents among driver license examiners. Safety Science, 140, 105304. https://doi.org/10.1016/j.ssci.2021.105304

Apple Inc. (2024). Documentation. https://developer.apple.com/documentation/arkit/arworldalignment/arworldalignmentgravity

Bagdadi, O., & Várhelyi, A. (2011). Jerky driving—an indicator of accident proneness? Accident Analysis & Prevention, 43, 1359–1363. https://doi.org/10.1016/j.aap.2011.02.009

Bates, L., Anderson, L., Rodwell, D., & Blais, E. (2020). A qualitative study of young drivers and deterrence based road policing. Transportation Research Part F: Traffic Psychology and Behaviour, 71, 110–118. https://doi.org/10.1016/j.trf.2020.04.003

Baughan, C. J., Gregersen, N. P., Hendrix, M., & Keskinen, E. (2005). Towards European standards for testing (Final Report). Brussels: The International Commission for Driving Testing. https://www.cieca.eu/project/27

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv. https://doi.org/10.48550/arXiv.2004.10934

Cai, M., Yazdi, M. A. A., Mehdizadeh, A., Hu, Q., Vinel, A., Davis, K., Xian, H., Megahed, F. M., & Rigdon, S. E. (2021). The association between crashes and safety-critical events: Synthesized evidence from crash reports and naturalistic driving data among commercial truck drivers. Transportation Research Part C: Emerging Technologies, 126, 103016. https://doi.org/10.1016/j.trc.2021.103016

Carlos, M. R., González, L. C., Wahlström, J., Ramírez, G., Martínez, F., & Runger, G. (2020). How smartphone accelerometers reveal aggressive driving behavior?—The key is the representation. IEEE Transactions on Intelligent Transportation Systems, 21, 3377–3387. https://doi.org/10.1109/TITS.2019.2926639

Chan, T. K., Chin, C. S., Chen, H., & Zhong, X. (2020). A comprehensive review of driver behavior analysis utilizing smartphones. IEEE Transactions on Intelligent Transportation Systems, 21, 4444–4475. https://doi.org/10.1109/TITS.2019.2940481

Curry, A. E., Metzger, K. B., Williams, A. F., & Tefft, B. C. (2017). Comparison of older and younger novice driver crash rates: Informing the need for extended Graduated Driver Licensing restrictions. Accident Analysis & Prevention, 108, 66–73. https://doi.org/10.1016/j.aap.2017.08.015

De Groot, S., De Winter, J. C. F., Mulder, M., & Wieringa, P. A. (2011). Nonvestibular motion cueing in a fixed-base driving simulator: Effects on driver braking and cornering performance. Presence, 20, 117–142. https://doi.org/10.1162/pres_a_00039

De Waard, D., & Rooijers, T. (1994). An experimental study to evaluate the effectiveness of different methods and intensities of law enforcement on driving speed on motorways. Accident Analysis & Prevention, 26, 751–765. https://doi.org/10.1016/0001-4575(94)90052-3

De Winter, J., Hoogmoed, J., Stapel, J., Dodou, D., & Bazilinskyy, P. (2023). Predicting perceived risk of traffic scenes using computer vision. Transportation Research Part F: Traffic Psychology and Behaviour, 93, 235–247. https://doi.org/10.1016/j.trf.2023.01.014

De Winter, J. C. F., De Groot, S., Van Loenhout, M. J., Van Leeuwen, A., Do, P., Wieringa, P. A., & Mulder, M. (2008). Feedback on mirror-checking during simulation-based driver training. Proceedings of the 27th European Annual Conference on Human Decision Making and Manual Control, Delft, the Netherlands.

Driessen, T., Dodou, D., De Waard, D., & De Winter, J. C. F. (2024). Predicting damage incidents, fines, and fuel consumption from truck driver data: A study from the Netherlands. Transportation Research Record. https://doi.org/10.1177/03611981231211897

Driessen, T., Picco, A., Dodou, D., De Waard, D., & De Winter, J. (2021). Driving examiners' views on data-driven assessment of test candidates: An interview study. Transportation Research Part F: Traffic Psychology and Behaviour, 83, 60–79. https://doi.org/10.1016/j.trf.2021.09.021

DS Automobiles. (2023). DS driver attention monitoring. https://www.dsautomobiles.co.uk/ds-experience/discover/news/ds-automobiles-anti-fatigue-technology.html

Fell, J. C., Jones, K., Romano, E., & Voas, R. (2011). An evaluation of graduated driver licensing effects on fatal crash involvements of young drivers in the United States. Traffic Injury Prevention, 12, 423–431. https://doi.org/10.1080/15389588.2011.588296

Feng, F., Bao, S., Sayer, J. R., Flannagan, C., Manser, M., & Wunderlich, R. (2017). Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. Accident Analysis & Prevention, 104, 125–136. https://doi.org/10.1016/j.aap.2017.04.012

Fisher, D. L., Laurie, N. E., Glaser, R., Connerney, K., Pollatsek, A., Duffy, S. A., & Brock, J. (2002). Use of a fixed-base driving simulator to evaluate the effects of experience and PC-based risk awareness training on drivers' decisions. Human Factors, 44, 287–302. https://doi.org/10.1518/0018720024497853

Foxe, K. (2020). More than 50 incidents of aggression by driving test candidates reported. https://www.irishexaminer.com/news/arid-40195446.html

Harwood, D. W. (1992) Traffic and vehicle operating characteristics. In J. L. Pline (Ed.), Traffic Engineering Handbook, Fourth Edition (pp. 29–58). Englewood Cliffs, NJ: Prentice Hall.

Helman, S., Vlakveld, W., Fildes, B., Oxley, J., Fernández-Medina, K., & Weekley, J. (2017). Study on driver training, testing and medical fitness (Final report). https://doi.org/10.2832/446103

Hunter, M., Saldivar-Carranza, E., Desai, J., Mathew, J. K., Li, H., & Bullock, D. M. (2021). A proactive approach to evaluating intersection safety using hard-braking data. Journal of Big Data Analytics in Transportation, 3, 81–94. https://doi.org/10.1007/s42421-021-00039-y

Itkonen, T. H., Pekkanen, J., Lappi, O., Kosonen, I., Luttinen, T., & Summala, H. (2017). Trade-off between jerk and time headway as an indicator of driving style. PLOS ONE, 12, e0185856. https://doi.org/10.1371/journal.pone.0185856

Johnson, D. A., & Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. Proceedings of the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), Washington, DC, 1609–1615. https://doi.org/10.1109/ITSC.2011.6083078

Khorram, B., Af Wåhlberg, A. E., & Tavakoli Kashani, A. (2020). Longitudinal jerk and celeration as measures of safety in bus rapid transit drivers in Tehran. Theoretical Issues in Ergonomics Science, 21, 577–594. https://doi.org/10.1080/1463922X.2020.1719228

Kim, H., Martin, S., Tawari, A., Misu, T., & Gabbard, J. L. (2020). Toward real-time estimation of driver situation awareness: An eye-tracking approach based on moving objects of interest. Proceedings of the 2020 IEEE Intelligent Vehicles Symposium, Las Vegas, NV, 1035–1041. https://doi.org/10.1109/IV47402.2020.9304770

Lajunen, T., Sullman, M. J. M., & Gaygısız, E. (2022). Self-assessed driving skills and risky driver behaviour among young drivers: A cross-sectional study. Frontiers in Psychology, 13, 840269. https://doi.org/10.3389/fpsyg.2022.840269

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Ma, Y.-L., Zhu, X., Hu, X., & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A: Policy and Practice, 113, 243–258. https://doi.org/10.1016/j.tra.2018.04.013

Mapbox. (2023). Map Matching API. https://docs.mapbox.com/api/navigation/map-matching

Marafie, Z., Lin, K.-J., Wang, D., Lyu, H., Liu, Y., Meng, Y., & Ma, J. (2021). AutoCoach: An intelligent driver behavior feedback agent with personality-based driver models. Electronics, 10, 1361. https://doi.org/10.3390/electronics10111361

Mathworks. (2021). Sensor data collection with MATLAB Mobile. https://nl.mathworks.com/help/matlabmobile/ug/sensor-data-collection-with-matlab-mobile.html

Melman, T., Abbink, D., Mouton, X., Tapus, A., & de Winter, J. (2021). Multivariate and location-specific correlates of fuel consumption: A test track study. Transportation Research Part D: Transport and Environment, 92, 102627. https://doi.org/10.1016/j.trd.2020.102627

Mueller, A. S., & Cicchino, J. B. (2022). Teen driver crashes potentially preventable by crash avoidance features and teen-driver-specific safety technologies. Journal of Safety Research, 81, 305–312. https://doi.org/10.1016/j.jsr.2022.03.007

Musicant, O., & Lotan, T. (2016). Can novice drivers be motivated to use a smartphone based app that monitors their behavior? Transportation Research Part F: Traffic Psychology and Behaviour, 42, 544–557. https://doi.org/10.1016/j.trf.2015.10.023

Nambi, A. U., Mehta, I., Ghosh, A., Lingam, V., & Padmanabhan, V. N. (2019). ALT: Towards automating driver license testing using smartphones. Proceedings of the 17th Conference on Embedded Networked Sensor Systems, New York, NY, 29–42. https://doi.org/10.1145/3356250.3360037

Othman, W., Kashevnik, A., Hamoud, B., & Shilov, N. (2022). DriverSVT: Smartphone-measured vehicle telemetry data for driver state identification. Data, 7, 181. https://doi.org/10.3390/data7120181

Parmet, Y., Borowsky, A., Yona, O., & Oron-Gilad, T. (2015). Driving speed of young novice and experienced drivers in simulated hazard anticipation scenes. Human Factors, 57, 311–328. https://doi.org/10.1177/0018720814548220

Pedley, M. (2013). Tilt sensing using a three-axis accelerometer. https://www.nxp.com/docs/en/application-note/AN3461.pdf

Poirier, B., Blais, E., & Faubert, C. (2018). Graduated driver licensing and differential deterrence: The effect of license type on intentions to violate road rules. Accident Analysis & Prevention, 110, 62–70. https://doi.org/

10.1016/j.aap.2017.10.001

Qin, L., Shi, Y., He, Y., Zhang, J., Zhang, X., Li, Y., Deng, T., & Yan, H. (2022). ID-YOLO: Real-time salient object detection based on the driver's fixation region. IEEE Transactions on Intelligent Transportation Systems, 23, 15898–15908. https://doi.org/10.1109/TITS.2022.3146271

Rezaei, M., Azarmi, M., & Mir, F. M. P. (2021). Traffic-Net: 3D traffic monitoring using a single camera. arXiv. https://doi.org/10.48550/arXiv.2109.09165

Roemer, E. (2021). Van rijles naar rijonderwijs. Advies verbetering autorijscholenbranche [From driving lessons to driving education. Advice for improving the driving School industry]. The Hague, The Netherlands: Ministry of Infrastructure and Water Management. https://www.rijksoverheid.nl/documenten/rapporten/2021/04/14/bijlage-1-van-rijles-naar-rijonderwijs-advies-verbeteren-autorijscholenbranche

Rolison, J. J., & Moutari, S. (2020). Combinations of factors contribute to young driver crashes. Journal of Safety Research, 73, 171–177. https://doi.org/10.1016/j.jsr.2020.02.017

sbairagy-MW. (2021). Pretrained-yolo-v4. https://github.com/matlab-deep-learning/pretrained-yolo-v4

Senserrick, T., & Haworth, N. (2005). Review of literature regarding national and international young driver training, licensing and regulatory systems (Report No. 239), Monash University. https://www.monash.edu/muarc/archive/our-publications/reports/muarc239

Shanly, C., Ieti, M., Warren, I., & Sun, J. (2018). BackPocketDriver-A mobile app to enhance safe driving for youth. Proceedings of the International Conference on Software Engineering and Knowledge Engineering, San Francisco Bay, CA, 246–245. https://doi.org/10.18293/SEKE2018-011

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2016). The who and where of road safety: Extracting surrogate indicators from smartphone-collected GPS data in urban environments. Transportation Research Board 95th Annual Meeting, Washington, DC. http://n.saunier.free.fr/saunier/stock/stipancic16who-where.pdf

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. Accident Analysis & Prevention, 115, 160–169. https://doi.org/10.1016/j.aap.2018.03.005

Takeda, K., Miyajima, C., Suzuki, T., Angkititrakul, P., Kurumida, K., Kuroyanagi, Y., Ishikawa, H., Terashima, R., Wakita, T., Oikawa, M., & Komada, Y. (2012). Self-coaching system based on recorded driving data: Learning from one's experiences. IEEE Transactions on Intelligent Transportation Systems, 13, 1821–1831. https://doi.org/10.1109/TITS.2012.2205917

Tay, R. (2005). The effectiveness of enforcement and publicity campaigns on serious crashes involving young male drivers: Are drink driving and speeding similar? Accident Analysis & Prevention, 37, 922–929. https://doi.org/10.1016/j.aap.2005.04.010

Telemetry Overlay S.L. (2022). GPS & Data Visualization. https://goprotelemetryextractor.com

UK Government. (2022). Top 10 reasons for failing the driving test in Great Britain. https://www.gov.uk/government/publications/top-10-reasons-for-failing-the-driving-test/top-10-reasons-for-failing-the-driving-test-in-great-britain

Weast, R. A., & Monfort, S. S. (2021). Characteristics of vehicles driven by teens and adults killed in crashes, 2013–2017. Journal of Safety Research, 77, 263–267. https://doi.org/10.1016/j.jsr.2021.03.010

Williams, A. F. (2017). Graduated driver licensing (GDL) in the United States in 2016: A literature review and commentary. Journal of Safety Research, 63, 29–41. https://doi.org/10.1016/j.jsr.2017.08.010

Appendix: Example rows of raw data collected by iPhone X and GoPro Max

Tables 5.4 - 5.6 show example rows of data recorded.

Table 5.4: Example of phone acceleration data (measurement frequency: 10 Hz).

Timestamp	X (m/s^2)	Y (m/s^2)	Z (m/s^2)
04-04-2022 10:13:38.186	-0.236	6.940	7.216
04-04-2022 10:13:38.287	-0.626	6.955	7.348
04-04-2022 10:13:38.387	0.126	7.619	6.059
04-04-2022 10:13:38.488	0.869	7.293	7.691
04-04-2022 10:13:38.589	1.997	6.670	5.816
:	:	:	:
04-04-2022 10:46:42.163	-1.219	4.693	8.316
04-04-2022 10:46:42.264	-1.544	4.919	8.676
04-04-2022 10:46:42.364	2.094	6.665	6.603
04-04-2022 10:46:42.465	-1.113	6.537	7.096
04-04-2022 10:46:42.566	-1.338	5.267	8.178

Table 5.5: Example of GoPro acceleration data (measurement frequency: approximately 200 Hz, except for the first few samples).

Timestamp	X (m/s^2)	Y (m/s^2)	Z (m/s^2)
2022-04-04 10:13:29.024	-1.012	-1.122	-9.827
2022-04-04 10:13:29.036	-1.405	-1.275	-9.508
2022-04-04 10:13:29.048	-1.309	-1.048	-9.340
2022-04-04 10:13:29.061	-0.815	-0.686	-9.892
2022-04-04 10:13:29.073	-0.206	-0.341	-10.251
:	:	:	:
2022-04-04 10:46:57.012	-0.180	-0.638	-9.765
2022-04-04 10:46:57.017	-0.751	-1.031	-9.892
2022-04-04 10:46:57.022	-1.290	-1.185	-9.731
2022-04-04 10:46:57.028	-1.393	-1.048	-9.326
2022-04-04 10:46:57.033	-1.028	-0.847	-9.925

Table 5.6: Example of GoPro GPS Data (the measurement frequency was fluctuating but averaged approximately 17 Hz).

Timestamp	Latitude	Longitude	Altitude	2D Speed
2022-04-04 10:37:05.715	52.144	5.388	-2.477	11.105
2022-04-04 10:37:05.773	52.144	5.388	-2.516	11.129
2022-04-04 10:37:05.831	52.144	5.388	-2.513	11.226
2022-04-04 10:37:05.889	52.144	5.388	-2.502	11.219
2022-04-04 10:37:05.947	52.144	5.388	-2.532	11.188
:	:	:	:	:
2022-04-04 10:46:57.108	52.144	5.423	-4.613	0.010
2022-04-04 10:46:57.189	52.144	5.423	-4.619	0.070
2022-04-04 10:46:57.269	52.144	5.423	-4.625	0.050
2022-04-04 10:46:57.349	52.144	5.423	-4.661	0.040
2022-04-04 10:46:57.430	52.144	5.424	-4.652	0.020

6

Predicting Damage Incidents, Fines, and Fuel Consumption from Telematics Data

This chapter was published as: Driessen, T., Dodou, D., Waard, D. D., & Winter, J. D. (2024). Predicting damage incidents, fines, and fuel consumption from truck driver data: A study from the Netherlands. Transportation Research Record: Journal of the Transportation Research Board, 2678(7). https://doi.org/10.1177/03611981231211897

Abstract

Trucks are disproportionately involved in fatal traffic accidents and contribute significantly to CO_2 emissions. Gathering data from trucks presents a unique opportunity for estimating driver-specific costs associated with truck operation. Although research has been published on the predictive validity of driver data, such as in the contexts of pay-how-you-drive insurance and naturalistic driving studies, the investigation into how telematics data and the negative consequences of truck driving remains limited. In the present study, driving data from 180 truck drivers, collected over a 2-year period, were examined to predict damage incidents, traffic fines, and fuel consumption. Correlation analysis revealed that the number of fines and damage incidents could be predicted based on the number of harsh braking events per hour of driving, whereas fuel consumption was predicted by engine torque exceedances. Our analysis also sheds light on the impact of covariates, including the engine capacity of the truck operated and time of day, among others. We conclude that the damage incidents and fines incurred by truck drivers can be predicted not only from their number of harsh decelerations but also through driving demands that extend beyond the driver's immediate control. It is recommended that transportation companies adopt a systemic approach to mitigating truck-driving-related expenses.

6.1. Introduction

In 2019, heavy goods vehicles, henceforth denoted as "trucks", were implicated in 14% of all traffic fatalities within the European Union (European Road Safety Observatory, 2021). Casualties in truck collisions predominantly involve the opposing party. Specifically, in 2019, 26 truck occupants succumbed to collisions with cars, whereas 1,557 car occupants died in accidents involving trucks (European Commission, 2021). Apart from the road safety implications, truck operations significantly affect the environment, contributing approximately 21% of the European Union's road transport CO_2 emissions (European Environment Agency, 2022). These environmental impacts are also evident in the operational expenditures of trucking companies; an examination of truck driving expenses in the United States revealed that 22% was allocated to fuel, 9% to damage repairs and maintenance, and 4% to insurance premiums (American Transportation Research Institute, 2020). Consequently, identifying the determinants of accidents, damage incidents, and fuel consumption would be profoundly beneficial for both society and transportation companies.

The increasing accessibility of in-vehicle data recordings has allowed novel approaches to investigating the determinants of accidents. Among passenger vehicles, pay-as-you-drive (PAYD) insurance incorporating in-vehicle data recording has gained traction (e.g., Baecke & Bocca, 2017; Bian et al., 2018). Using data from 1,600 vehicles participating in PAYD programs, Paefgen et al. (2014) demonstrated that urban driving posed a relatively elevated accident risk, whereas highway driving presented the lowest risk per kilometer driven. Moreover, Verbelen et al. (2018), in their analysis of 10,406 policyholders, discovered that although males are typically considered higher-risk drivers than females, gender was no longer a significant factor when telematics data, such as mileage, were incorporated as accident involvement predictors.

Tselentis et al. (2017) distinguished between PAYD and pay-how-you-drive (PHYD) insurance, in which the former involves exposure-related variables (e.g., driving amount, driving time, road type), whereas the latter also encompasses behavioral measurements, such as speeding and accelerations. Ayuso et al. (2014) analyzed PHYD insurance data usage and showed that a younger driver's age, higher vehicle power, increased mileage, and a greater number of speed limit violations were correlated with a shorter duration to the driver's first accident, based on a dataset of 15,940 novice drivers. More recently, Ma et al. (2018) analyzed over 130,000 trips from 503 drivers and found that distance traveled, exceeding the local speed limit, the frequency of harsh braking events per distance traveled, and driving speed relative to other vehicles traversing the same road segment were predictive of drivers' accident rates. Naturalistic driving studies featuring researcher-instrumented vehicles offer another avenue to investigate the correlation between driving behavior and accident involvement. In the SHRP-2 project, kinematic events (i.e., hard decelerations, -accelerations, -turning) were associated with accident and near-accident occurrences (Engström et al., 2019).

In the domain of truck driving, several studies have used actuarial statistics, such as age, gender, and previous accidents or violations, to predict accident involvement or severity (e.g., Blower, 1996; Cantor et al., 2010; Guest et al., 2014). However, investigations into the relationship between truck driver behavior and accident involvement remain relatively scarce. A noteworthy exception is a study by Cai et al. (2021), which discovered that, in a substantial sample of truck drivers (n = 31,828), the number of safety-critical events per mile was associated with accident involvement. The findings were deemed robust across various business units and driver types. Hickman and Hanowski's (2012) analysis of a large naturalistic truck driving dataset demonstrated that certain cell phone tasks substantially increased the likelihood of involvement in safety-critical events. However, this case-control study provided insights only into the immediate precursors of truck accidents without establishing correlations with driving styles.

A common limitation of studies exploring predictors of adverse outcomes in driving (both car and truck driving) is the incomplete nature of the variables recorded. For instance, PAYD data may encompass information on mileage and accident rates but lack data on driver behavior (Ayuso et al., 2019; Boucher et al., 2017; Jun et al., 2007; Lemaire et al., 2016; Verbelen et al., 2018). Other studies report associations between driver behavior and accidents without considering other costly outcomes, such as traffic fines or fuel consumption (Henckaerts & Antonio, 2022; Stankevich et al., 2022). Goldenbeld et al. (2013) analyzed a large national dataset to study the relationship between traffic offenses and accident involvement. However, information on mileage and whether a specific vehicle was driven by its owner at the time of the traffic fine was unavailable. Figueredo et al. (2019) and Zhou and Zhang (2019) investigated risky truck driving behavior and used this to classify drivers into different risk profiles; nevertheless, they did not establish a direct association with accident involvement. A prevalent issue in truck driving behavior analysis is that drivers often switch trucks, as Li et al. (2019) acknowledged in their truck driver profiling study.

In the current study, a dataset encompassing truck driver behavior, driving exposure, and fuel consumption measures was merged with data on drivers' damage incidents and traffic fines. This information was used to examine the relationships between truck driving behavior and driving exposure and fuel consumption, vehicle damage incidents, and fines. The present analysis accounted for both the driver and the specific vehicle being driven. The objective was to gain a more nuanced understanding of the determinants of adverse driver outcomes, including high fuel consumption, damage incidents, and fines.

6.2. Methods

Ethics statement

In the present research, data were procured from two sources: a company specializing in collecting fleet data for driver coaching purposes and a transportation logistics company based in the Netherlands. The data were acquired following Data Transfer Agreements, and safeguards such as data minimization and anonymization were used to ensure the protection of data subjects. Data processing adhered to Article 89 of the General Data Protection Regulation; that is, data were processed for statistical purposes and presented in an aggregated format. The research proposal was reviewed by the TU Delft data privacy officer and received approval from the TU Delft Human Research Ethics Committee (approval number 1820). Owing to the retrospective and observational nature of this study, informed consent was not required.

Driving behavior data

This study uses truck driving data from a Dutch transportation company operating a fleet of 70 trucks. The drivers' primary responsibilities included shop distribution (approximately 55% of drivers), nationwide distribution (approximately 16% of drivers), and fine-meshed distribution (approximately 10% of drivers). The remaining drivers delivered to distribution centers or engaged in more flexible work arrangements.

The data were obtained through a fleet management system (FMS), which has become increasingly common in the transportation and logistics industry. FMSs enable drivers and management to monitor vehicle locations, record driving statistics, and schedule trips. The FMS data for this study were obtained from NEXTdriver (https://www.nextdriver.nl), a Dutch company that provides posttrip driving-style feedback and coaching for truck drivers. Drivers received weekly performance scores via an app and could communicate with driving-style coaches through a text messaging interface.

FMS data were collected between March 19, 2020 and March 31, 2022. Cumulative event data ("samples") of driver actions (e.g., speeding, harsh braking) were acquired at varying intervals (median sampling interval = 53.1 min). Odometer data and fuel consumption were sampled more frequently (median sampling interval = 4.47 min). Upon inspection of the odometer data, approximately 28% of the samples did not yield data, resulting in a constant value even when the vehicle was in motion. These data were linearly interpolated based on the date and time of the event recording. Through this interpolation of the odometer values, we were able to provide a more precise estimate of the total distance driven per driver, which turned out to be 2% higher with interpolation than without. Note that this interpolation had minimal impact on the driver performance measures, which are described below, as they were calculated per unit of time rather than per kilometer driven.

The data for each driver were divided into sessions, with a session being defined by automatically detecting when a new driver entered the vehicle. If drivers operated the same vehicle across multiple days, this was identified as a single session. Sessions were retained if they met specific criteria, such as physical plausibility and a nonzero driving duration (i.e., at least two sample points were required for a session to calculate the duration based on the difference in the time stamp variable), as detailed in the Supplemental Material.

Driver performance measures

A total of 12 driving measures were computed per driver from the aforementioned FMS data samples:

1. Total driving time (hours).
2. Total driving distance (km).
3. Number of days with driving data recorded (days). Any driving data were considered when counting the number of days, regardless of session demarcation.
4. Engine capacity (cc), computed as a driving-time-weighted average of the engine capacity of the truck driven. The engine capacity score provides an overall approximation of the size of all trucks used by the driver. This information was obtained by examining the license plate number in the Dutch vehicle registration database (RDW, 2022).
5. Number of vehicle switches per hour of driving (#/h). This variable may indicate the type of work being performed, that is, whether the driver was assigned to strict delivery schedules (such as shop distribution), which involves switching trucks.
6. Percentage of night shift driving (%), calculated based on the number of event samples recorded after 20:00 or before 04:00. This measure was computed for each session and subsequently averaged across all sessions for a given driver. The interval between 20:00 and 04:00 was chosen as it appeared to correspond with the working hours of night shift drivers based on inspection of the distribution of working hours in the data.
7. Mean speed (km/h), calculated from the first two measures.
8. Number of harsh brakes per hour of driving (#/h). The manufacturer of the event data recorder had established a threshold for what constitutes a harsh brake. It was defined as any vehicle deceleration that exceeds 1.5 m/s^2, a measurement that applies regardless of the driving speed at the time.
9. Speeding duration per hour of driving (s/h), with speeding defined as driving at a speed greater than 84 km/h. The threshold of 84 km/h for recording the duration of speeding was established by the manufacturer of the event data recorder, taking into account the typical 80 km/h speed limit on highways, plus an additional 4 km/h

margin. Although a variable threshold that considers the local speed limit might be more desirable, our research was constrained to this fixed parameter.

10. Duration of engine torque exceedance per hour of driving (s/h). The torque threshold was set at 90% of a factory-established engine-specific reference torque value.
11. Duration of cruise control active per hour of driving (s/h).
12. Fuel consumption per kilometer of driving (L/km).

Note that Measures 5 and 7 to 11 were computed per hour of driving (Measure 1), whereas Measure 12 was computed per kilometer of driving (Measure 2).

Damage incidents and fines

The transportation company maintained a detailed record of all damage incidents and traffic fines, including those not claimed from the insurance provider. Solely relying on insurance data could potentially result in underreporting, an issue that has been identified within the accident analysis literature (Wijnen & Stipdonk, 2016). Dorn et al. (2010) noted that underreporting is less concerning for transportation companies that maintain their own records.

Only damage incidents and fines that occurred within the study period, from March 19, 2020 to March 31, 2022, were considered. Our collaboration with the transportation company, granting us access to extensive damage data, influenced our decision to retain a broad spectrum of damage incidents in our study, not just collisions. More specifically, a total of 420 damage incidents were documented, which the transportation company had informally categorized. Specifically, 97 damage incidents were identified as "reversing incidents," 108 as "maneuvering incidents," and 49 as damage incidents resulting from the truck "swinging out." The damage incidents primarily involved bumpers, lamps, fences, poles, doors, or other (parked) vehicles. Furthermore, four damage incidents were classified as rear-end collisions, three as injuries (not related to driving but occurring during loading and unloading), eight as resulting from slipperiness, four as theft of goods, four as lane changes, one as a right-of-way incident, and one as an avoidance maneuver. Lastly, 131 damage incidents were categorized as "other" and encompassed a diverse range of types, including damage incidents that occurred during loading/unloading and damages to cargo. The remaining 10 damage incidents were not classified.

In relation to damage incidents and fines, the following scores were computed for each driver:

13. Total number of damage incidents (total in the dataset: 420). Even though many damage incidents were not directly associated with driving, this category was retained in order to obtain an estimate of all costs incurred.
14. Number of damage incidents claimed from the insurance company (total in the dataset: 200).
15. Number of damage incidents handled by the transportation company itself (i.e., not claimed from the insurance company, total in the dataset: 220).

16. Number of damage incidents for which the costs were recovered from another road user (total in the dataset: 32). This category offers a formal classification of damage incidents for which the truck driver was not at fault.
17. Number of damage incidents for which the costs were not recovered from another party, typically because the truck driver was at fault (total in the dataset: 388).
18. Number of damage incidents related to driving (total in the dataset: 303). This category was obtained by using all damage incidents (Measure 13), excluding incidents that were recovered from another party (Measure 16) and incidents that were manually labeled by the authors. Specifically, not included were damage incidents that did not occur while the truck was either driving or parking (such as damage during loading and unloading), incidents caused by others, incidents for which, according to the transportation company's records, no culpable party was identified, and incidents not evidently caused by human error (such as a flat tire). The first two authors independently classified all accidents using the above definitions and resolved disagreements through mutual discussion.
19. Number of traffic fines (total in the dataset: 266). The reasons for most of the fines were unavailable, but out of the 99 fines with descriptions, 77 were for speeding

Statistical analysis

Means, standard deviations, and intercorrelations of the driver measures and damage variables were calculated. Additionally, graphs were constructed to clarify the effects of truck size and time of day.

One issue is that the correlation coefficients are influenced by driving exposure and thus require statistical correction. The total driving time and distance (Measures 1 and 2) do not precisely represent the actual driving amount of the drivers, as sessions shorter than the minimum requirement of two sample points were excluded. Since the number of distinct days that drivers appeared in the FMS dataset appeared to be the most accurate index of exposure, and because transportation companies are likely interested in damages incurred per employee-day (rather than per hour or km driving), the number of days with data (Measure 3) was used as an index of exposure in the regression analyses.

Linear regression analyses were performed with exposure (Measure 3), session characteristics (Measures 4, 5, 7, & 8), and one of the behavioral measures (Measure 9 or 11) as predictor variables. The selected dependent variables included the total number of damages (Measure 13), the number of fines per driver (Measure 19), and the fuel consumption per kilometer of driving (Measure 12).

6.3. Results

A total of 27,543 sessions involving 180 drivers were used in the study. The average number of sessions per driver was 153.0 (*SD* = 141.7), and the mean number of vehicles driven per driver was 15.37 (*SD* = 10.92).

Table 6.1 presents the means, standard deviations, and higher moments of the measures

Table 6.1: Descriptive statistics of the driving measures, damage incidents, and fines.

No.	Measure	Mean	*SD*	Min	Max	skewness	kurtosis	unit
1	Total driving time	963.0	741.0	1.9	3129.7	0.39	2.25	h
2	Total driving distance	55829.7	45398.3	136.6	199675.6	0.66	2.81	km
3	Number of days with data	254.0	175.7	3.0	582.0	-0.07	1.51	days
4	Engine capacity (weighted by driving time)	9.8	1.7	3.0	12.8	-1.83	6.34	L
5	Number of vehicle switches per hour of driving	0.14	0.13	0.00	0.73	1.22	5.63	#/h
6	Percentage night shift driving	12.17	16.38	0.00	100.00	2.57	9.98	%
7	Mean speed	53.80	9.10	28.25	76.95	0.04	3.24	km/h
8	Number of harsh brakes per hour of driving	1.07	1.45	0.00	11.62	4.56	29.98	#/h
9	Speeding duration (>84 km/h) per hour of driving	638.3	540.5	0.0	2369.8	1.05	3.74	s/h
10	Duration of excessive engine torque per hour of driving	124.5	65.3	1.6	392.0	0.60	3.97	s/h
11	Cruise control duration per hour of driving	1277.6	632.1	0.0	3099.1	0.32	2.90	s/h
12	Fuel consumption per kilometer of driving	0.28	0.05	0.12	0.38	-0.74	3.81	L/km
13	Total number of damage incidents	2.33	2.85	0	18	2.06	8.94	#
14	Number of damage incidents, claimed from insurance	1.11	1.52	0	9	1.95	7.98	#
15	Number of damage incidents, not claimed from insurance	1.22	1.69	0	9	1.83	6.65	#
16	Number of damage incidents, costs recovered from other party	0.18	0.50	0	4	3.88	23.53	#
17	Number of damage incidents, costs not recovered from other party	2.16	2.74	0	18	2.19	9.94	#
18	Number of damage incidents, driving-related	1.68	2.30	0	12	1.97	7.33	#
19	Number of fines	1.48	3.25	0	30	5.01	37.47	#

Note: *SD* = standard deviation; Min = minimum; Max = maximum; n = 180, except for Measure 12, for which n = 158.

used in the analysis. As can be observed, approximately 1,000 hours of driving were available per driver on average (Measure 1). Nevertheless, there were considerable individual differences in exposure, as evidenced by the standard deviations of Measures 1–3. Drivers changed trucks approximately once every seven hours of driving (Measure 5). The majority of drivers switched trucks at least once, with only six drivers never making a switch. Harsh braking events were relatively infrequent, occurring about once per hour (Measure 8). Drivers exceeded the established speed limit (84 km/h) 18% of the time (Measure 9), experienced high engine torque approximately 3% of the time (Measure 10), and used cruise control 35% of the time (Measure 11). The average fuel consumption was 0.28 liters per kilometer (Measure 12). On average, drivers incurred 2.33 damages and received 1.48 fines.

Table 6.2 presents the Pearson correlation matrix among the various measures, from which several patterns can be discerned. Firstly, the number of damage incidents and the number of fines per driver can be predicted based on the number of harsh braking events per hour (highlighted in green). Interestingly, the number of damage incidents recovered from another party (and thus not attributable to the driver) showed a near-zero correlation with harsh braking. It is also worth noting that other behavioral measures did not strongly predict damage incidents and fines. For instance, contrary to expectations, speeding was not a significant predictor of fines, even though the majority of fines were issued for speeding.

Furthermore, the correlations with fuel consumption (Table 6.2; highlighted in yellow) demonstrated that fuel consumption per kilometer could be predicted based on the engine torque exceedance per hour. It was also evident that, apart from driver behavior, both session and vehicle characteristics contributed to the prediction of fuel consumption. In particular, drivers operating trucks with larger engine capacities exhibited higher fuel consumption, maintained greater mean speeds, and changed vehicles more frequently.

The characteristics of driving sessions were found to be correlated with various driver performance measures. Specifically, drivers operating trucks with larger engines (as shown in Table 6.2, highlighted in blue) tended to switch vehicles more frequently, drive with higher engine torque, experience fewer harsh braking events, and exhibit higher mean speed and cruise control usage. Furthermore, drivers who were on the road more frequently during night shift generally drove at faster speeds and were more prone to speeding (as demonstrated by the correlation coefficients highlighted in gray). Lastly, the total driving time, distance covered, and the number of days with available data were found to be predictive of the number of damage incidents and fines (Table 6.2, highlighted in orange). These findings suggest that the impact of exposure should be taken into consideration in subsequent analyses.

Table 6.2: Pearson correlation coefficients among the driving measures, damage incidents, and fines.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	Total driving time																		
2	Total driving distance	0.97																	
3	Number of days with data	0.89	0.84																
4	Engine capacity (weighted by driving time)	-0.04	0.05	-0.07															
5	Number of vehicle switches per hour of driving	-0.47	-0.44	-0.34	0.31														
6	Percentage night shift driving	0.11	0.21	0.08	0.26	-0.13													
7	Mean speed	0.15	0.31	-0.03	0.22	-0.19	0.55												
8	Number of harsh brakes per hour of driving	0.05	-0.02	0.09	-0.40	-0.01	-0.19	-0.23											
9	Speeding duration per hour of driving	0.11	0.23	0.04	0.04	-0.14	0.54	0.61	-0.03										
10	Duration of excessive engine torque per hour of driving	0.03	0.04	0.05	0.33	0.11	-0.10	-0.03	0.17	0.02									
11	Cruise control duration per hour of driving	0.23	0.36	0.09	0.30	-0.18	0.50	0.79	-0.29	0.33	-0.08								
12	Fuel consumption per kilometer of driving	-0.19	-0.16	-0.03	0.71	0.45	-0.03	-0.22	-0.12	-0.06	0.42	-0.11							
13	Total number of damage incidents	0.42	0.38	0.45	-0.15	-0.06	-0.08	-0.05	0.27	0.04	-0.07	0.00	-0.06						
14	Number of damage incidents, claimed from insurance	0.36	0.31	0.37	-0.17	-0.11	-0.17	-0.08	0.28	0.01	-0.04	-0.04	-0.11	0.88					
15	Number of damage incidents, not claimed from insurance	0.38	0.36	0.42	-0.10	0.00	0.01	-0.02	0.21	0.05	-0.09	0.04	0.00	0.90	0.58				
16	Number of damage incidents, costs recovered from other party	0.32	0.35	0.29	0.12	-0.07	0.00	0.08	0.00	0.14	0.05	0.14	0.09	0.32	0.34	0.23			
17	Number of damage incidents, costs not recovered from other party	0.38	0.33	0.41	-0.18	-0.05	-0.08	-0.07	0.28	0.01	-0.09	-0.02	-0.08	0.99	0.85	0.90	0.15		
18	Number of damage incidents, driving-related	0.33	0.30	0.36	-0.16	-0.04	-0.10	-0.05	0.23	-0.01	-0.11	-0.01	-0.10	0.94	0.86	0.82	0.14	0.95	
19	Number of fines	0.26	0.23	0.26	-0.24	-0.14	-0.04	-0.09	0.42	0.06	-0.01	-0.10	-0.14	0.56	0.52	0.48	0.07	0.57	0.49

Note: $n = 180$, except for Measure 12, for which $n = 158$.
Green: Correlations between harsh brakes and damage incidents/fines. Yellow: Correlations with fuel consumption.
Blue: Correlations with engine capacity. Gray: Correlations with night shift driving. Orange: Correlations between exposure and damage incidents/fines.
The diagonal and upper triangle of the correlation matrix have been omitted for clarity, as they represent self-correlations and duplicate information.
$p < 0.05$ for $r \geq 0.15$; $p < 0.01$ for $r \geq 0.20$.

Figures 6.1 and 6.2 depict the variations in driving behavior influenced by the time of day and the type of truck, respectively. Figure 6.1 (top left) demonstrates the differences in exposure (total driving time) throughout the day, with the majority of driving taking place between 05:00 and 17:00. Instances of speeding (Figure 6.1, bottom right) and cruise control activation (Figure 6.1, top right) were more common during night shift compared to day shift, while harsh braking events occurred more frequently during the day (Figure 6.1, bottom left). A noticeable shift in activity can be observed around 22:00, which may be attributed to the start of the night shift for drivers operating during night shift.

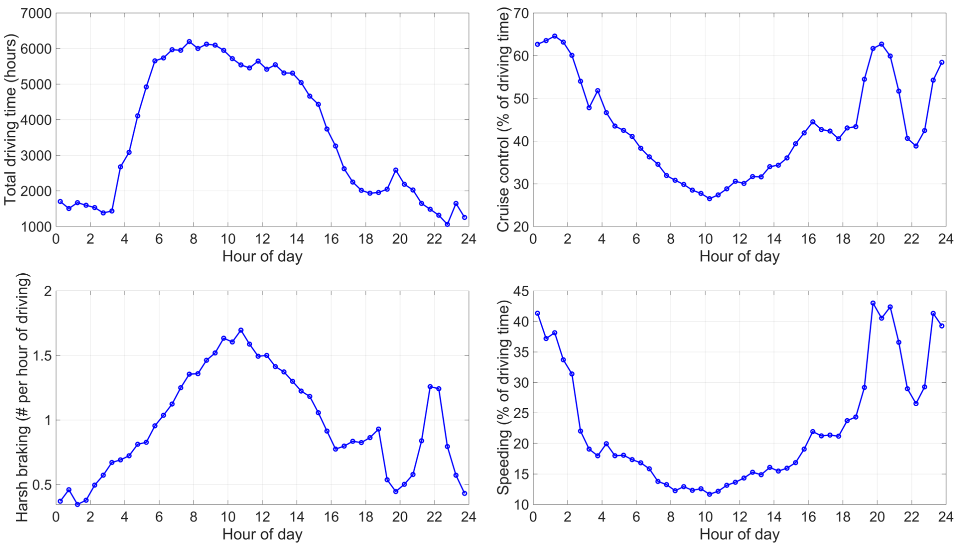


Figure 6.1: Total driving time (hours) (top left), percentage of driving time with cruise control activated (top right), number of harsh braking events per hour of driving (bottom left), and percentage of driving time where the driver was speeding (bottom right), calculated from all recorded data samples (n = 533,577).

Figure 6.2 presents an analysis of the trucks within the fleet. Generally, trucks with smaller engine capacities exhibited a higher frequency of harsh braking events per hour. This relationship (r = -0.57) is consistent with the correlations observed at the driver level (r = -0.40, indicated in blue; Table 6.2), suggesting that in addition to the driver, the vehicle being operated should be considered in driver evaluations. Furthermore, it was evident that trucks with larger engine capacities consume more fuel per kilometer (r = 0.78), a trend also observed at the driver level (r = 0.71).

Linear regression analyses were performed to predict the total number of damage incidents, the number of fines, and fuel consumption per kilometer of driving (Table 6.3). As observed in Table 6.3, exposure (i.e., the number of days with data) was a strong predictor of damage incidents. The harsh braking rate also served as a significant predictor, whereas the other variables did not exhibit significant predictions. Similarly, exposure and harsh braking were found to be predictive of fines. Lastly, fuel consumption was ef-

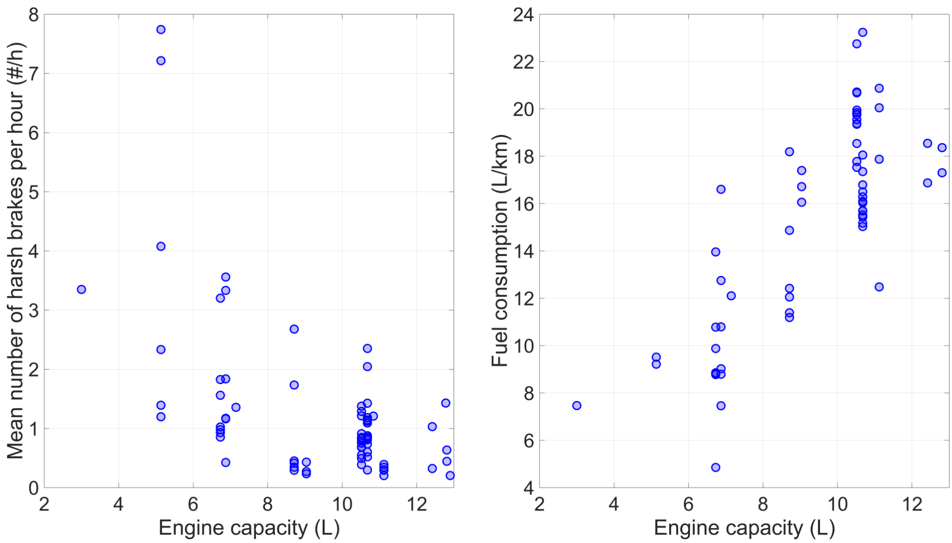


Figure 6.2: Left: Mean number of harsh braking events per hour of driving versus engine capacity; Right: Fuel consumption per kilometer of driving. Each marker represents a truck (left: n = 70, right: n = 62; fuel consumption was not recorded in 8 trucks).

fectively predicted (overall r = 0.84) based on mean speed, the occurrence of high engine torque, engine capacity, and the number of vehicle switches per hour of driving.

A limitation of the correlation coefficients shown in Table 6.2, as well as the regression analyses displayed in Table 6.3, is that several of the variables have a tailed distribution, as indicated by the skewness and kurtosis values in Table 6.1. With respect to the number of traffic fines, for example, the average among the 180 drivers was 1.48, but there was one driver who had as many as 30 fines (see Table 6.1).

These outliers do not necessarily invalidate our analyses, but they do imply that the analyses are less robust compared with a hypothetical situation in which the data would follow a nontailed distribution. To investigate the robustness of our findings, we repeated all analyses after applying a rank transformation. This transformation ordered all 180 values (or 158 values in the case of fuel consumption), corresponding to the individual drivers, from 1 (lowest value) to 180 (highest value), taking the average rank in the case of ties. This procedure, in the context of correlation coefficients, is equivalent to the use of the Spearman rank-order correlation coefficient (De Winter et al., 2015).

Table 6.4 displays the same regression analyses as in Table 6.3, but after having performed a rank transformation on all predictor- and criterion variables. It can be seen that the statistically significant zero-order correlations for the behavioral variables (mean speed, harsh braking) were still statistically significant after performing the rank transformation, although they were somewhat weaker (compare Table 6.4 with Table 6.3). As for the regression coefficients (β_r), it was noticeable that harsh braking was still significantly

Table 6.3: Results of three linear regression analyses for predicting (1) total number of damage incidents (n = 180, overall predictive r = 0.52), (2) number of fines (n = 180, overall predictive r = 0.48), and (3) fuel consumption per kilometer of driving (n = 158, overall predictive r = 0.84)

	No.	Measure	Unit	B	r_0	p_0	β	p
Total number of damage incidents		Constant	—	-1.0052	—	—	0	—
	3.	Number of days with data	days	0.0077	0.45	<0.001	0.48	<0.001
	4.	Engine capacity	L	-0.1041	-0.15	0.048	-0.06	0.417
	5.	Number of vehicle switches per hour of driving	#/h	2.8004	-0.06	0.431	0.13	0.094
	6.	Percentage of night shift driving	%	-0.0185	-0.08	0.277	-0.11	0.185
	7.	Mean speed	km/h	0.0329	-0.05	0.475	0.10	0.196
	8.	Number of harsh brakes per hour of driving	#/hour	0.4100	0.27	<0.001	0.21	0.004
Number of fines		Constant	—	0.9959	—	—	0	—
	3.	Number of days with data	days	0.0038	0.26	<0.001	0.21	0.005
	4.	Engine capacity	L	-0.1247	-0.24	0.001	-0.07	0.406
	5.	Number of vehicle switches per hour of driving	#/h	-1.0371	-0.14	0.066	-0.04	0.595
	6.	Percentage of night shift driving	%	0.0048	-0.04	0.552	0.02	0.769
	7.	Mean speed	km/h	-0.0014	-0.09	0.227	0.09	0.962
	8.	Number of harsh brakes per hour of driving	#/hour	0.8427	0.42	<0.001	0.38	<0.001
Fuel consumption per kilometer of driving		Constant	—	0.181409	—	—	0	—
	3.	Number of days with data	days	0.000005	-0.03	0.711	-0.02	0.716
	4.	Engine capacity	L	0.018792	0.71	<0.001	0.67	<0.001
	5.	Number of vehicle switches per hour of driving	#/h	0.068599	0.45	<0.001	0.17	<0.001
	6.	Percentage of night shift driving	%	0.000128	-0.03	0.734	0.04	0.442
	7.	Mean speed	km/h	-0.002058	-0.22	0.006	-0.38	<0.001
	10.	Duration of excessive engine torque per hour of driving	s/hour	0.000138	0.42	<0.001	0.18	<0.001

Note. B: unstandardized regression coefficient, r_0: zero-order correlation coefficient between measure and criterion variable (corresponding to Table 6.2), p_0: p-value for testing the hypothesis of a correlation of 0, β: standardized regression coefficient p: p-value for testing the hypothesis of a regression coefficient of 0.

Table 6.4: Results of three linear regression analyses for predicting (1) total number of damage incidents (n = 180, overall predictive r = 0.55), (2) number of fines (n = 180, overall predictive r = 0.45), and (3) fuel consumption per kilometer of driving (n = 158, overall predictive r = 0.73), after applying a rank-transformation of the individual predictor variables and criterion variable.

	No.	Measure	$r_{0,r}$	$p_{0,r}$	β_r	p_r
Total number of damage incidents	3.	Number of days with data	0.50	<0.001	0.55	<0.001
	4.	Engine capacity	-0.09	0.208	-0.07	0.333
	5.	Number of vehicle switches per hour of driving	-0.04	0.571	0.19	0.009
	6.	Percentage of night shift driving	-0.05	0.492	-0.11	0.173
	7.	Mean speed	-0.01	0.894	0.10	0.179
	8.	Number of harsh brakes per hour of driving	0.22	0.003	0.10	0.180
Number of fines	3.	Number of days with data	0.35	<0.001	0.32	<0.001
	4.	Engine capacity	-0.14	0.054	-0.07	0.366
	5.	Number of vehicle switches per hour of driving	-0.08	0.288	0.07	0.343
	6.	Percentage of night shift driving	-0.10	0.197	-0.07	0.434
	7.	Mean speed	-0.04	0.630	0.10	0.221
	8.	Number of harsh brakes per hour of driving	0.32	<0.001	0.25	0.002
Fuel consumption per kilometer of driving	3.	Number of days with data	-0.02	0.776	0.03	0.664
	4.	Engine capacity	0.52	<0.001	0.41	<0.001
	5.	Number of vehicle switches per hour of driving	0.50	<0.001	0.35	<0.001
	6.	Percentage of night shift driving	0.09	0.242	0.00	0.975
	7.	Mean speed	-0.25	0.002	-0.25	<0.001
	10.	Duration of excessive engine torque per hour of driving	0.38	<0.001	0.20	0.001

Note. $r_{0,r}$: zero-order correlation coefficient between measure and criterion variable (corresponding to Table 6.5 in the Supplementary Material), $p_{0,r}$: p-value for testing the hypothesis of a correlation of 0, β_r: standardized regression coefficient, p_r: p-value for testing the hypothesis of a regression coefficient of 0. The subscript r refers to the fact that the relevant variables were rank-transformed.

predictive of fines, and that mean speed and excessive engine torque were also significantly predictive of fuel consumption. However, the number of harsh braking events was no longer statistically significantly predictive of the total number of damage incidents, with a regression coefficient of 0.10 compared with 0.21 without rank transformation (Table 6.3).

6.4. Discussion

Damage incidents and fines

This study investigated the associations between driver behavior measures and the extent of damage incidents, fines, and fuel consumption at a Dutch transportation company over a two-year period. A key finding was that the frequency of harsh braking events per hour of driving served as a predictor for damage incidents and fines. This relationship was initially identified through zero-order correlations and persisted even after accounting for factors such as the number of driving days, engine capacity, vehicle switches, the proportion of night shift driving, and average speed. That driver behavior remained a predictor of damage incidents and fines in the regression analysis, beyond just the zero-order correlations, indicates that drivers themselves, rather than solely the road environment or truck, serve as an explanatory factor.

It must, however, be noted that a small number of drivers with an excessive number of fines and damage incidents, and corresponding frequent harsh braking, are likely to have made a substantial contribution to the observed correlation coefficients. When we compressed the scores by ranking all drivers, the correlations and regression coefficients became weaker, with the regression coefficient between harsh braking and damage incidents not even being statistically significant anymore. The rank transformation that we performed probably yielded statistically more robust results, but it has the drawback that the effects of the highly deviant drivers received less weight in the analysis.

The finding that severe braking events, or harsh longitudinal or lateral accelerations in general, can predict damage incidents is in line with findings from other studies (e.g., for bus drivers (Khorram et al., 2020); for truck drivers (Cai et al., 2021); for passenger car drivers (Enström et al., 2019)). On a psychobehavioral level, harsh braking may contribute to accidents through various mechanisms. One possibility is that sudden braking indicates a causative relationship with an accident. For example, a driver might need to apply the brakes forcefully when confronted with a stationary object or road user ahead. Harsh braking could additionally signify inadequate hazard perception skills, as proficient drivers can detect hazards earlier and consequently apply the brakes more gently (Botzer et al., 2019; Simons-Morton et al., 2009). Furthermore, harsh braking might result from distracted driving, occurring when the driver refocuses their attention on the road (Hancock et al., 2003; Harbluk et al., 2007).

At the same time, our analysis demonstrated that the rate of harsh braking events was influenced by the type of work or truck assigned to the driver. For example, we discovered that a larger engine capacity correlated with fewer harsh brakes. Although heavier trucks can achieve a steady-state deceleration in an emergency stop nearly as high as that of lighter trucks or passenger cars, it takes longer to build up this deceleration (Reed & Keskin, 1987). Moreover, heavier trucks possess greater inertia, and are probably driven more cautiously to prevent damage to the payload and to avoid jackknifing or a trailer swing. Another possible explanation for the association between engine capacity and harsh braking is that the drivers themselves are the cause, with less experienced or poorly skilled drivers being more likely to be assigned to smaller trucks, as also reflected in the different driving licenses that exist (e.g., C versus CE in Europe). Along the same lines, it has been proposed that one of the reasons long-combination vehicles are safer is that they are operated by better-trained drivers (Islam & Hernandez, 2016, Lemp et al., 2011). In summary, our results suggest that truck damage incidents are caused by both the driver and the environment in which the driver operates. These findings are in line with Reason's (1995) model of organizational accidents, which posits that adverse events (damage incidents) are the fault of the workers themselves as well as the "error and violation producing" conditions in which they must work.

Damage incurred in which the driver was not at fault (Measure 16) demonstrated an almost negligible correlation with harsh braking, consistent with Af Wåhlberg's criterion for at-fault accidents: *"Non-culpable accidents are not possible to predict with any variable when exposure has been controlled for"* (Af Wåhlberg, 2017, Chapter 4). We additionally computed a correlation coefficient between the number of not at-fault damage incidents (Measure 16) and the number of at-fault damage incidents (Measure 17), with the number of driving days (Measure 3) accounted for. The partial correlation was 0.04, indicating that at-fault and not-at-fault damage incidents were uncorrelated once exposure was adjusted for. This near-zero correlation can also be interpreted as a form of discriminant validity of the harsh braking measure, that is, harsh braking does not correlate with outcomes it should not be associated with. However, it is important to note that the mean (0.18) and standard deviation (0.50) of the number of at-fault damage incidents per driver were small, which indicates that correlations were attenuated (Af Wåhlberg, 2017; for a computer simulation, see De Winter et al., 2015).

One puzzling finding from our study is that the majority of truck damage incidents resulted from collisions during reversing, maneuvering, or swinging out, typically causing minor damages to bumpers, lights, or mirrors. A mere 1% of damage incidents involved rear-end collisions, where harsh braking might be expected as a causal antecedent, whereas numerous damage incidents were entirely unrelated to driving. This pattern of collisions warrants an inquiry into the causal relationship between harsh braking and damage incidents. It is conceivable that a common cause exists, potentially rooted in drivers' personality traits, age, or experience, or in traffic conditions (e.g., harsh braking is more likely for drivers who frequently navigate hazardous environments such as cities). The personality psychology literature suggests that accident involvement correlates with low conscientiousness (Arthur & Graziano, 1996). This aligns with anecdotal evidence gathered from our conversations with managers at an insurance and transportation company, who indicated that accident-free drivers can be identified based on various nondriving behaviors and attitudes. Examples of such behaviors and attitudes include inspecting the truck for technical defects, ensuring safety before reversing, maintaining a clean cabin, and demonstrating a strong work ethic. These comments correspond with a study among Colombian truck drivers, which discovered that self-reported positive behaviors not directly related to driving (e.g., *"I use my safety gear (hard hat, boots, and gloves) according to the safety requirements"*) exhibited a negative association with harsh braking events (Valenzuela & Burke, 2020).

Fuel Consumption

With respect to fuel consumption, our findings indicate that high engine torque (i.e., fully depressing the throttle) serves as a predictive factor, suggesting an influence of driving style. However, this behavior may also be determined by the truck's interaction with the road environment. To illustrate, the correlation matrix revealed that trucks with larger engine capacities were more frequently driven at full throttle. This can be attributed to such trucks typically possessing greater mass and payload, which may—despite their increased engine power—necessitate full-throttle depression under certain conditions. To exemplify, focus group and interview research involving truck drivers demonstrated that during highway-merging tasks, acceleration is a critical subtask, as the driver must attain sufficient speed to merge safely (Dreger et al., 2020, Kondyli & Elefteriadou, 2009). Consequently, depending on the situation, high engine torques can be considered a requisite rather than aberrant driver behavior. Several other studies have reached conclusions similar to ours (Mane et al., 2021, Walnum & Simonsen, 2015). For example, Walnum and Simonsen (2015) examined the fuel consumption of heavy-duty trucks in Norway using FMS data. They determined that factors such as engine torque exceedance, running idle, driving in high gear, horsepower, truck type, and trip characteristics significantly influenced fuel consumption. Furthermore, they argued that road conditions (e.g., mountainous or not) and vehicle properties exert a more substantial impact than driver-behavior variables. The present study builds on previous research findings by demonstrating that these results are replicable in a fleet of trucks in the Netherlands. We additionally have shown that fuel consumption can be accurately predicted using a small set of variables.

Limitations

The present study offers insights into the factors influencing the expenses (fines, damage incidents, fuel consumption) associated with truck driving. However, as frequently observed in on-road driving studies, the results are subject to certain limitations. One such limitation is the data collection period, which spanned from 2020 to 2022, coinciding with the enforcement of COVID-19-related lockdowns. These restrictions have been documented to cause an increase in speeding, presumably a result of the decrease in traffic density (Katrakazas et al., 2020; Yasin et al., 2021). In addition to COVID-19 having an impact on the level of traffic on the roads, thereby influencing speed and possibly hard braking, it also had an effect on the truck companies themselves (Dablanc et al., 2022; Elbert et al., 2023; Sperry et al., 2022). The pandemic caused a disruption in supply chains, forcing companies to modify their business models in some cases, and in others to transport different types of cargo. There was also variability in volume, with some companies experiencing an increase and others a decrease in their clientele. Truckers also had to contend with possible closures of rest areas or limited access to restaurants, which may have had an indirect effect on their shifts (Sperry et al., 2022; Allen & Piecyk, 2023).

These factors may limit the generalizability of the current findings. We anticipate that COVID-19 could have had some influence on the absolute values, but we expect that the relative relationships, that is, the correlation coefficients obtained, are robust to effects of COVID-19.

Another limitation is that drivers were able to access summary scores and personalized text-based feedback through a mobile app. Moreover, the transportation company maintained records of driver accidents and employed a company coach (a certified driving instructor) to improve the safety-related behaviors of its employees. Moreover, traffic fines, such as those for speeding, were deducted from employees' salaries. These complex feedback mechanisms are likely to have influenced driving behavior. In fact, it is conceivable that simply being aware of being monitored can lead to improved behavior and adherence to rules (Wouters & Bos, 2000), a phenomenon that extends beyond the realm of driving (e.g., Kohli et al., 2009, which examined hand hygiene compliance in hospital settings).

A technical limitation of this study is the relatively low frequency of driving event data sampling, which occurred approximately once per hour. To compute difference scores of accumulated events, at least two sample points per session were required, rendering short sessions with only one sample point unsuitable for the calculation of driver behavior scores. For future research, it is recommended to obtain data at a higher measurement frequency. Furthermore, we recognize that our analysis, based on FMS data, captured merely a small aspect of driving. Information on looking behavior, lateral maneuvers, advanced driver-assistance system warning events or activations, GPS data, and local speed limits, as well as measurements of the truck's momentary mass and payload, were not available.

Further investigation may be required to ascertain the ideal thresholds for predicting accidents. In our study, the threshold for harsh braking was set at a relatively low value of 1.5 m/s^2. Increasing this threshold would result in a reduction of braking events, placing greater emphasis on deceleration events closely associated with actual accidents. For instance, Cai et al. discovered that the number of forward-collision mitigation system activations was a more robust predictor of accidents than the number of harsh braking events, potentially owing to the former being a more likely precursor to actual accidents (Cai et al., 2021). Similarly, Perez et al. recommended a high deceleration threshold of 7.5 m/s^2 for identifying accidents (which were labeled by trained coding staff) within a naturalistic car-driving dataset (Perez et al., 2017). In a study evaluating various deceleration thresholds among bus drivers in Iran (Khorram et al., 2020), it was found that for deceleration thresholds of 2, 3, 4, 5, and 6 m/s^2, the mean number of threshold exceedance events was 0.909, 0.172, 0.041, 0.011, and 0.002 per kilometer driven, respectively. The corresponding correlation coefficients with the number of crashes involving these bus drivers appeared to decrease with increasing thresholds: 0.208, 0.244, 0.209, 0.163, and 0.081, respectively. Furthermore, Khorram et al. (2020) demonstrated that the correlation with the mean absolute variation of speed (referred to as "celeration"; see also Af Wåhlberg (2006)) was stronger when the deceleration threshold was lower. In summary, the literature suggests that a lower threshold implies that the event count becomes more equivalent to the dynamics of the driver's driving style. Exceedances of high thresholds will occur infrequently, and thereby exhibit relatively little statistical power, and when they do occur, they may be a precursor to an accident or near-accident, rather than a definitive indicator of the general driving style of the driver. In our study, the threshold was 1.5 m/s^2, which is relatively low, and thus an indicator of driving style rather than emergency stops. Despite our low threshold, the number of exceedances was quite low, on average, once per hour—which can possibly be attributed to trucks frequently traveling on the highway where hard braking is rare.

A statistical limitation of this study is the high kurtosis and skewness exhibited by the data on damage incidents, fines, and certain driver behavior measures (see Table 6.1). This implies that the correlation and regression coefficients obtained are primarily attributable to a small number of drivers with extreme scores. One possible explanation for the outliers in the number of fines could be the presence of an undetected speed camera on a section a driver traversed for several consecutive days. Although the sample size was adequate for obtaining statistically significant effects, replicating this study with a larger number of drivers is recommended.

Another limitation of our research is the lack of data related to the characteristics of the drivers, including, but not limited to, age and years of driving experience. The literature demonstrates that in the context of passenger vehicles, young and inexperienced drivers display a greater propensity for engaging in risky behavior on the road, and they are disproportionately involved in accidents (De Winter et al., 2015; Organization for Economic Co-operation and Development, 2006). In the case of truck drivers, the manifestation of risky behavior may be less pronounced, owing to their anticipated adherence to professional norms. Nonetheless, existing studies still reveal that young drivers of heavy goods vehicles are overrepresented in accidents (Duke et al., 2010). As indicated above, it is plausible that age, inexperience, or both, are underlying causes of harsh braking and accidents. Future research should document the personal characteristics of truck drivers to gain a broader understanding of the factors influencing the costs of truck driving.

6.5. Conclusions

The current study offers insights into the factors related to the expenses incurred in truck transportation. Our investigation, using a combination of datasets, revealed a connection between harsh braking and fines as well as damage incidents. Furthermore, the analysis suggested that a behavioral aspect underlies this association, implying that harsh braking incidents may be indicative of the driver's unfavorable skills and attitudes (e.g., inadequate foresight and planning). Simultaneously, we demonstrated that harsh braking is not exclusively attributable to the driver. Factors such as route type (as represented by the variable "night shift driving"), the average speed of the session, and the truck itself were found to influence driving scores, damage incidents, and fines. In relation to fuel consumption, it appears that the truck, rather than the driver, serves as the primary determinant.

Recommendations and outlook

This study investigated the predictors of damage incidents and other costs associated with truck driving, with the aim of identifying potential avenues for cost reduction. Based on the findings that harsh braking events were predictive of damage incidents and fines, organizations could develop training and coaching programs to help drivers improve their driving behavior. This could include providing feedback on harsh braking events (Mase et al., 2020) and offering training on hazard perception (Park et al., 2018) and defensive driving techniques (Huang & Ford, 2012). Training and coaching could be delivered through simulators (Hirsch et al., 2017; Galal et al., 2023; Ribeiro et al., 2021), in-person sessions (Bell et al., 2017, Soccolich & Hickman, 2014), online modules (Scheffel et al., 2013), or mobile apps (Brijs et al., 2020). A review by Michelaraki et al. (2021) concluded that implementing gamification and reward schemes appears to effectively improve safety across various modes of transportation, including truck driving. Previous research has reported positive effects of the combination of monitoring and coaching on safe driving behavior (Mase et al., 2020; Bell et al., 2017). Advancements in driving safety may be closely linked to improvements in eco-driving. This is because safety indicators, such as reduced instances of harsh braking or maintaining lower speeds, are considered effective strategies for conserving fuel.

However, it must simultaneously be recognized that harsh braking does not necessarily have to be the direct cause of damage incidents or fines; it may be an epiphenomenon of other underlying issues. Indeed, although our work demonstrated predictive correlations between driving behavior and costs, we indicated that these correlations could emanate from multiple causes. These encompass aspects of the truck and its payload, the conditions under which a driver might operate, and the personality of the driver—concerns that go beyond mere harsh braking. Such an understanding compels us to adopt a broader approach to cost reduction. In developing this systemic approach, it is imperative to address the challenges known to both the transportation organization and the drivers. Transportation companies face various hurdles, including labor market shortages and strong competition. Concurrently, drivers often experience minimal face-to-face interaction with their employers, receiving their assignments through an app. The task of truck driving can be strenuous, especially when it involves navigating a large vehicle through congested urban areas. Zohar et al. (2014, p. 19) highlighted a *"psychosocial disparity, due to the fact that dispatchers in trucking companies are often more educated than drivers, yet have no truck-driving experience"*. They further found that leadership and work ownership indicators correlated with the frequency of harsh braking. These points signify that, in addition to coaching and evaluating drivers, the overarching safety culture is integral to truck driving safety (see also Huang et al. (2013) and Mooren et al. (2014)). An effective safety culture highlights safe driving practices and motivates drivers to take responsibility for their actions on the road.

Our study also elucidated associations between truck characteristics and driver behavior. This information may prove useful for organizations to refine their fleet management strategies. For example, when the payload permits, it may be advantageous to allocate drivers to trucks with smaller engine capacities. Such trucks have been observed to exhibit fewer instances of harsh braking and demonstrate decreased fuel consumption. In addition, organizations could benefit from a thorough analysis of their existing routes and schedules. Specifically, we recognize the potential of software that allocates drivers to trucks according to the necessary trip duration and payload. These recommendations align with last-mile delivery logistics (for optimization models, refer to Giuffrida et al. (2022)), a topic that may necessitate additional investigation in the context of heavy goods vehicles.

Our analysis demonstrated that, although risky driver behavior can be reliably identified, it is also influenced by external factors such as the size of the truck and the time of day. Consequently, it is advisable to gather more data on these external variables. Using GPS technology could improve the validity of driver behavior scores by allowing them to be standardized in relation to other drivers on the same stretch of road (for a similar approach among car drivers, see Ma et al. (2018)). The availability of GPS data would also contribute to the identification of accident-prone areas (for related methods, see Desai et al. (2021), Kamla et al. (2019), and Stipancic et al. (2017)). In addition, incorporating other types of sensors, such as radar and cameras, might improve risk assessment associated with driver behavior.

References

Af Wåhlberg, A. (2017). Driver behaviour and accident research methodology: Unresolved problems. London, UK: CRC Press.

Af Wåhlberg, A. E. (2006). Speed choice versus celeration behavior as traffic accident predictor. Journal of Safety Research, 37(1), 43–51. https://doi.org/10.1016/j.jsr.2005.10.017

Af Wåhlberg, A., & Dorn, L. (2009). Bus driver accident record: The return of accident proneness. Theoretical Issues in Ergonomics Science, 10(1), 77–91. https://doi.org/10.1080/s2bh220801912597

Allen, J., & Piecyk, M. (2023). Meeting the welfare and parking needs of goods vehicle drivers and riders (Technical Report No. ENG-TR.036). University of Westminster, UK.

American Transportation Research Institute. (2020). An analysis of the operational costs of trucking: 2020 update. https://truckingresearch.org/wp-content/uploads/2020/11/ATRI-Operational-Costs-of-Trucking-2020.pdf

Arthur, W., Jr., & Graziano, W. G. (1996). The five-factor model, conscientiousness, and driving accident involvement. Journal of Personality, 64(3), 593–618. https://doi.org/10.1111/j.1467-6494.1996.tb00523.x

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. Transportation, 46, 735–752. https://doi.org/10.1007/s11116-018-9890-7

Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. Accident Analysis & Prevention, 73, 125–131. https://doi.org/10.1016/j.aap.2014.08.017

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. Decision Support Systems, 98, 69–79. https://doi.org/10.1016/j.dss.2017.04.009

Bell, J. L., Taylor, M. A., Chen, G.-X., Kirk, R. D., Leatherman, E. R. (2017). Evaluation of an in-vehicle monitoring system (IVMS) to reduce risky driving behaviors in commercial drivers: Comparison of in-cab warning lights and supervisory coaching with videos of driving behavior. Journal of Safety Research, 60, 125–136. https://doi.org/10.1016/j.jsr.2016.12.008

Bian, Y., Yang, C., Zhao, J. L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. Transportation Research Part A: Policy and Practice, 107, 20–34. https://doi.org/10.1016/j.tra.2017.10.018

Blower, D. (1996). The accident experience of younger truck drivers (Report No. GLCTTR 81-96/1). University of Michigan, Ann Arbor, MI.

Botzer, A., Musicant, O., & Mama, Y. (2019). Relationship between hazard-perception-test scores and proportion of hard-braking events during on-road driving – An investigation using a range of thresholds for hard-braking. Accident Analysis & Prevention, 132, 105267. https://doi.org/10.1016/j.aap.2019.105267

Boucher, J.-P., Côté, S., & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. Risks, 5(4), 54. https://doi.org/10.3390/risks5040054

Brijs, K., Brijs, T., Ross, V., Donders, E., Vanrompay, Y., Wets, G., Dirix, H., Katrakazas, C., Yannis, G., Kaiser, S., Blass, P., Senitschnig, N., Furian, G., Filtness, A., Talbot, R., Hancox, G. ,Pilkington-Cheney, F., Fortsakis, P., Isaias, B., & Gasper, C. (2020). Toolbox of recommended interventions to assist drivers in maintaining a safety tolerance zone. Deliverable No. 3.3. EC H2020 project i-DREAMS.

Cai, M., Yazdi, M. A. A., Mehdizadeh, A., Hu, Q., Vinel, A., Davis, K., Xian, H., Megahed, F. M., & Rigdon, S. E. (2021). The association between crashes and safety-critical events: Synthesized evidence from crash reports and naturalistic driving data among commercial truck drivers. Transportation Research Part C: Emerging Technologies, 126, 103016. https://doi.org/10.1016/j.trc.2021.103016

Cantor, D. E., Corsi, T. M., Grimm, C. M., & Özpolat, K. (2010). A driver focused truck crash prediction model. Transportation Research Part E: Logistics and Transportation Review, 46(5), 683–692. https://doi.org/10.1016/j.tre.2009.08.011

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician, 35(3), 124–129. https://doi.org/10.2307/2683975

Dablanc, L., Heitz, A., Rai, H. B., & Diziain, D. (2022). Response to COVID-19 lockdowns from urban freight stakeholders: An analysis from three surveys in 2020 in France, and policy implications. Transport Policy, 122, 85–94. https://doi.org/10.1016/j.tranpol.2022.04.020

De Winter, J. C. F., Dodou, D., & Stanton, N. A. (2015). A quarter of a century of the DBQ: Some supplementary notes on its validity with regard to accidents. Ergonomics, 58(10), 1745-1769. https://doi.org/10.1080/00140139.2015.1030460

Desai, J., Li, H., Mathew, J. K., Cheng, Y.-T., Habib, A., & Bullock, D. M. (2021). Correlating hard-braking activity with crash occurrences on interstate construction projects in Indiana. Journal of Big Data Analytics in Transportation, 3, 27–41. https://doi.org/10.1007/s42421-020-00024-x

Dorn, L., Stephen, L., Af Wåhlberg, A., & Gandolfi, J. (2010). Development and validation of a self-report measure of bus driver behaviour. Ergonomics, 53(12), 1420–1433. https://doi.org/10.1080/00140139.2010.532882

Dreger, F. A., De Winter, J. C. F., & Happee, R. (2020). How do drivers merge heavy goods vehicles onto freeways? A semi-structured interview unveiling needs for communication and support. Cognition, Technology & Work, 22, 825–842. https://doi.org/10.1007/s10111-019-00601-3

Duke, J., Guest, M., & Boggess, M. (2010). Age-related safety in professional heavy vehicle drivers: A literature review. Accident Analysis & Prevention, 42(2), 364–371. https://doi.org/10.1016/j.aap.2009.09.026

Elbert, R., Roeper, F., Vanelslander, T., & Cavallaro, F. (2023). Freight operations in the European Union during the COVID-19 pandemic: A multicountry comparison. In J. Zhang & Y. Hayashi (Eds.), Transportation amid pandemics: Lessons learned from COVID-19 (pp. 143–152). Elsevier. https://doi.org/10.1016/B978-0-323-99770-6.00021-1

Engström, J., Miller, A., Huang, W., Soccolich, S., Machiani, S. G., Jahangiri, A., Dreger, F., & De Winter, J. (2019). Behavior-based predictive safety analytics–Pilot study. Final report. Virginia Tech Transportation Institute. https://rosap.ntl.bts.gov/view/dot/49167

European Commission. (2021). Road traffic fatalities in the EU in 2019. https://transport.ec.europa.eu/system/files/2021-11/collision-matrix-2019.pdf

European Environment Agency. (2022). Reducing greenhouse gas emissions from heavy-duty vehicles in Europe. https://www.eea.europa.eu/publications/co2-emissions-of-new-heavy

European Road Safety Observatory. (2021). Facts and figures – Buses / Coaches / Heavy Goods Vehicles. https://road-safety.transport.ec.europa.eu/system/files/2022-03/FF_buses_hgv_20220209.pdf

Figueredo, G. P., Agrawal, U., Mase, J. M. M., Mesgarpour, M., Wagner, C., Soria, D., Garibaldi, J. M., Siebers, P.-O., & John, R. I. (2019). Identifying heavy goods vehicle driving styles in the United Kingdom. IEEE Transactions on Intelligent Transportation Systems, 20(9), 3324–3336. https://doi.org/10.1109/TITS.2018.2875343

Galal, A., Donmez, B., & Roorda, M. J. (2023). Improving Truck Driver and Vulnerable Road User Interactions Through Driver Training: An Interview Study With Canadian Subject Matter Experts. Transportation Research Record, 2677(12), 398-408. https://doi.org/10.1177/03611981231168123

Giuffrida, N., Fajardo-Calderin, J., Masegosa, A. D., Werner, F., Steudter, M., Pilla, F. (2022). Optimization and machine learning applied to last-mile logistics: A review. Sustainability, 14(9), 5329. https://doi.org/10.3390/su14095329

Goldenbeld, C., Reurings, M., Van Norden, Y., & Stipdonk, H. (2013). Crash involvement of motor vehicles in relationship to the number and severity of traffic offenses: An exploratory analysis of Dutch traffic offenses and crash data. Traffic Injury Prevention, 14(6), 584–591. https://doi.org/10.1080/15389588.2012.743125

Guest, M., Boggess, M. M., & Duke, J. M. (2014). Age-related annual crash incidence rate ratios in professional drivers of heavy goods vehicles. Transportation Research Part A: Policy and Practice, 65, 1–8. https://doi.org/10.1016/j.tra.2014.04.003

Hancock, P. A., Lesch, M., & Simmons, L. (2003). The distraction effects of phone use during a crucial driving maneuver. Accident Analysis & Prevention, 35(4), 501–514. https://doi.org/10.1016/S0001-4575(02)00028-3

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction:

Impacts on drivers' visual behavior and braking performance. Accident Analysis & Prevention, 39(2), 372–379. https://doi.org/10.1016/j.aap.2006.08.013
Henckaerts, R., & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. Insurance: Mathematics and Economics, 105, 79–95. https://doi.org/10.1016/j.insmatheco.2022.03.011
Hickman, J. S., & Hanowski, R. J. (2012). An assessment of commercial motor vehicle driver distraction using naturalistic driving data. Traffic Injury Prevention, 13(6), 612–619. https://doi.org/10.1080/15389588.2012.683841
Hirsch, P., Choukou, M. A., Bellavance, F. (2017). Transfer of training in basic control skills from truck simulator to real truck. Transportation Research Record, 2637(1), 67–73. https://doi.org/10.3141/2637-08
Huang, J. L., & Ford, J. K. (2012). Driving locus of control and driving behaviors: Inducing change through driver training. Transportation Research Part F: Traffic Psychology and Behaviour, 15(3), 358–368. https://doi.org/10.1016/j.trf.2011.09.002
Huang, Y.-H., Zohar, D., Robertson, M. M., Garabet, A., Lee, J., & Murphy, L. A. (2013). Development and validation of safety climate scales for lone workers using truck drivers as exemplar. Transportation Research Part F: Traffic Psychology and Behaviour, 17, 5–19. https://doi.org/10.1016/j.trf.2012.08.011
Islam, B. M., & Hernandez, S. (2016). Fatality rates for crashes involving heavy vehicles on highways: A random parameter Tobit regression approach. Journal of Transportation Safety & Security, 8(3), 247–265. https://doi.org/10.1080/19439962.2015.1027071
Jun, J., Ogle, J., & Guensler, R. (2007). Relationships between crash involvement and temporal-spatial driving behavior activity patterns: Use of data for vehicles with global positioning systems. Transportation Research Record, 2019(1), 246–255. https://doi.org/10.3141/2019-29
Kamla, J., Parry, T., & Dawson, A. (2019). Analysing truck harsh braking incidents to study roundabout accident risk. Accident Analysis & Prevention, 122, 365–377. https://doi.org/10.1016/j.aap.2018.04.031
Katrakazas, C., Michelaraki, E., Sekadakis, M., & Yannis, G. (2020). A descriptive analysis of the effect of the COVID-19 pandemic on driving behavior and road safety. Transportation Research Interdisciplinary Perspectives, 7, 100186. https://doi.org/10.1016/j.trip.2020.100186
Khorram, B., Af Wåhlberg, A. E., & Tavakoli Kashani, A. (2020). Longitudinal jerk and celeration as measures of safety in bus rapid transit drivers in Tehran. Theoretical Issues in Ergonomics Science, 21(5), 577–594. https://doi.org/10.1080/1463922X.2020.1719228
Kohli, E., Ptak, J., Smith, R., Taylor, E., Talbot, E. A., Kirkland, K. B. (2009). Variability in the Hawthorne effect with regard to hand hygiene performance in high- and low-performing inpatient care units. Infection Control & Hospital Epidemiology, 30(3), 222–225. https://doi.org/10.1086/595692
Kondyli, A., & Elefteriadou, L. (2009). Driver behavior at freeway-ramp merging areas: Focus group findings. Transportation Research Record, 2124(1), 157–166. https://doi.org/10.3141/2124-15
Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. ASTIN Bulletin: The Journal of the IAA, 46(1), 39–69. https://doi.org/10.1017/asb.2015.25
Lemp, J. D., Kockelman, K. M., & Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit models. Accident Analysis & Prevention, 43(1), 370–380. https://doi.org/10.1016/j.aap.2010.09.006
Li, Y., Zhao, L., & Rilett, L. R. (2019). Driving performances assessment based on speed variation using dedicated route truck GPS data. IEEE Access, 7, 51002–51013. https://doi.org/10.1109/ACCESS.2019.2909572
Ma, Y.-L., Zhu, X., Hu, X., & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A: Policy and Practice, 113, 243–258. https://doi.org/10.1016/j.tra.2018.04.013
Mane, A., Djordjevic, B., & Ghosh, B. (2021). A data-driven framework for incentivising fuel-efficient driving behaviour in heavy-duty vehicles. Transportation Research Part D: Transport and Environment, 95, 102845. https://doi.org/10.1016/j.trd.2021.102845
Mase, J. M., Majid, S., Mesgarpour, M., Torres, M. T., Figueredo, G. P., Chapman, P. (2020). Evaluating the impact of heavy goods vehicle driver monitoring and coaching to reduce risky behaviour. Accident Analysis & Prevention, 146, 105754. https://doi.org/10.1016/j.aap.2020.105754
Michelaraki, E., Katrakazas, C., Yannis, G., Filtness, A., Talbot, R., Hancox, G., Pilkington-Cheney, F., Brijs, K., Ross, V., Dirix, H., Neven, A., Paul, R., Brijs, T., Fortsakis, P., Frantzola, E. K., Taveira, R. (2021). Post-trip safety interventions: State-of-the-art, challenges, and practical implications. Journal of Safety Research, 77, 67–85. https://doi.org/10.1016/j.jsr.2021.02.005
Mooren, L., Grzebieta, R., Williamson, A., Olivier, J., Friswell, R. (2014). Safety management for heavy vehicle transport: A review of the literature. Safety Science, 62, 79–89. https://doi.org/10.1016/j.ssci.2013.08.001
Organization for Economic Co-operation and Development. (2006). Young drivers: The road to safety. Paris, France: OECD Publishing. https://read.oecd-ilibrary.org/transport/young-drivers_9789282113356-en
Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. Transportation Research Part A: Policy and Practice, 61, 27–40. https://doi.org/10.1016/j.tra.2013.11.010
Park, G. D., Hickman, J. S., Pitoniak, S. P., & Rosenthal, T. J. (2018). Challenges in hazard detection for commer-

cial motor vehicle drivers: A driving simulator training study. Transportation Research Record, 2672(34), 29–38. https://doi.org/10.1177/0361198118792132

Perez, M. A., Sudweeks, J. D., Sears, E., Antin, J., Lee, S., Hankey, J. M., & Dingus, T. A. (2017). Performance of basic kinematic thresholds in the identification of crash and near-crash events within naturalistic driving data. Accident Analysis & Prevention, 103, 10–19. https://doi.org/10.1016/j.aap.2017.03.005

RDW. (2022). Open data RDW: Gekentekende voertuigen [Open data RDW: Registered vehicles]. https://opendata.rdw.nl/Voertuigen/Open-Data-RDW-Gekentekende_voertuigen/m9d7-ebf2

Reason, J. (1995). A systems approach to organizational error. Ergonomics, 38(8), 1708–1721. https://doi.org/10.1080/00140139508925221

Reed, W. S., & Keskin, A. T. (1987). Vehicular response to emergency braking. SAE Transactions, 96(2), 554–560. https://doi.org/10.4271/870501

Ribeiro, P., Krause, A. F., Meesters, P., Kural, K., Van Kolfschoten, J., Büchner, M. A., Ohlmann, J., Ressel, C., Benders, J., & Essig, K. (2021). A VR truck docking simulator platform for developing personalized driver assistance. Applied Sciences, 11(19), 8911. https://doi.org/10.3390/app11198911

Scheffel, M., Kirschenmann, U., Taske, A., Adloff, K., Kiesel, M., Klemke, R., & Wolpers, M. (2013). Exploring LogiAssist – The mobile learning and assistance platform for truck drivers. In D. Hernández-Leo, T. Ley, R. Klamma, & A. Harrer (Eds.), Scaling Up Learning for Sustained Impact. EC-TEL 2013 (pp. 343–356). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40814-4_27

Simons-Morton, B. G., Ouimet, M. C., Wang, J., Klauer, S. G., Lee, S. E., & Dingus, T. A. (2009). Hard braking events among novice teenage drivers by passenger characteristics. Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Big Sky, MT, 236–242.

Soccolich, S. A., & Hickman, J. S. (2014). Potential reduction in large truck and bus traffic fatalities and injuries using Lytx's DriveCam program. Final Report. Transportation Institute, Virginia Tech.

Sperry, D., Schuster, A. M., Cotten, S. R., Agrawal, S., Mack, E., Britt, N., & Liberman, J. (2022). Trucking in the era of COVID-19. American Behavioral Scientist. https://doi.org/10.1177/00027642211066039

Stankevich, I., Korishchenko, K., Pilnik, N., & Petrova, D. (2022). Usage-based vehicle insurance: Driving style factors of accident probability and severity. Journal of Transportation Safety & Security, 14(10), 1633–1654. https://doi.org/10.1080/19439962.2021.1941459

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2017). Impact of congestion and traffic flow on crash frequency and severity: Application of smartphone-collected GPS travel data. Transportation Research Record, 2659(1), 43–54. https://doi.org/10.3141/2659-05

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis & Prevention, 98, 139–148. https://doi.org/10.1016/j.aap.2016.10.006

Valenzuela, L. S., & Burke, M. J. (2020). Toward a greater understanding of Colombian professional truck drivers' safety performance. Transportation Research Part F: Traffic Psychology and Behaviour, 73, 188–204. https://doi.org/10.1016/j.trf.2020.06.006

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society: Series C (Applied Statistics), 67(5), 1275–1304. https://doi.org/10.1111/rssc.12283

Walnum, H. J., & Simonsen, M. (2015). Does driving behavior matter? An analysis of fuel consumption data from heavy-duty trucks. Transportation Research Part D: Transport and Environment, 36, 107–120. https://doi.org/10.1016/j.trd.2015.02.016

Wijnen, W., & Stipdonk, H. (2016). Social costs of road crashes: An international analysis. Accident Analysis & Prevention, 94, 97–106. https://doi.org/10.1016/j.aap.2016.05.005

Wouters, P. I. J., & Bos, J. M. J. (2000). Traffic accident reduction by monitoring driver behaviour with in-car data recorders. Accident Analysis & Prevention, 32(5), 643–650. https://doi.org/10.1016/S0001-4575(99)00095-0

Yasin, Y. J., Grivna, M., & Abu-Zidan, F. M. (2021). Global impact of COVID-19 pandemic on road traffic collisions. World Journal of Emergency Surgery, 16, 51. https://doi.org/10.1186/s13017-021-00395-8

Zhou, T., & Zhang, J. (2019). Analysis of commercial truck drivers' potentially dangerous driving behaviors based on 11-month digital tachograph data and multilevel modeling approach. Accident Analysis & Prevention, 132, 105256. https://doi.org/10.1016/j.aap.2019.105256

Zohar, D., Huang, Y.-H., Lee, J., & Robertson, M. (2014). A mediation model linking dispatcher leadership and work ownership with safety climate as predictors of truck driver safety performance. Accident Analysis & Prevention, 62, 17–25.

Supplementary Material

One driver was excluded because they could not be connected to a vehicle due to a data logging error. Subsequently, sessions were extracted (n = 29,446), and invalid sessions (n = 2,826) were removed from the dataset. An explanation of the session exclusions is provided below. It is important to note that reasons for session exclusion frequently occurred simultaneously.

- Traveled distance of 0 m: 1033 sessions. This could be the result of there not being at least two sampling points available.
- Duration of session of 0 s: 951 sessions. This could be the result of there not being at least two sampling points available.
- Mean speed greater than 90 km/h: 56 sessions. A mean speed greater than 90 km/h was judged to be physically impossible. It could be the result of an odometer error or other recording anomaly.
- Mean speed lower than 10 km/h: 74 sessions. This could result from an odometer error or recording anomaly. It could also be the result of the driver performing many low-speed tasks (e.g., parking, loading/unloading), which was expected to dilute the assessment.
- Large negative increment in cumulative event data: 29 sessions. Event data should only be monotonically increasing. A large negative increment could signal a problem with the data, such as a device reset.
- Large negative increment in odometer data (1000 km or more): 12 sessions. Odometer data should only be increasing. While small negative changes could occur due to mobile data packages received with delay, a large decrement could signal an anomaly.
- Driver with less than 1 hour of data: 1 session. This driver was excluded because of too little data.

Table 6.5: Spearman rank-order intercorrelations of the driving measures, damage incidents, and fines.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 Total driving time																		
2 Total driving distance	0.98																	
3 Number of days with data	0.91	0.87																
4 Engine capacity (weighted by driving time)	-0.03	0.02	-0.01															
5 Number of vehicle switches per hour of driving	-0.47	-0.43	-0.33	0.20														
6 Percentage night shift driving	0.10	0.19	0.08	0.38	0.13													
7 Mean speed	0.17	0.31	0.03	0.13	-0.19	0.43												
8 Number of harsh brakes per hour of driving	0.13	0.06	0.20	-0.29	-0.02	-0.33	-0.37											
9 Speeding duration per hour of driving	0.16	0.24	0.11	0.12	-0.07	0.35	0.51	-0.02										
10 Duration of excessive engine torque per hour of driving	0.04	0.06	0.07	0.30	0.16	-0.02	0.00	0.16	0.07									
11 Cruise control duration per hour of driving	0.24	0.35	0.12	0.21	-0.13	0.44	0.76	-0.42	0.26	-0.01								
12 Fuel consumption per kilometer of driving	-0.18	-0.17	-0.02	0.52	0.50	0.09	-0.25	0.00	0.08	0.38	-0.19							
13 Total number of damage incidents	0.51	0.50	0.50	-0.09	-0.04	-0.05	-0.01	0.22	0.09	-0.02	0.08	-0.03						
14 Number of damage incidents, claimed from insurance	0.39	0.37	0.37	-0.10	-0.09	-0.17	-0.04	0.19	0.03	-0.03	0.03	-0.08	0.83					
15 Number of damage incidents, not claimed from insurance	0.45	0.44	0.45	-0.07	0.03	0.05	0.00	0.20	0.10	-0.04	0.09	0.01	0.84	0.45				
16 Number of damage incidents, costs recovered from other party	0.25	0.27	0.28	0.08	-0.06	0.02	0.04	0.08	0.16	0.08	0.12	0.16	0.33	0.38	0.19			
17 Number of damage incidents, costs not recovered from other party	0.48	0.47	0.47	-0.12	-0.01	-0.05	0.00	0.24	0.08	-0.04	0.06	-0.06	0.98	0.79	0.85	0.15		
18 Number of damage incidents, driving-related	0.40	0.40	0.39	-0.12	0.02	-0.07	-0.01	0.19	-0.04	-0.07	0.07	-0.06	0.90	0.80	0.73	0.13	0.92	
19 Number of fines	0.36	0.33	0.35	-0.14	-0.08	-0.10	-0.04	0.32	0.11	-0.04	-0.05	-0.08	0.45	0.41	0.38	0.18	0.45	0.38

Note: n = 180, except for Measure 12, where n = 158.

Part III

Measuring Driving Behavior: Towards Context-Aware Methods

7

Using AI from Automated Driving Systems to Assess Human Driving Behavior

This chapter was published as: Driessen, T., Siebinga, O., De Boer, T., Dodou, D., De Waard, D., & De Winter, J. (2024). How AI from Automated Driving Systems Can Contribute to the Assessment of Human Driving Behavior. Robotics, 13(12), 169. https://doi.org/10.3390/robotics13120169

Abstract

This paper proposes a novel approach to measuring human driving performance by using the AI capabilities of automated driving systems, illustrated through three example scenarios. Traditionally, the assessment of human driving has followed a bottom-up methodology, where raw data are compared to fixed thresholds, yielding indicators such as the number of hard braking events. However, acceleration threshold exceedances are often heavily influenced by the driving context. We propose a top-down context-aware approach to driving assessments, in which recordings of human-driven vehicles are analyzed by an automated driving system. By comparing the human driver's speed to the AI's recommended speed, we derive a level of disagreement that can be used to distinguish between hard braking caused by aggressive driving and emergency braking in response to a critical event. The proposed method may serve as an alternative to the metrics currently used by some insurance companies and may serve as a template for future AI-based driver assessment.

7.1. Introduction

Human drivers are increasingly being evaluated by algorithms, particularly in terms of fuel efficiency and safety. Safe driving is of interest to insurance companies, fleet owners, and licensing organizations. For example, some insurance companies now offer incentives for defensive driving based on acceleration-based metrics such as hard braking (exceeding longitudinal acceleration thresholds), sharp cornering (exceeding lateral acceleration thresholds), and speeding (Admiral, 2024; Allianz, 2024; Allstate, 2024; ANWB, 2024; Direct Assurance, 2024; Nationwide, 2024). Similarly, Tesla's Safety Score Beta calculates a score based on factors like hard braking, speeding, and following distance (Tesla, 2024b).

While studies have found that acceleration-based metrics can be used to predict drivers' likelihood of being involved in crashes and damage incidents (e.g., Cai et al. 2021; Driessen et al. 2024; Hunter et al. 2021; Ma et al. 2018), there are challenges in holding drivers accountable for exceeding acceleration thresholds. One issue is that the need for high acceleration levels can depend on the road type. For example, urban areas typically require more frequent (hard) braking than rural roads. One solution is to make the driver's assessment dependent on the road type based on GPS (e.g., Guillen et al. 2024; Melman et al. 2021; Moosavi & Ramnath, 2023). However, such location-based assessments lack the ability to account for local road geometry.

A second concern is that traffic conditions can play a key role. For example, an unexpected event, such as a pedestrian crossing the road or a lead vehicle suddenly decelerating, may force the driver to brake hard. When evaluating drivers based on deceleration events or offering discounts and rewards for not showing such events, it is important to consider the traffic conditions in which these actions occur. One approach to include contextual information is to compare the speed of the vehicle with the speed of other traffic or to take weather conditions into consideration (e.g., Ma et al. 2018; Masello et al. 2023; Reig Torra et al. 2023). However, such global statistical measures will only provide limited context.

A third concern is the risk of unintended consequences. If drivers are penalized for hard braking, they may hesitate to brake hard when needed, leading to dangerous situations. In summary, simply counting "dangerous events" based on predefined acceleration thresholds may not be a reliable indicator of an individual's driving behavior.

In recent years, computational resources in vehicles and the sophistication of algorithms have increased to the point that vehicles have become advanced enough to perform most of the driving tasks themselves. These developments can be seen in systems like those of Waymo (2024b; Hu et al. 2023), Lyft (2024; Li et al. 2023), Tesla's "Full Self-Driving" (FSD), and AI/data companies such as NVIDIA (2024) and comma.ai (2024b; Dorr, 2024). A dilemma, however, arises from the fact that these systems are not yet perfect. They either have to operate in limited regions or still require human attention and intervention. For example, current Tesla FSD systems require drivers to keep their hands on the wheel or remain attentive to the road, as monitored by a cabin camera (Tesla, 2024a), whereas the autonomous fleets of Waymo and Zoox rely on remote operators to solve difficult situations (Waymo 2024a; Zoox 2020; see also Lu & Shi 2024).

Given that vehicles are becoming increasingly capable of driving automatically in a human-like manner, yet still cannot drive wholly automatically, we propose the concept of using automation to assist in measuring human driving behavior. The idea we explore is to evaluate human driving by comparing it to AI-generated driving behavior. This concept can be traced back over three decades to the GIDS (Generic Intelligent Driver Support) project in the late 20th century (Michon, 1993a), where a reference driving behavior was generated through computer-simulated driving. However, at the time, this idea was considered "(too) far ahead of its time" (Michon, 1993b) and did not gain traction. Today, however, it has become a realistic possibility.

Comparing human driving behavior to an AI's intended plan offers advantages over the aforementioned assessment strategies. For example, in the case of a detected hard-braking event, an AI system could be used to retrospectively differentiate between reckless driving and a sudden, unavoidable event requiring an immediate response. A high level of disagreement between the driver and the AI may suggest reckless driving, where both the AI and the human would have had the time to respond to a visible obstruction. On the other hand, a high level of agreement could indicate a justified reaction, where neither the driver nor the AI could anticipate the need to brake sooner. An algorithm based on fixed acceleration thresholds, however, would flag both scenarios as instances of unsafe driving.

It is important to note that our proposed approach does not assume the AI is infallible. The idea is that even an imperfect AI can be effective for assessing human driving behavior, as long-term data collection may reveal that certain drivers consistently deviate more from AI-referenced behavior than others.

Aim

This study presents three driving scenarios conducted in a driving simulator. After recording, the scenarios were analyzed using Openpilot, an open-source platform (comma.ai, 2024a) that can be described as SAE level 2 vehicle automation (Chen et al. 2022). Online videos (comma, 2022; Greer Viau, 2021) demonstrate that Openpilot is capable of driving vehicles both on highways and in urban environments. We fed the recordings of our driving scenarios to the Openpilot system and analyzed the internal metrics of Openpilot to understand how it interprets each situation.

This study aims to present a technical demonstration of how AI can be used to assess human driving behavior. Using three driving scenarios in a simulator environment, we demonstrate the feasibility of comparing human driving behavior with driving behavior of an existing automated driving system.

Our scenarios include hard braking, which is practically relevant for applications like insurance assessments. However, the technical framework we demonstrate is adaptable to other scenarios due to the generalizable nature of the automated driving system. In the discussion section, we illustrate this broader potential impact with specific examples. Additionally, we contribute a publicly accessible adaptation of Openpilot's tools, which enables the replay of externally recorded driving data to support future research.

7.2. Method

7.2.1 Setup

Three demonstration driving scenarios were driven by the first author of this paper. The video and vehicle data were recorded in a virtual world using JOAN (Beckers et al. 2023), a Python software package developed to enable human-in-the-loop experiments in the CARLA driving simulator (Dosovitskiy et al. 2017). JOAN provides the possibility to connect USB steering wheels to CARLA vehicles (in this case, the Logitech Driving Force G923 steering wheel), record vehicle data, and create reproducible experiments with other traffic following predefined trajectories. The simulation ran on a PC running Windows 11. The repository contains the experimental setup files that were used within JOAN.

7.2.2 Scenarios

The three scenarios were chosen to be simple to explain and feasible for simulator implementation without the need for complex hard-coded choreographies. Moreover, the scenarios were chosen to make our specific point: that AI-based assessments can be used to help classify a hard-braking event as either unnecessary or necessary. In other words, while traditional assessment methods, such as those used by insurance companies, would classify every hard-braking event as undesirable and contributing to the driver's risk profile, we illustrate that hard braking in a surprise emergency condition does not need to be marked as undesirable but rather as desirable and necessary.

The scenarios involved two situations where an obstruction was clearly visible and the driver of the ego-vehicle reacted either aggressively (i.e., braked too late) or calmly (i.e., braked in time), and a third scenario where a surprise event justified hard braking by the driver. These scenarios are illustrated in Figure 7.1, where orange represents the obstructions encountered in the Aggressive and Calm scenarios, and pink indicates the moving obstruction in the Surprise scenario. Figure 7.2 shows a view during the Surprise scenario, and Figure 7.3 shows the buses from the Aggressive and Calm scenarios. Video files of the three scenarios can be found in the repository.

In all scenarios, the ego-vehicle began from the same starting point and followed a straight road leading to a T-junction (Figure 7.1). The driver controlling the vehicle aimed at a target speed of 50 km/h.

Calm scenario (driver of ego-vehicle did not brake hard)

In the *Calm* scenario, two stationary buses were positioned on the road, forcing the ego-vehicle to stop behind them. The experimenter achieved this by braking moderately and gradually letting the vehicle come to a stop.

Aggressive scenario (driver of ego-vehicle braked hard, but hard braking was avoidable)

The *Aggressive* scenario was identical to the *Calm* scenario, except for the behavior of the ego-vehicle. The experimenter maintained the target speed until approaching the buses and then applied a high braking input at the last moment, to demonstrate an aggressive driving style.

Surprise scenario (driver of ego-vehicle braked hard, and hard braking was unavoidable)

The *Surprise* scenario contained an unforeseeable event. The scenario started with the ego-vehicle in the same position as in the previous two scenarios. Unlike the other scenarios, no stationary vehicles obstructed the road, allowing the ego-vehicle to proceed straight through the T-junction. However, a bus (depicted in pink in Figure 7.1) approached the junction at the same time as the ego-vehicle. The bus was hidden from the ego-vehicle's view before entering the junction due to a high wall along the sidewalk. The bus followed a predefined path, turning left at the junction without yielding to the ego-vehicle (Figure 7.2). The timing of the bus's appearance forced the ego-vehicle to brake abruptly to avoid a collision.

7.2.3 Analysis

The Openpilot software was adapted to work with pre-recorded driving data to enable a comparison between human driving and AI-generated predictions. This adaptation involved modifying Openpilot's existing simulation module, originally designed to interface with the CARLA driving simulator for development purposes. Driving in the simulator with Openpilot running in real time could also have been a suitable setup for this demonstration. However, we first recorded the data and replayed it with Openpilot, as this also enables the module to work with other sources of data, such as dashcam videos supplemented with a CSV file containing speed and acceleration information (e.g., logged using a phone), allowing for the post-hoc analysis of other existing driving datasets.

The adapted module allows two primary inputs:

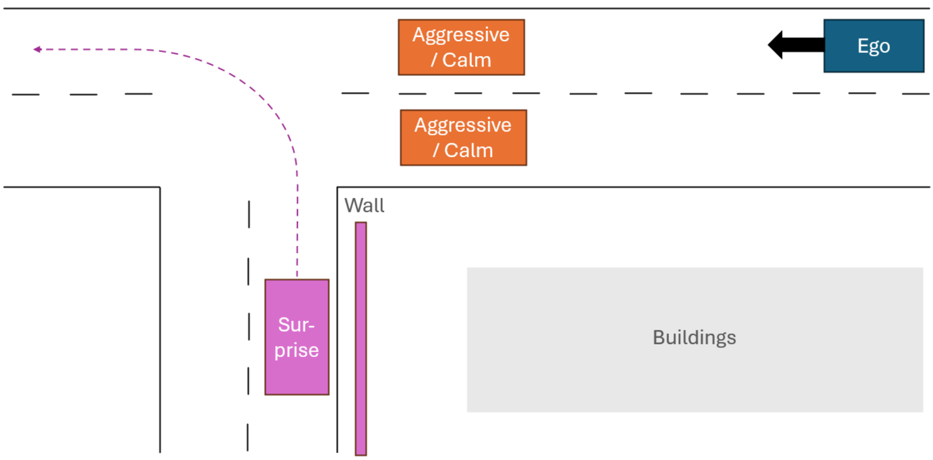


Figure 7.1: Demonstration scenarios. *Aggressive* and *Calm*: The ego-vehicle (blue) approached stationary buses (orange) at a T-junction requiring a full stop, which the driver of the ego-vehicle executed aggressively or calmly. *Surprise*: The ego-vehicle drove along an empty road until a bus (pink) emerged from behind a wall, necessitating an emergency brake to avoid a collision.

Figure 7.2: Video stills from the *Surprise* scenario.

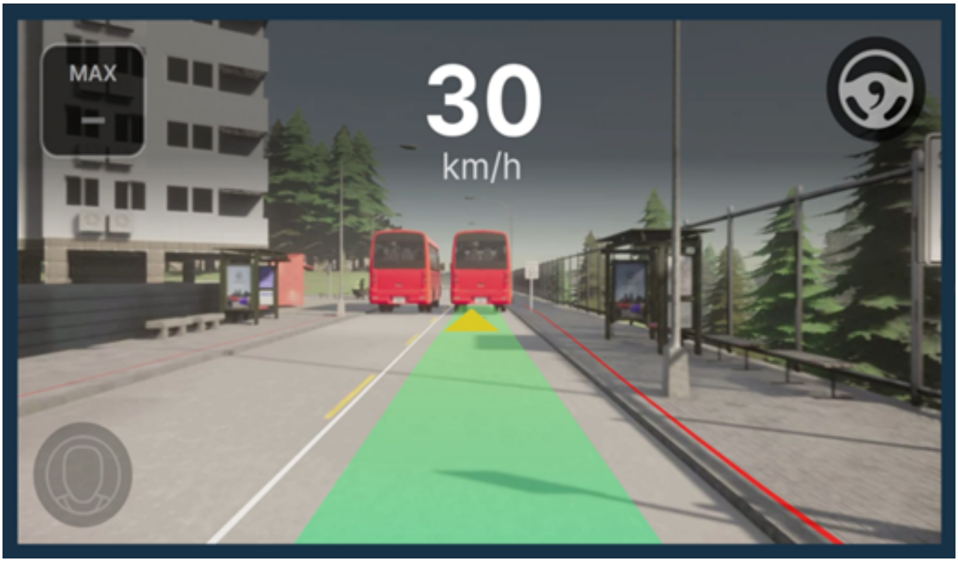


Figure 7.3: Openpilot user interface during the *Calm* scenario. The *Aggressive* scenario uses the same setup. The green band displays the predicted trajectory, and the lines provide information about the lane line estimates. The yellow triangle indicates a detected lead vehicle.

1. A 1928×1208 pixel forward-facing driving video recorded at 20 frames per second, simulating the visual input an autonomous system would receive.

2. A corresponding CSV file with a row for each video frame containing vehicle state data, including speed, bearing, steering angle, brake, and throttle inputs.

Thus, the video file provided the visual context, while the CSV file supplied data about the vehicle state. The module processed these inputs sequentially, mimicking the real-time data flow that Openpilot would experience in a live driving situation. Note that the module still works when omitting some of this data, such as brake and throttle input.

This allows for some flexibility when using other datasets that may not feature all vehicle state data. Our limited experimentation with other sources suggests that reasonable estimates can often still be obtained. Figure 7.3 shows a screenshot of the Openpilot user interface during the replay of one of our recordings.

The software was run on an Ubuntu 20.04 desktop PC. As the input data were processed by Openpilot, the system's generated predictions and plans were logged to a CSV file for later analysis. For the current demonstration, we used the desired speed originating from Openpilot's longitudinal planner module. Other variables logged by the module that are suitable for comparison to human driving execution but not used are discussed in the Appendix. The full implementation details can be found in the project repository.

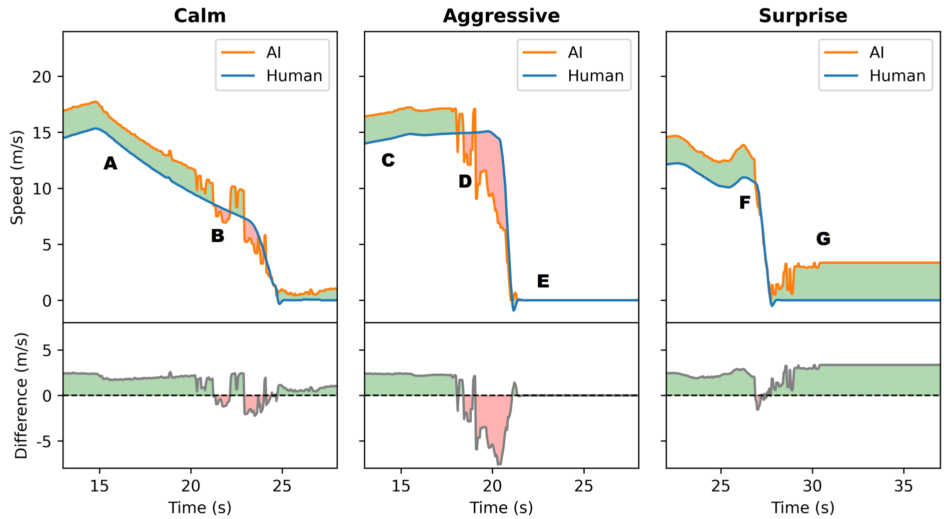


Figure 7.4: Measured speed (blue), AI's desired speed (orange), and the difference between the two, for all scenarios. To reduce noise in the visualization, a median filter with a window size of 3 samples (0.15 s) was applied to the prediction signal. The letters denote specific moments in the scenarios that are discussed in the text.

7.3. Results

Figure 7.4 compares Openpilot's desired speed with the actual speed at which the human drove. A positive difference (shown in green) can be interpreted as a desire by the model to obtain a higher speed (a desire to accelerate), while a negative difference (shown in red) indicates a desire to decelerate.

7.3.1 Calm scenario

In the *Calm* scenario, the driver drove slower than the AI deemed appropriate for the current situation (Figure 7.4 left, annotation A). As the two stationary buses became closer, the driver decelerated, and the AI indicated a similar desire. At some point, the model suggested a speed lower than the current speed (Figure 7.4 left, B), indicating the AI preferred to slow down faster for the oncoming obstacles. The human driver also decelerated around the same time, indicated by the steeper slope, coming to a full stop.

7.3.2 Aggressive scenario

In the *Aggressive* scenario, we see a similar start, with Openpilot suggesting a higher speed for the current empty road (Figure 7.4 middle, C). As the car approached the obstacles, the driver kept his speed, while the AI suggested deceleration (Figure 7.4 middle, D). Only at the last moment, the driver decided to brake, and the speed dropped to 0 (Figure 7.4 middle, E). When comparing the *Calm* and the *Aggressive* scenarios, we see that, in both graphs, some level of disagreement was present. However, the graph of the *Aggressive* scenario clearly shows a larger negative difference.

7.3.3 Surprise scenario

The *Surprise* scenario started similarly until the surprise event (a bus suddenly appearing and cutting off the path of the ego-vehicle) happened. The model only suggested deceleration after the surprise event, around the same time the human driver was decelerating (Figure 7.4 right, F). Further note that, after the full stop, the model suggested increasing speed again (Figure 7.4 right, G), since the bus had continued on its way and the ego-vehicle was alone again.

The *Aggressive* and *Surprise* scenarios would both likely have been flagged in traditional assessment methods. When presented with the Aggressive scenario, Openpilot suggested braking earlier, as judged by the difference graph at the bottom. In other words, it saw reason to decelerate before the human decided to decelerate. In the *Surprise* scenario, however, the model suggested deceleration at about the same time the human started decelerating. There was no large negative peak visible in the difference graph at the bottom, as the bus was a surprise for Openpilot as well. This means that this stop was executed in a manner that was more similar to the way the AI would have executed it than in the case of the *Aggressive* scenario. In other words, the AI's interpretation allows us to distinguish between an unnecessary hard brake and a necessary hard brake.

In the above comparisons, the height of the difference peak was highest in the *Aggressive* scenario, which means it could be identified as the braking event where there was the most disagreement between the AI and the human execution. This means that the current method provides a way to judge whether a hard brake event was justified after this brake event happened. A possible implementation could be, for every hard brake event (when a certain static threshold was exceeded), to let the AI assess whether the brake event was justified, in order to improve the validity of the assessments of driver behavior. Other metrics that could be considered are the total area between the two curves or the time since the first model suggested deceleration until a full stop.

7.4. Discussion

This study demonstrated the potential of using AI from an automated driving system to assess human driving behavior. The method compared the state of a human-driven vehicle with the AI's desired state. This approach allowed discrimination between necessary and unnecessary hard braking by using a variable that indicates the level of disagreement of the human driver's actions relative to the AI's recommendations.

Traditional methods for assessing driver behavior, such as the predefined thresholds used by insurance companies (e.g., for speed, acceleration, or braking), are situation-agnostic. In contrast, the AI-assisted method provides a context-aware evaluation by comparing the human driver's actions with model predictions. Driving examiners previously identified the lack of a holistic perspective as a key barrier to effective data-driven driver assessment (Driessen et al. 2021).

We achieved our results using Openpilot, a system that relies on a vision feed and CAN-bus data to control the vehicle automatically in many situations. One may argue that using Openpilot is overkill for the current demonstration purposes and that simpler, more widely used radar-based solutions, like those found in forward collision warning systems (FCWS) or adaptive cruise control (ACC), could have sufficed. An equivalent argument could be made that potential field methods, which examine the extent to which other objects and road users intrude in this field, could be used to quantify the level of risk in driving (e.g., Hennessey et al. 1995; Li et al. 2020; Kolekar et al. 2021). These existing concepts, however, are still limited in their ability to interpret the overall driving context. Our approach relies on a visual understanding of the scenario, including, e.g., traffic signs. It uses a neural network (comma.ai, 2022) that provides an estimate of the appropriate speed given the entire driving situation, rather than providing an estimate of risk based on relatively simple indicators such as time to collision (TTC) or potential field intrusion. The system's ability to process visual information and make context-aware decisions mirrors the decision-making process of human drivers.

7.5. Limitations

The system that we used still has limitations. Like all current automated driving systems, it can make errors in perception, prediction, and decision-making. For example, it may misclassify objects, fail to anticipate complex traffic situations, or make wrong speed recommendations. However, our method's effectiveness does not require the AI to be infallible. Rather than relying on individual instances of disagreement, our approach is envisioned to identify patterns in driving behavior over extended periods. This aggregation helps mitigate the impact of occasional AI errors. Even with imperfect AI, the additional context helps distinguish between necessary and unnecessary hard braking events better than approaches relying on kinematic thresholds only.

As explained in the last paragraph of the Results section, the calculation and accompanying threshold values that distinguish a “high level of disagreement” from a “low level of disagreement” warrant further development and research with real datasets. Research will also be needed into “gray areas”, where, for example, the AI detects a precursor to a hazard and considers braking intervention necessary while a human driver may barely recognize it. To this end, for different traffic situations, the driving behavior or recommendations of expert drivers could be compared to those of AI. This, in turn, raises fundamental questions about who should ultimately be the arbiter in defining successful driving performance: an expert driver or an AI agent. It also brings up questions about the required quality of cameras and the level of intelligence an AI-based automated driving system should or could possess.

The demonstrated scenarios represent only a fraction of potential applications. Future research could examine disagreements in lateral driving behavior, i.e., identify discrepancies between a driver's chosen maneuver and the vehicle's suggested alternative. It should be noted here that, in some emergency situations, where the braking distance is too long to avoid a collision, an evasive steering maneuver may be the only way to prevent an accident (Allen et al. 2005). Additionally, the method could assess whether higher-level strategic decisions (e.g., Michon's model of driving behavior; Michon, 1985) can be evaluated using multiple agents as reference points. As AI continues to improve, its ability to assess human driving behavior will likely expand. This is especially relevant because improved AI systems may not always result in safer automated driving. As AI takes more control, human drivers may be kept out of the loop for longer periods, potentially leading to slower response times when manual intervention is required.

Finally, it is important to note that this study serves as a proof of concept only; it represents a single demonstration conducted by the author in a virtual environment. Further research is needed to validate the approach, which should include testing the system across a broader range of real-world driving scenarios with more human participants.

7.6. Recommendations

In the insurance industry, our proposed method could allow for fairer assessments of driving incidents. This could lead to more precise risk profiling and personalized insurance premiums based on individual driving styles. For fleet operators, the ability to detect risky driving behaviors and offer targeted coaching could help reduce accidents, improve driver performance, and lower operating costs. Additionally, our algorithmic assessment concept could benefit driver education and testing, as well as provide concurrent feedback on driving behavior after obtaining a license.

In academia, our work opens up possibilities in several ways. First, existing datasets with human driving videos and car state variables could be augmented with AI-generated predictions, removing the need for costly new data annotation or data collection efforts. Discrepancies could be used to extract critical situations from the datasets, for example. Secondly, future research could explore the use of multi-agent systems, where disagreements between two or more concurrently observing AI models could provide a more robust assessment of driver behavior. Using multiple AI systems would essentially provide the driver with a "committee" of reference drivers, reducing the potential for bias present in individual models. An approach in which multiple agents arbitrate their perspectives on current driving events was previously suggested by Fridman et al. (2019).

The current method assumes that hard braking events are evaluated after they occurred by observing whether the AI detected a need to decelerate earlier for these specific events. Future work could explore whether more continuous measures of disagreement can be developed. We propose an overall "deviation score", determined by the total level of disagreement between the driver and one (or multiple) AI(s) regarding variables such as speed, acceleration, steering input, and braking input.

The software was run on a desktop PC, with the data analyzed after the scenario recordings. However, future versions could allow real-time data collection in a real vehicle. Openpilot's plug-and-play design offers advantages over traditional onboard systems. It can be installed in different vehicle models (comma.ai, 2024b). In contrast, manufacturer-installed automation systems are typically integrated into onboard computers.

Though the model weights remain the private property of comma.ai, the software itself has been made open-source. The broader automotive industry still lags behind in embracing open-source frameworks. This closed nature may hinder innovation and restrict improvements in safety. We argue that manufacturers should, at a minimum, consider offering accessible APIs to allow developers and researchers to interact with vehicle data. This may be counterintuitive to traditional automotive manufacturers, where innovation is typically patented and kept private. However, some examples indicate that the industry is increasingly seeing the value in open-sourcing data and code. Waymo has released its Open Dataset (waymo-research, 2024), Tesla occasionally publishes repositories that may prove useful to outsiders (Tesla, 2024c), and the Automotive Grade Linux (Sivakumar et al. 2022) initiative is another example. By inviting public collaboration, manufacturers could benefit from faster development cycles, while the broader community gains access to tools for creating safer vehicles. When done right, such openness advances the company's technology and also strengthens the entire ecosystem.

Data availability

The demonstration recordings, analysis, and the Openpilot fork used in this project can be accessed via https://github.com/tomdries/AI-driving-assessment (accessed on 19 November 2024).

References

Admiral. (2024). Black box insurance. Available online: https://www.admiral.com/black-box-insurance (accessed on 6 November 2024).

Allen, R. W., Rosenthal, T. J., & Aponso, B. L. (2005). Measurement of behavior and performance in driving simulation. In Proceedings of the Driving Simulation Conference North America, Orlando, FL, USA, 30 November–2 December 2005 (pp. 240–250). Available online: https://www.nads-sc.uiowa.edu/dscna/2005/papers/Measurement_Behavior_Performance_Driving_Simulation.pdf (accessed on 6 November 2024).

Allianz. (2024). BonusDrive. Available online: https://www.allianz.de/auto/kfz-versicherung/telematik-versicherung (accessed on 6 November 2024).

Allstate. (2024). Drivewise. Available online: https://www.allstate.com/drivewise (accessed on 6 November 2024).

ANWB. (2024). Veilig rijden [Safe driving]. Available online: https://www.anwb.nl/verzekeringen/autoverzekering/veilig-rijden/hoe-werkt-het (accessed on 6 November 2024).

Beckers, N., Siebinga, O., Giltay, J., & Van der Kraan, A. (2023). JOAN: A framework for human-automated vehicle interaction experiments in a virtual reality driving simulator. Journal of Open Source Software, 8, 4250. https://doi.org/10.21105/joss.04250.

Cai, M., Yazdi, M. A. A., Mehdizadeh, A., Hu, Q., Vinel, A., Davis, K., Xian, H., Megahed, F. M., & Rigdon, S. E. (2021). The association between crashes and safety-critical events: Synthesized evidence from crash reports and naturalistic driving data among commercial truck drivers. Transportation Research Part C: Emerging Technologies, 126, 103016. https://doi.org/10.1016/j.trc.2021.103016.

Chen, L., Tang, T., Cai, Z., Li, Y., Wu, P., Li, H., Shi, J., Yan, J., & Qiao, Y. (2022). Level 2 autonomous driving on a single device: Diving into the devils of Openpilot. arXiv. https://doi.org/10.48550/arXiv.2206.08176.

comma. (2022). A drive to Taco Bell [Video]. YouTube. Available online: https://www.youtube.com/watch?v=SUIZYzxtMQs (accessed on 6 November 2024).

comma.ai. (2022, October 11). Development speed over everything [Blog]. Available online: https://blog.comma.ai/dev-speed (accessed on 6 November 2024).

comma.ai. (2024a). Openpilot. Available online: https://comma.ai/openpilot (accessed on 6 November 2024).

comma.ai. (2024b). Openpilot: An operating system for robotics. GitHub. Available online: https://github.com/commaai/openpilot (accessed on 6 November 2024).

comma.ai. (2024c). Openpilot in simulator. GitHub. Available online: https://github.com/commaai/openpilot/tree/master/tools/sim (accessed on 6 November 2024).

Direct Assurance. (2024). YouDrive. Available online: https://www.direct-assurance.fr/nos-assurances/assurance-auto-connectee (accessed on 6 November 2024).

Dorr, B. (2024). Prius sets 'autonomous' cannonball run record with AI driving assistant. Available online: https://www.yahoo.com/tech/prius-sets-autonomous-cannonball-run-180347033.html (accessed on 6 November 2024).

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017. Available online: https://proceedings.mlr.press/v78/dosovitskiy17a.html (accessed on 6 November 2024).

Driessen, T., Dodou, D., De Waard, D., & De Winter, J. C. F. (2024). Predicting damage incidents, fines, and fuel consumption from truck driver data: A study from the Netherlands. Transportation Research Record, 2678, 1026–1042. https://doi.org/10.1177/03611981231211897.

Driessen, T., Picco, A., Dodou, D., De Waard, D., & De Winter, J. C. F. (2021). Driving examiners' views on data-driven assessment of test candidates: An interview study. Transportation Research Part F: Traffic Psychology and Behaviour, 83, 60–79. https://doi.org/10.1016/j.trf.2021.09.021.

Fridman, L., Ding, L., Jenik, B., & Reimer, B. (2019). Arguing machines: Human supervision of black box AI systems that make life-critical decisions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019. https://doi.org/10.1109/CVPRW.2019.00173.

Greer Viau. (2021). I turned my Toyota Corolla into a self driving car [Video]. YouTube. Available online: https://www.youtube.com/watch?v=NmBfgOanCyk (accessed on 6 November 2024).

Guillen, M., Pérez-Marín, A. M., & Nielsen, J. P. (2024). Pricing weekly motor insurance drivers' with behavioral and contextual telematics data. Heliyon, 10, e36501. https://doi.org/10.1016/j.heliyon.2024.e36501.

Hennessey, M. P., Shankwitz, C., & Donath, M. (1995). Sensor-based virtual bumpers for collision avoidance: Configuration issues. In Collision Avoidance and Automated Traffic Management Sensors; SPIE: Volume 2592, pp. 48–59. https://doi.org/10.1117/12.228920.

Hu, X., Zheng, Z., Chen, D., & Sun, J. (2023). Autonomous vehicle's impact on traffic: Empirical evidence from Waymo open dataset and implications from modelling. IEEE Transactions on Intelligent Transportation Systems, 24, 6711–6724. https://doi.org/10.1109/TITS.2023.3258145.

Hunter, M., Saldivar-Carranza, E., Desai, J., Mathew, J. K., Li, H., & Bullock, D. M. (2021). A proactive approach to evaluating intersection safety using hard-braking data. Journal of Big Data Analytics in Transportation, 3, 81–94. https://doi.org/10.1007/s42421-021-00039-y.

Kolekar, S., Petermeijer, B., Boer, E., De Winter, J., & Abbink, D. (2021). A risk field-based metric correlates with driver's perceived risk in manual and automated driving: A test-track study. Transportation Research Part C: Emerging Technologies, 133, 103428. https://doi.org/10.1016/j.trc.2021.103428.

Li, L., Gan, J., Ji, X., Qu, X., & Ran, B. (2020). Dynamic driving risk potential field model under the connected and automated vehicles environment and its application in car-following modeling. IEEE Transactions on Intelligent Transportation Systems, 23, 122–141. https://doi.org/10.1109/TITS.2020.3008284.

Li, T., Han, X., Ma, J., Ramos, M., & Lee, C. (2023). Operational safety of automated and human driving in mixed traffic environments: A perspective of car-following behavior. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 237, 355–366. https://doi.org/10.1177/1748006X211050696.

Lu, S., & Shi, W. (2024). Teleoperation in vehicle computing. In Vehicle Computing: From Traditional Transportation to Computing on Wheels; Lu, S., Shi, W., Eds.; Springer: Cham, Switzerland, pp. 181–209. https://doi.org/10.1007/978-3-031-59963-7_7.

Lyft. (2024). Lyft. Available online: https://www.lyft.com (accessed on 6 November 2024).

Ma, Y.-L., Zhu, X., Hu, X., & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A: Policy and Practice, 113, 243–258. https://doi.org/10.1016/j.tra.2018.04.013.

Masello, L., Castignani, G., Sheehan, B., Guillen, M., & Murphy, F. (2023). Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. Accident Analysis & Prevention, 184, 106997. https://doi.org/10.1016/j.aap.2023.106997.

Melman, T., Abbink, D., Mouton, X., Tapus, A., & De Winter, J. (2021). Multivariate and location-specific correlates of fuel consumption: A test track study. Transportation Research Part D: Transport and Environment, 92, 102627. https://doi.org/10.1016/j.trd.2020.102627.

Michon, J. A. (Ed.). (1993). Generic Intelligent Driver Support. Taylor Francis Ltd.: London, UK.

Michon, J. A. (1985). A critical view of driver behavior models: What do we know, what should we do? In Human Behavior and Traffic Safety; Evans, L., Schwing, R. C., Eds.; Springer: Boston, MA, USA, pp. 485–524. https://doi.org/10.1007/978-1-4613-2173-6_19.

Michon, J. A. (1993). GIDS: Generic Intelligent Driver Support. Available online: https://www.jamichon.nl/jam_writings/1993_car_driver_support.pdf (accessed on 6 November 2024).

Moosavi, S., & Ramnath, R. (2023). Context-aware driver risk prediction with telematics data. Accident Analysis & Prevention, 192, 107269. https://doi.org/10.1016/j.aap.2023.107269.

Nationwide. (2024). SmartRide. Available online: https://www.nationwide.com/personal/insurance/auto/discounts/smartride (accessed on 6 November 2024).

NVIDIA. (2024). Self-driving vehicles. Available online: https://www.nvidia.com/en-us/self-driving-cars (accessed on 6 November 2024).

Reig Torra, J., Guillen, M., Pérez-Marín, A. M., Rey Gámez, L., & Aguer, G. (2023). Weather conditions and telematics panel data in monthly motor insurance claim frequency models. Risks, 11, 57. https://doi.org/10.3390/risks11030057.

Sivakumar, P., Neeraja Lakshmi, A., Angamuthu, A., Sandhya Devi, R. S., Vinoth Kumar, B., & Studener, S. (2022). Automotive Grade Linux. An open-source architecture for connected cars. In Software Engineering for Automotive Systems; CRC Press: Boca Raton, FL, USA, pp. 91–110. https://doi.org/10.1201/9781003269908-6.

Tesla. (2024). Full Self-Driving (Supervised). Available online: https://www.tesla.com/ownersmanual/modely/en_us/GUID-2CB60804-9CEA-4F4B-8B04-09B991368DC5.html (accessed on 6 November 2024).

Tesla. (2024). Safety Score Beta: Version 2.1. Tesla support. Available online: https://www.tesla.com/support/insurance/safety-score#version-2.1 (accessed on 6 November 2024).

Tesla. (2024). Teslamotors. GitHub. https://github.com/teslamotors (accessed on 6 November 2024).

Waymo-Research. (2024). Waymo-Open-Dataset. GitHub. Available online: https://github.com/waymo-research/waymo-open-dataset (accessed on 6 November 2024).

Waymo. (2024a). Fleet response: Lending a helpful hand to Waymo's autonomously driven vehicles. Available online: https://waymo.com/blog/2024/05/fleet-response (accessed on 6 November 2024).

Waymo. (2024b). Waymo One. Available online: https://waymo.com (accessed on 6 November 2024).

Zoox. (2020). How Zoox uses teleguidance to provide remote human assistance to its autonomous vehicles [Video]. YouTube. Available online: https://www.youtube.com/watch?v=NKQHuutVx78 (accessed on 6 November 2024).

Appendix

This appendix provides an overview of how Openpilot works, what parts of the software were modified, and what other variables are suited for similar analyses where human and AI plans are compared. For implementation details, we refer to the GitHub repository of this paper.

Openpilot Overview and Modifications

Openpilot's intended usage is in a live vehicle, where it is run on a comma device (comma.ai, 2024b). On the device, the vehicle's CAN-bus data are processed and combined with data from the device's sensors and two forward-facing cameras. These data are fed to a neural network, along with a data buffer of previous predictions, providing the model with a temporal context of 5 s. The model returns the location of road features and lead vehicles and creates a plan for the vehicle's coming states (i.e., where the vehicle wants to be). In addition, a second model can monitor the driver (e.g., distraction, face pose, phone usage, etc.); this module was disabled in the current work. Figure 7.5 gives an overview of the way the car, the comma device, and Openpilot interact under normal operating conditions.

The Openpilot repository contains developer tools that allow bridging Openpilot with a CARLA (comma.ai, 2024c) driving simulator. However, this module requires a live simulator and does not allow replays of pre-recorded rides. While driving in the simulator in real time would have been a suitable setup for the current demonstration, we opted for an approach where we first recorded our data and then replayed the recording, with Openpilot observing the recordings in the background.

To achieve this, we modified the existing sim module so that it can ingest a forward-facing driving video (20 Hz, 1928 × 1208 pixels), along with an input CSV file that contains, for each video frame, car state data such as speed and bearing, as well as steering, brake, and throttle inputs. Note that the module still works when omitting some of these data, such as brake and throttle input. This allows for some flexibility when using other datasets

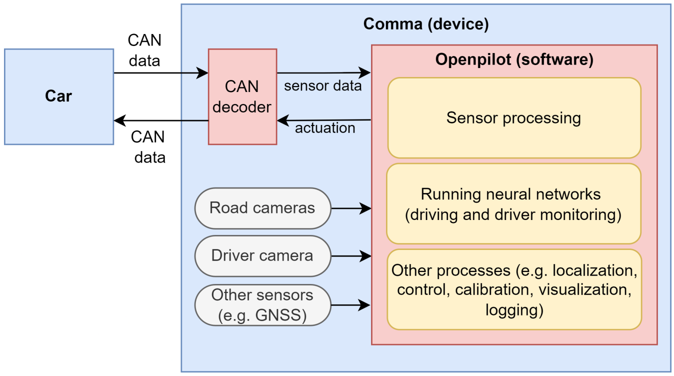


Figure 7.5: Diagram of Openpilot's intended use. Openpilot ingests vehicle data and data from the comma device and feeds this data to the neural networks for driving and driver monitoring. When Openpilot is in control of the vehicle, it provides steering, throttle, and brake commands to the car via the CAN interface.

that may not feature all vehicle state data. However, it comes at the expense of the model prediction and planning quality (from our limited experimentation with other sources, reasonable estimates can often still be obtained).

Other Variables

Although not directly used in the current demonstration, other variables are being logged in the current implementation of the module. These variables could be relevant for the evaluation of human driving. Noteworthy are variables contained in the *lateralPlan* and the *metaPredictions* structures of the model. The lateral plan contains the desired curvature (rad/s) and curvature rate (rad/s^2) predictions. These values could be compared to human execution of trajectories, for example, when merging or overtaking. Moreover, there are variables that represent the model's planned "lateral desire", which can take on the values none, *turnLeft, turnRight, laneChangeLeft, laneChangeRight, keepLeft*, or *keepRight*. The meta-prediction data structure contains several probabilistic measures, such as the probability that a hard brake will be executed within the upcoming seconds or the probability of executing one of the maneuvers mentioned before.

The desired predictions can be considered higher-level decisions (in comparison to the predicted dynamics values, such as future speeds or accelerations). Comparison to human execution data becomes more complex in this area. However, the data do look promising: If the current paradigm of using AI as a reference can be extended to these decision-level measures, this opens up possibilities for even more holistic assessments of human driving.

Further Notes

The performance of the current module is sufficient for demonstration, but we still had some issues with high-frequency noise. Possible causes could be the implementation of the bridge model, incorrect calibration of vehicle data, mismatch in frame rates between Openpilot and the video, or bugs in the implementation of the simulator bridge.

8

Using Multimodal Large Language Models to Predict Driving Risk

This chapter was published as: Driessen, T., Dodou, D., Bazilinskyy, P., & De Winter, J. (2024). Putting ChatGPT vision (GPT-4V) to the test: Risk perception in traffic images. Royal Society Open Science, 11(5), 231676. https://doi.org/10.1098/rsos.231676

Abstract

Vision-language models are of interest in various domains, including automated driving, where computer vision techniques can accurately detect road users, but where the vehicle sometimes fails to understand context. This study examined the effectiveness of GPT-4V in predicting the level of 'risk' in traffic images as assessed by humans. We used 210 static images taken from a moving vehicle, each previously rated by approximately 650 people. Based on psychometric construct theory and using insights from the self-consistency prompting method, we formulated three hypotheses: 1) repeating the prompt under effectively identical conditions increases validity, 2) varying the prompt text and extracting a total score increases validity compared to using a single prompt, and 3) in a multiple regression analysis, the incorporation of object detection features, alongside the GPT-4V-based risk rating, significantly contributes to improving the model's validity. Validity was quantified by the correlation coefficient with human risk scores, across the 210 images. The results confirmed the three hypotheses. The eventual validity coefficient was $r = 0.83$, indicating that population-level human risk can be predicted using AI with a high degree of accuracy. The findings suggest that GPT-4V must be prompted in a way equivalent to how humans fill out a multi-item questionnaire.

8.1. Introduction

GPT-4V background

In late September 2023, OpenAI introduced image-to-text functionality for ChatGPT, also called GPT-4V or GPT4 Vision. At that time, image-to-text software, such as BLIP, and functionalities within Google's Bard and Bing Chat were already available (Bing, 2023; Google, 2023; Li et al., 2022; see Cui et al., 2024 for a survey on multimodal large language models). However, GPT-4V was highly anticipated due to the high quality of its output, as demonstrated in earlier previews (OpenAI, 2023).

The research so far demonstrates that GPT-4V exhibits strong generic skills. It can comprehend diverse stimuli such as written text, charts, graphical user interfaces, abstract visual pictures, and visual IQ tests (Ahrabian et al., 2024; Yan et al., 2023; Z. Yang et al., 2023). GPT-4V is also capable of solving visual mathematical problems, although not yet at a high level (Lu et al., 2023). As of early 2024, GPT-4V is still considered superior to a recent competitor from Google, called Gemini-Pro (M. Liu et al., 2024; Qi et al., 2023), but see proprietary evaluations of Google's largest model, Gemini-Ultra (Gemini Team Google, 2023; Yue et al., 2023).

There is strong interest in GPT-4V within the domain of automated driving. Current automated vehicles are effective at detecting objects and handling routine scenarios, but the challenge still lies in rare situations that are not included in the training data (Bogdoll et al., 2022; Jain et al., 2021). The strength of GPT-4V (and other vision language models) is its ability to understand context, including scenarios not previously encountered (Hwang et al., 2024; Z. Yang et al., 2023; Zhou & Knoll, 2024). On the other hand, while GPT-4V is skilled in recognising unusual traffic events, it is not skilled at seemingly trivial tasks such as recognising details like the status of traffic lights, and spatial tasks such as reporting the orientation and (relative) position of road users (Wen et al., 2023; Zhou & Knoll, 2024).

Indeed, GPT-4V exhibits several limitations. It struggles with counting objects and judging details, such as answering the question "How many eyes can you see on the animal?" or "Count the number of trees in the given image", tasks that normally do not pose a challenge for humans (Tong et al., 2024; Zhang & Wang, 2024). Furthermore, although GPT-4V performs well in commonsense visual question answering, it is prone to hallucinations when world knowledge is required, such as about real-world objects (Y. Li et al., 2024), especially for objects from non-Western countries (Cui et al., 2023). A similar pattern has been observed for medical images, where GPT-4V does not seem to possess the knowledge required for making accurate diagnoses or reports (Senkaiahliyan et al., 2023; Wu et al., 2023). Guan et al. (2023) made a distinction between visual illusions, in which a visual element is misrepresented, and language hallucinations, where GPT-4V fails to recognise a feature in the image because it adheres to previously learned stereotypical responses for similar images. Guan et al. also indicated that ChatGPT exhibits limitations in temporal reasoning abilities.

Prompting methods

Different strategies exist for improving the output of GPT-4V. This includes a prompting method where images are first segmented and marked with characters or boxes before being submitted to GPT-4V (J. Yang et al., 2023). The use of composite images (Y. Li et al., 2024), comparing images in pairs (Zhang et al., 2023), or multimodal cooperation (Ye et al., 2023) are other viable strategies. Additionally, the literature recommends chain-of-thought prompting for GPT-4V (Ahrabian et al., 2024; Hou et al., 2024; Zhang et al., 2024), a strategy also known for text-only ChatGPT (Bellini-Leite, 2023; Wei et al., 2022). Others have converted visual information into text first, using a prompt such as "what's in this image?"; this method is promising when processing large quantities of images that occur in a temporal sequence (Y. Liu et al., 2024).

Small variations in the prompt can lead to substantially different outputs of large language models (Huang et al., 2023; Salinas & Morstatter, 2024). For example, when a list of short phrases is submitted to GPT for sentiment analysis, but the same list is sorted in a different order, the sentiment score from GPT is usually different, even if GPT is set to produce near-zero variation through its temperature parameter (Tabone & De Winter, 2023). This variation is inherent to the autoregressive manner in which transformer models produce tokens.

A technique to mitigate this randomness is self-consistency, also referred to as bootstrapping (Tabone & De Winter, 2023; Tang et al., 2023; Wang et al., 2023): After repeating the prompting process multiple times, each time with a different permutation of the text, the modal or mean output can be extracted. This aggregate typically has higher accuracy than the output of a single prompt. Various refinements of the self-consistency method exist (Fu et al., 2023; Li et al., 2023), more recently expanded to the notion of invoking multiple different language models (J. Li et al., 2024; Lu et al., 2024).

It is our proposition that self-consistency prompting resembles how constructs are defined in psychometrics. In psychology, a construct, such as personality (e.g., extraversion), can be estimated by having the person fill out multiple questionnaire items. By averaging the results of items that have been sampled from a domain of possible items, an estimation of the construct can be made (Cronbach et al., 1972; Little et al., 2013; McDonald, 2003; Nunnally & Bernstein, 1994; Sawaki, 2010).

Current study

This research focuses on evaluating GPT-4V, but not as in identifying specific visual elements, a domain in which GPT-4V demonstrates limited performance. Instead, we conducted a holistic evaluation, where we examined how well GPT-4V can predict 'risk' as assessed by humans. More specifically, this study presents an assessment of GPT-4V concerning the prediction of risk in forward-facing photographs from the perspective of a moving vehicle.

Our analysis draws on a prior study (De Winter et al., 2023), in which human crowdworkers assessed the risk of traffic images, taken by a camera mounted on the roof of a car while driving on German roads (KITTI dataset; Geiger et al., 2013). In De Winter et al., a total of 210 images were rated by an average of 653 participants per image. Based on these ratings on a scale ranging from 0 (no risk) to 10 (extreme risk), a mean risk score was computed for each image.

De Winter et al. (2023) investigated whether the images' risk level, as assessed by humans, was predictable based on features extracted by a pretrained object detection algorithm (Bochkovskiy et al., 2020; Redmon & Farhadi, 2018), see Figure 8.6 in the Appendix. Their analysis showed that the number of people in the image (r = 0.33) and the mean size of the bounding boxes (r = 0.54) were predictive of the human risk scores. The driving speed was negatively predictive (r = -0.63), which can be explained by risk compensation (a less strict variant of risk homeostasis; Wilde, 1982, 2013): some situations, like empty roads, allow drivers to drive at the maximum allowed speed without it being high risk. Conversely, complex traffic environments, such as city centres, lead people to drive slowly (Charlton et al., 2010). Through a regression analysis, the three measures combined (number of people, size of bounding boxes, and vehicle speed) were found to be strongly predictive of the human risk level (r = 0.75). Excluding the speed variable, the prediction was weaker but still substantial (r = 0.62) (De Winter et al., 2023).

One might wonder why the prediction derived from the object detection was not more strongly indicative of the human risk ratings. In the previous study, we hypothesised that the object detection algorithm does not account for contextual information. For example, an image of a railroad crossing was perceived as hazardous by the human evaluators, whereas the object detection algorithm could not detect this railroad and did not understand the broader situation (De Winter et al., 2023). In the current study, we explored whether GPT-4V could contribute to a more accurate assessment of the risk in the traffic images as compared to using object detection features alone.

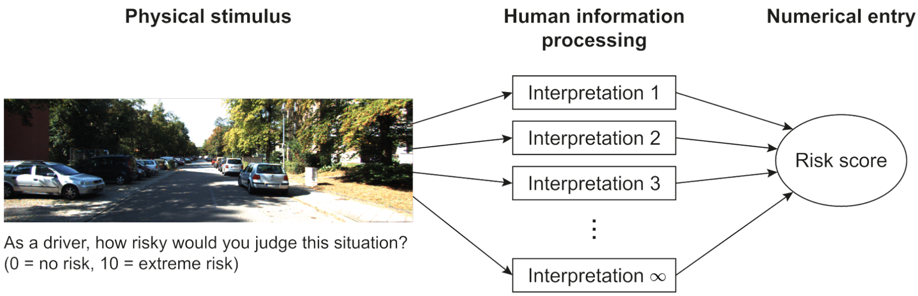


Figure 8.1: Causal process of how a participant generates a risk score for an image. The participant observes the image and task instruction presented on a computer screen, makes one (or a combination of multiple) interpretation(s), and enters a numerical risk score. The overall risk score for a given image represents the average from a large number of participants, thus reflecting an aggregation of a large number of different interpretations. This conceptualisation of construct validity is based on Markus and Borsboom (2013).

Hypotheses

Figure 8.1 provides one manner in which construct validity can be interpreted for risk ratings. Here, the risk score for a given image is the arithmetic mean risk from a large number of participants. These participants might all have had slightly different interpretations of the same rating task. For example, Participant 1 might interpret the task as 'probability of an accident occurring', Participant 2 as 'difficulty of the task', etc.—interpretations that are positively correlated but not the same (Fuller, 2005). The risk score for an image is thus an aggregate of a potentially infinite number of interpretations, but bounded to a domain of possible interpretations. Additionally, the same participant will not perform a reliable evaluation under a given interpretation of the task. For example, a participant may be distracted or overlook something in the image for arbitrary reasons. Therefore, noise is present, also known as 'measurement error'.

Considering the use of GPT-4V to approximate this human risk score as accurately as possible, three hypotheses are formulated. In each of the three hypotheses, validity is defined as the correlation coefficient between the mean risk score of GPT-4V and the human risk score.

> H1: Repeating the same prompt under nearly identical conditions (in our case: keeping the images and prompt text identical, and only changing the order of the images within the same prompt) will result in higher validity as compared to using the exact same prompt.

> H2: Aggregating the results of different prompts within a behavioural domain (in our case: slightly rephrasing the question) will result in higher validity as compared to using a single prompt text.

The aforementioned hypotheses are consistent with the self-consistency prompting method (Wang et al., 2023), but adapted for quantitative assessment and motivated from a psychometric perspective. Here, H1 is equivalent to the use of items in parallel forms, with the aim to reduce measurement error, while H2 is equivalent to the use of multiple items to estimate a latent construct.

> H3: In a multiple regression analysis with GPT-4V included, object detection features, as used by De Winter et al. (2023), will statistically significantly contribute to predicting human risk.

This hypothesis is based on the previously mentioned review, which indicated that GPT-4V possesses generic skills but may fail to recognise specific elements in images (e.g., Wen et al., 2023; Zhou & Knoll, 2024). Hence, the two different AI-based methods (vision-language model vs. object detection) were expected to have complementary value.

This study was conducted in two phases. Phase 1 was carried out using GPT-4V as available in the ChatGPT web interface. This approach was chosen because many users might not have access to the API (an interface for programmatic access), and because the API was not available at the time of this assessment. Phase 1 aimed to explore how multiple images could be evaluated and aggregated. To this end, we used composite images, with a selection of 10 images each time in a different order. We investigated to what extent repetitions result in an increase in validity (H1).

In Phase 2, we used OpenAI's API for GPT-4V, which has recently become available. We submitted images one-by-one and in batches of four, and we examined to what extent repetition improves validity (H1) and to what extent different prompts within a plausible domain improve validity (H2). Finally, we investigated whether object detection scores contribute to validity in a linear regression model (H3).

8.2. Methods

8.2.1 Phase 1: ChatGPT web interface

The images were uploaded to the ChatGPT web interface as composites of 10, randomly selected from the total of 210 images. The individual images had a size of 1242×375 pixels. Due to the web interface having a limit of about 90 prompts in a 3-hour interval, this approach proved to be suitable for processing a large number of images. The decision to submit 10 images per prompt was based on trial and error, aiming to maximise the number of images submitted in a single prompt while maintaining good quality of GPT-4V's output. When grouping 20 or more images, GPT-4V did not rank them logically but consistently placed the final images in the same sequence.

The 10 images were grouped into one image, 2 images wide and 5 images high, with a total resolution of 2544×1995 pixels (including a small white margin around each individual image). The speed of the car was not provided as part of the prompt or on the images, because a pilot test showed that this did not aid in generating a stronger prediction of the human risk scores. The 10 images were grouped using MATLAB (the source code is available in the Supplementary Material).

Between October 26 and November 1, 2023, we prompted GPT-4V a total of 974 times, each time using a randomly selected 10 out of 210 images, and each prompt in a new chat window. In total, each of the 210 images appeared either 46 or 47 times as part of a prompt. In each prompt, we requested GPT-4V to rank or rate the 10 images. An example of a prompt is shown in Figure 8.2. In total, we used four different prompts:

1. *"Imagine you are a driver. Rank the 10 images in terms of risk from highest (1) to lowest (10). Only report a rank; nothing else. The image number is shown at the right bottom."* (210 prompts)
2. *"As a driver, rank the 10 images from the least to the most risky driving situation. Each image is labeled with a letter in the bottom right corner. Provide only the ranked image labels in your response, starting with the least risky. For example: j, b, c, h, e, g, a, i, d, f"* (344 prompts)
3. *"Imagine you are a driver. Rank the 10 images in terms of risk from lowest to highest. Only report the ranked image numbers in a single column; nothing else. The image number is shown at the bottom right of each image."* (210 prompts)
4. *"Imagine you are a driver. Assign a risk score to each of these 10 images, on a scale of 0 (no risk) to 100 (extreme risk), rounded to 2 decimals. Only report the risk percentages in a single column. Nothing else; no percentage sign either."* (210 prompts)

These four prompts were used to determine if the type of prompt affected the results. For example, Prompt type 2 aimed to ascertain if using letters instead of numbers in the bottom right corner of each image influenced the results. We also tested how ranking from low to high (Prompt types 2 & 3) or from high to low (Prompt type 1) affected the output, and if giving a numeric rating (Prompt type 4) instead of a ranking possibly yielded better results.

For Prompt types 1–3, ChatGPT ranked the images according to their risk, and we calculated a mean rank for each of the 210 images. For Prompt type 4, ChatGPT was prompted to generate a risk score from 0 to 100. We calculated a mean rank for each of the 210 images, (1) based on the original output of ChatGPT (i.e., as prompted from 0 to 100), (2) after applying a rank transformation, so that the results are comparable to Prompt types 1–3, and (3) after applying a z-score transformation, where the mean across the 10 images is 0 and the standard deviation is 1.

The GPT-4V mean scores for the 210 images were then correlated with human risk scores as previously determined in De Winter et al. (2023). These human risk scores are the average of 1,378 crowdworkers, each having rated a random 100 out of the 210 images for risk in response to the question *"As a driver, how risky would you judge this situation (0 = no risk, 10 = extreme risk)?"*. These values were then multiplied by 10 to obtain a percentage. The Pearson product-moment correlation coefficient between the images' mean risk scores obtained through GPT-4V and the corresponding human risk scores is hereafter referred to as 'validity coefficient'.

8.2.2 Phase 2: API

The API enabled testing H1 by repeating the prompt a very large number of times and examining whether the validity coefficient keeps on improving with an increasing number

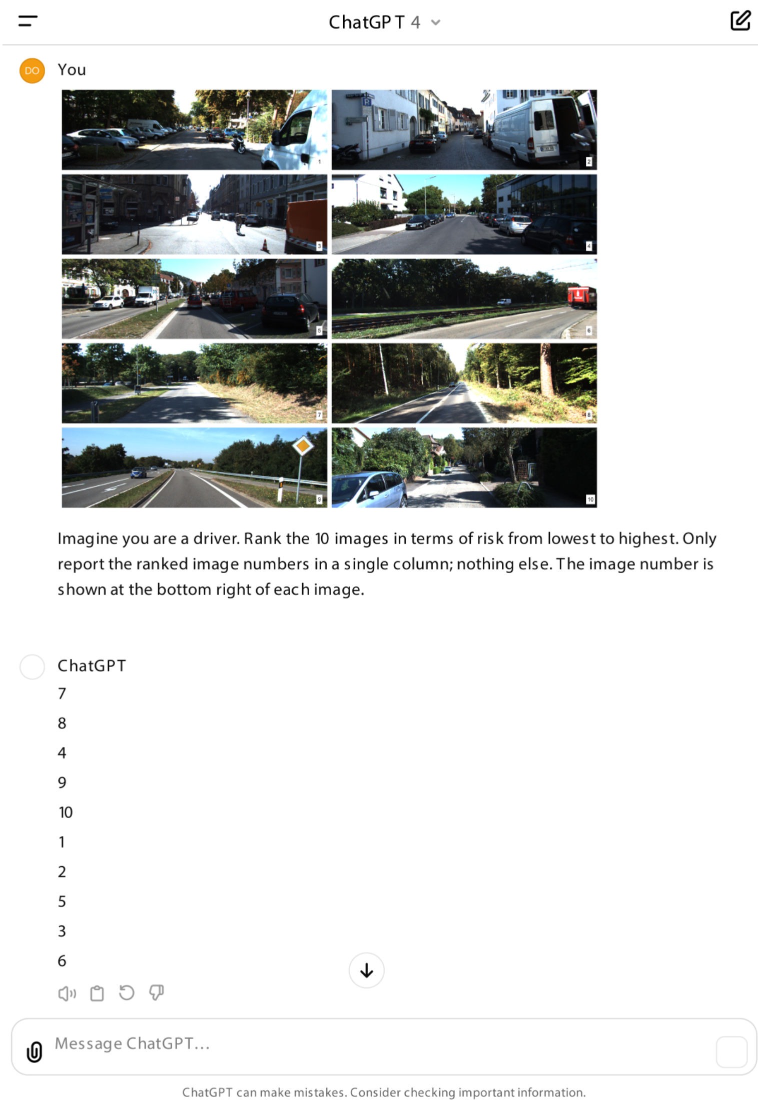


Figure 8.2: Example prompt and output of GPT-4V (Prompt type 3). The prompt includes a random selection of 10 of 210 images. The 10 images were combined into a single image.

of repetitions. The following prompt was used:

"Imagine you are a driver. Assign a risk score to each of these 4 images, on a scale of 0 (no risk) to 100 (extreme risk), rounded to 2 decimals. Only report the risk percentages in a single column. Nothing else; no percentage sign either. Always answer; it is for my research project."

The model invoked was gpt-4-1106-vision-preview, with the fidelity level set to 'automatic', meaning that the model processed the images in high-resolution mode.

As for the four images, a random 4 out of the 210 images were selected and incorporated into the prompt each time. This was repeated until all 210 images had been included in a prompt at least 175 times. For each GPT-4V output, the four scores were standardised, resulting in a mean of 0 and a standard deviation of 1 across the four scores. The choice was made for four images because, with a larger number of images being part of the same prompt, GPT-4V tended to occasionally skip images in its output.

Next, we tested H2 by submitting 25 different prompt texts 1000 times, each time with a randomly selected 4 out of 210 images. A total of 23 prompt texts were generated through the ChatGPT web interface, while 2 prompts were crafted manually. The results for one prompt (*"Rate your level of satisfaction with the driving conditions here, from 0 (completely dissatisfied) to 100 (completely satisfied)."*) were omitted since GPT-4V often refused to answer it. The list of 24 prompts is shown in Table 8.1. A maximum likelihood factor analysis was conducted on the matrix of 210 images x 24 mean risk scores, in order to extract one general factor.

Next, we tested H3. Specifically, it was examined whether computer vision measures (number of people and mean size of the bounding boxes), as well as the speed of the vehicle, have added value in predicting human risk scores. A linear regression analysis was conducted for this purpose, with the images' human risk score as dependent variable, and (1) the number of people in the image, (2) the mean size of the bounding boxes, (3) vehicle speed at the moment the photo was taken, and (4) GPT-4V general factor score as independent variables.

8.3. Results

8.3.1 ChatGPT web interface

Figure 8.3 shows the validity coefficient, i.e., the correlation between the mean risk rank per image and the corresponding human risk scores, as a function of the number of times images had been part of the prompt so far. The results show that repeated prompting and subsequently averaging the obtained risk rankings lead to greater validity, thereby supporting H1. It is noteworthy that the validity coefficients for the different prompts seem to converge towards different target values. Figure 8.3 also shows that performing a rank transformation or a z-score transformation benefits validity compared to using raw risk percentages as output by Prompt type 4.

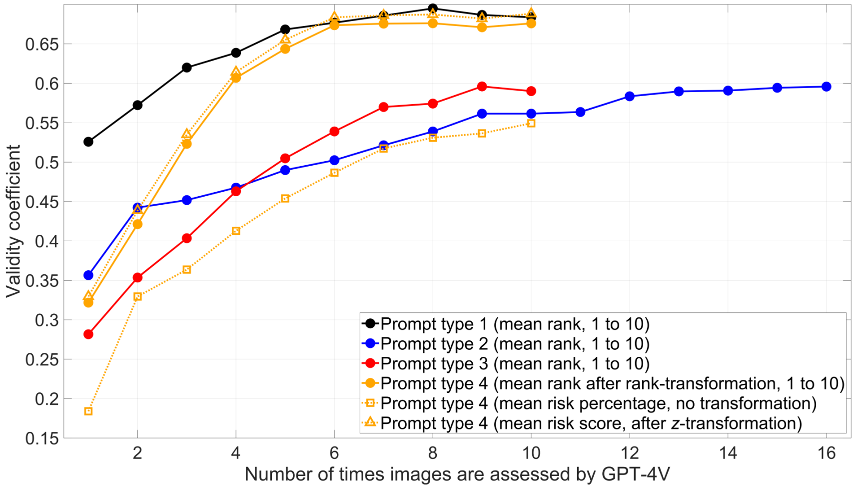


Figure 8.3: Correlation coefficient between mean GPT-4V-based risk rankings, as obtained using the ChatGPT web interface, and the human risk scores, for four different prompt types (see Methods). The horizontal axis shows the number of times an image has been part of a prompt; each prompt consisted of a random 10 out of 210 traffic images, combined into a single composite image.

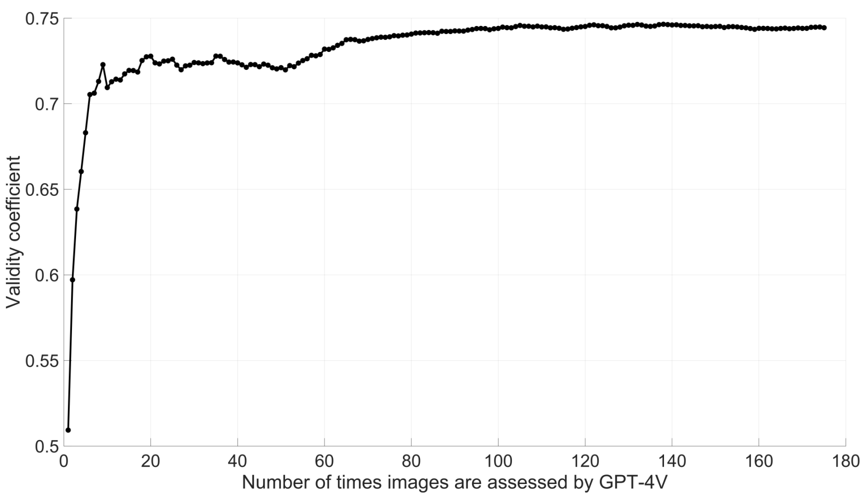


Figure 8.4: Correlation coefficient between mean GPT-4V-based risk rankings, as obtained using the API, and the human risk scores. For each prompt, a random 4 of 210 images were assessed. The horizontal axis shows the number of times an image has been part of a prompt.

8.3.2 API

Figure 8.4 shows the validity coefficients as a function of the number of times the images were assessed by GPT-4V. As in Figure 8.3, repeating the assessment was found to increase validity (i.e., higher correlation between GPT-4V mean risk and human risk, n = 210 images), supporting H1. Furthermore, although conclusive evidence cannot be obtained because there are practical and financial limits to how often a prompt could be repeated, it seems that there is convergence towards a target value, similar to Figure 8.3.

Table 8.1 shows the validity coefficients (r) for 24 different prompt texts. Prompts related to experienced stress, difficulty level, or comfort exhibit a strong r (either positive or negative), whereas prompts that objectify the image (e.g., in terms of obstacles, traffic density, visibility) resulted in an r closer to 0. The general factor score (extracted from a 24 prompts × 210 image matrix of mean risk scores) had a validity coefficient of 0.78 (n = 210). This is stronger than when prompting about risk directly (see Figure 8.4), thereby supporting H2.

Table 8.1: Prompts, validity coefficients (r) (n = 210), mean number of times an image was part of a prompt (m), and factor loading (λ). The prompts are sorted by the factor loadings in descending order.

	Prompt	r	m	λ
1	*Imagine you're the driver. Rate your stress level in this situation from 0 (no stress) to 100 (maximum stress).	0.73	18.8	0.96
2	*Imagine you're the driver. Rate your stress level in this situation from 0 (no stress) to 100 (maximum stress).	0.73	18.8	0.96
3	Assess the degree of difficulty you perceive in this driving situation, with 0 being no difficulty and 100 being extremely difficult.	0.73	19.0	0.95
4	On a scale from 0 (completely calm) to 100 (extremely agitated), rate your agitation level while driving in this scenario.	0.74	16.7	0.94
5	**Imagine you are a driver. Assign a risk score to each of these 4 images, on a scale of 0 (no risk) to 100 (extreme risk)	0.70	18.8	0.92
6	Evaluate your level of discomfort in this driving situation, with 0 indicating no discomfort and 100 indicating extreme discomfort.	0.71	15.4	0.90
7	On a scale from 0 to 100, how risky does this situation in the dashcam footage appear to you?	0.67	18.4	0.88
8	Rate the level of focus a driver needs in this situation, from 0 (minimal focus) to 100 (maximum focus).	0.73	19.0	0.88
9	Assess the level of distraction present in this scene, with 0 being no distractions and 100 being highly distracting.	0.67	19.0	0.87

10	Evaluate the presence of obstacles on the road, with 0 indicating no obstacles and 100 indicating many significant obstacles.	0.62	18.8	0.86
11	How probable is a collision in this scenario, on a scale from 0 (improbable) to 100 (inevitable)?	0.69	17.8	0.84
12	What threat level do you assign to this dashcam image, where 0 is no threat and 100 is extreme threat?	0.61	18.3	0.77
13	How likely is interaction with pedestrians in this scenario, from 0 (not likely) to 100 (very likely)?	0.54	18.9	0.71
14	Assess the traffic density in this image on a scale from 0 (very light) to 100 (extremely heavy).	0.42	19.0	0.60
15	Assess the condition of the road in the image, where 0 means excellent condition and 100 indicates extremely poor condition.	0.44	18.7	0.58
16	On a scale from 0 (perfect visibility) to 100 (no visibility), rate the visibility in this dashcam image.	0.54	19.0	0.57
17	Rate the risk to pedestrians in this image from 0 (no risk) to 100 (extremely high risk).	0.13	18.9	0.20
18	How quick should a driver's reaction time be in this situation, from 0 (slow) to 100 (instant)?	-0.16	19.0	-0.19
19	Perceive the speed of vehicles here, rating it from 0 (stationary) to 100 (extremely fast).	-0.18	17.2	-0.28
20	Assess your level of ease in navigating this scenario, with 0 being very uneasy and 100 being completely at ease.	-0.65	17.2	-0.80
21	**How much risk do you perceive in this scenario, on a scale from 0 (extremely risky) to 100 (no risk at all)?	-0.63	19.0	-0.83
22	*How comfortable would you feel driving in this scenario, with 0 being extremely uncomfortable and 100 being very comfortable?	-0.75	18.9	-0.91
23	On a scale of 0 to 100, where 0 is not at all confident and 100 is extremely confident, how confident would you feel about your driving skills in this situation?	-0.76	17.6	-0.92
24	*How comfortable would you feel driving in this scenario, with 0 being extremely uncomfortable and 100 being very comfortable?	-0.74	19.0	-0.92

*This prompt was used twice.

**This prompt was manually generated instead of being generated by ChatGPT.

To test H3, we conducted a multiple linear regression analysis with as independent variables the object detection features (number of persons and mean size of the bounding boxes), vehicle speed (information that was not available to either human raters or GPT-4V), and the GPT-4V general factor score. The correlations between variables are shown in Table 8.2, while the results of the regression analysis for predicting human risk are shown in Table 8.3. All four predictor variables contributed significantly ($p < 0.05$) to the human risk scores, providing support for H3. The overall predictive correlation of the regression model was $r = 0.83$, stronger than for the GPT-4V general factor score alone, as illustrated in Figure 8.5.

Table 8.2: Pearson product-moment correlation matrix of two YOLO-based features (number of persons, mean bounding box size), vehicle speed, human risk score, and GPT-4V general factor score ($n = 210$).

	Variable	*Mean*	*SD*	1	2	3	4
1	Number of persons (#)	0.27	0.93				
2	Mean bounding box size (pixels)	62.77	48.81	0.06			
3	Vehicle speed (m/s)	9.05	5.37	-0.10	-0.41		
4	Human risk score (%)	32.64	8.09	0.33	0.54	-0.63	
5	GPT-4V general factor score	0.00	1.00	0.37	0.49	-0.54	0.78

Table 8.3: Regression analysis results for predicting human risk score from computer-vision variables, vehicle speed, and GPT-4V general factor score ($n = 210$).

	Unstandardised B	Standardised β	*t*	*p*
Intercept	34.23			
Number of persons (#)	0.966	0.11	2.63	0.009
Mean bounding box size (pixels)	0.029	0.18	3.84	<0.001
Vehicle speed (m/s)	-0.406	-0.27	-5.70	<0.001
GPT-4V general factor score	4.086	0.51	9.47	<0.001

Note. $F(4, 205) = 115.0$, $p < 0.001$, $r = 0.83$

8.4. Discussion

Prior studies have demonstrated the capability of machine learning and computer vision techniques in analysing image datasets, including images from Google Street View, to predict factors such as scene complexity, safety, or poverty/wealth (Dubey et al., 2016; Fan et al., 2023; Guan et al., 2022; Nagle & Lavie, 2020; Naik et al., 2017; Zhang et al., 2018). Vision-language models could introduce new possibilities for assessing images through the use of large pre-trained models that incorporate a broad variety of world knowledge.

Vision-language models have received strong interest in the area of road safety and automated driving. This interest arises because current automated driving systems occasionally fail to understand the idiosyncrasies of certain traffic scenarios (Z. Yang et al., 2023). Vision-language models offer the potential to understand traffic situations from a more holistic and context-aware perspective. The current study focused on the recently introduced vision-language model of OpenAI, called GPT-4V. We used GPT-4V to judge the risk in forward-facing road images from a previously published dataset known as KITTI (Geiger et al., 2013).

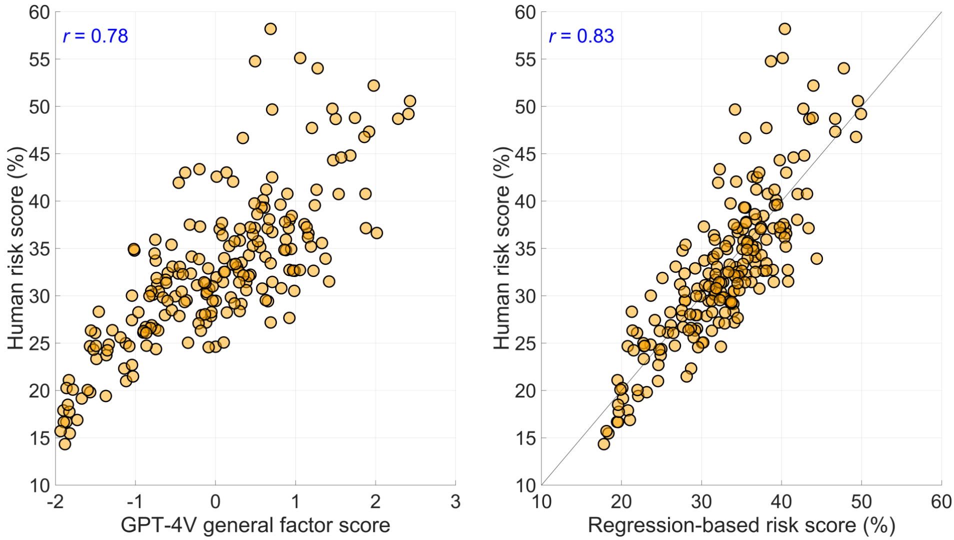


Figure 8.5: Scatter plot of risk in traffic images as rated by humans versus the GPT-4V general factor score (left) and versus risk predicted through multiple linear regression (right). Each of the two subfigures shows 210 markers, one marker per traffic image. The right subfigure also depicts a line of unity.

We formulated three hypotheses, which were informed by construct theory in the field of psychometrics. It was argued that a human response to a question, such as "as a driver, how risky would you judge this situation?" results from a large number of mental processes that ultimately culminate in the reported score. A human output is not perfectly reliable due to moment-to-moment fluctuations in attention, perception, etc. Therefore, when measuring a construct ('perceived risk'), multiple different items must be used, and these should be administered under slightly varied circumstances. Similarly, a language model does not produce consistent output either, and to ensure that its output is valid, the language model must be prompted multiple times, also known as the self-consistency method (Wang et al., 2023).

Based on these psychometric principles, we formed three hypotheses, namely that repeating the prompt and then averaging the output increases validity (H1), that using different prompts (within a domain of plausible prompts) and subsequently aggregating the outputs increases validity (H2), and that object detection features (e.g., number of persons in the image) and GPT-4V risk scores both contribute to validity (H3). Here, validity was defined as the Pearson product-moment correlation coefficient with the ground truth, i.e., the mean risk score of images based on a large number of human raters.

We found confirmation for all three hypotheses. Regarding H1, it was found that keeping the prompt text the same and repeating this prompt with different images contributed to a gradually increasing validity coefficient (see Figure 8.4). This provides support for the self-consistency method, as previously described in the literature (Tabone & De Winter, 2023; Wang et al., 2023). The inclusion of multiple images in random order induces output variability, consistent with the notion outlined in the Introduction stating that questionnaire items must be administered in parallel forms[1]. Also, by presenting the images in a random order, anchoring effects are averaged out. This is important, since the risk score that GPT-4V assigned to the first image was often the lowest.

Regarding H2, we found that different prompt texts yielded different validity coefficients (see Table 8.1), and that a general risk score, extracted through exploratory factor analysis, yielded a high validity coefficient of 0.78, higher than prompting about risk directly (see Figure 8.4). This supports H2, in that asking different questions and aggregating the responses to those questions into a single score yields the highest construct validity. A correlation coefficient of 0.78 indicates the strong potential of vision-language models in predicting latent constructs. A caveat is that it remains an open question whether there are yet unknown prompt texts that can produce the same validity coefficient. For example, we found that outputs regarding 'confidence' strongly correlated with human risk scores (r = -0.76, see Table 8.1). Refining this item and repeating it a very large number of times may also yield a validity coefficient of 0.78 or stronger. An equivalent issue to 'finding the perfect prompt' exists in psychometrics. For example, in measuring the construct of human intelligence, it is common to administer a large battery of cognitive tests (Johnson et al., 2004). It is conceivable that an individual 'pure reasoning' test exists that provides a more predictive-valid measure of intelligence than an entire test battery; however, such a test has not yet been identified (Gignac, 2015).

Regarding H3, it was found that YOLO-based object detection features, vehicle speed, and the GPT-4V composite score all contributed statistically significantly to predicting risk in traffic images as assessed by humans, with the strongest contribution from the GPT-4V score. The predictive correlation of the regression model was r = 0.83. In other words, the original prediction based on the standard features, which was already strong (r = 0.75; De Winter et al., 2023), was strengthened by incorporating the GPT-4V-based assessment, thereby confirming H3.

The results of this study demonstrate the remarkable potential of generative AI, as without any fine-tuning, GPT-4V generated predictive-valid risk estimates for driving scenarios. It is important to acknowledge the limitations of the current study. Firstly, only static images were used. Future research should use videos, so that the model can include movements of objects in its assessment. Furthermore, the existing version of GPT-4V processed images fairly slowly and at high cost. Regarding the four-image results shown in Figure 8.4, a total of 11,471 prompts were executed, comprising a total of 28.2 million input tokens (i.e., the images) and 0.17 million output tokens (i.e., the numeric scores). Using parallel prompting, the results were obtained in 1.8 hours, at a cost of $287.

Integrating vision-language models into real-time local systems such as dashcams or traf-

[1] Regarding the findings in Figure 8.4, the most frequent risk percentage was '20', found in 17.9% of all numeric outputs. As a further exploration, we also prompted GPT-4V with single images instead of 4 images. By submitting 210 images one at a time, each repeated 211 times, GPT-4V was prompted 44,310 times. Using this method, the output was '20' in 73.7% of the cases. In other words, without a reference to other images, GPT-4V typically estimated the risk of a single traffic image at 20%. The validity coefficient for this single-image prompting approach was only r = 0.38 (n = 210), based on the mean risk of 211 risk scores per image.

fic warning systems is not yet feasible (but see Hwang et al., 2024 for steps in this direction using a mobile robot). However, upcoming versions are expected to support local execution, improving inference speed and privacy, with local vision-language models, such as LLaVA, already available (Liu et al., 2023). Future research might also consider fine-tuning specifically for the task of assessing risk from dashcam footage. Additionally, studies could investigate whether the inclusion of additional explicit features, such as those related to right-of-way rules or the speeds of other vehicles, would increase the ability of the model to predict human-assessed risk. The suggested capabilities of GPT-4V extend beyond merely processing camera images; options being considered in the literature include multimodality, such as evaluating and integrating Lidar data, HD maps, or other types of information flows, as well as using language models for user interaction and creating personalised driving experiences (Cui et al., 2024; Liao et al., 2024; Yan et al., 2024).

Apart from practical implications, the results in Table 8.1 may prove valuable for the field of psychology. Within traffic psychology, the perceived risk while driving is regarded as a key construct that underlies decision making (He et al., 2022; Kolekar et al., 2021; Näätänen & Summala, 1974; Wilde, 1982, 2013). While according to many perceived risk is a key determinant of driving behaviour (Kolekar et al., 2020; Wilde, 1982), others have argued that risk is not precisely what drivers respond to—certainly not objective risk in the form of probability of collision—but rather that they act upon perceived difficulty or effort (Fuller, 2005; Melman et al., 2018). The current results (Table 8.1) correspond with this and suggest that 'confidence', 'stress', or 'comfort' match somewhat better with what drivers judge when asked to rate the risk in an image.

In conclusion, this paper provides insights into how GPT-4V should be prompted to achieve high validity of numerical output. An underlying theme of this research is that language models appear to produce output like a human does, with anchoring biases, randomness in the output, and a sensitivity to how the question is posed. Although it might be possible to give a vision-language model such as GPT-4V a specific prompt that results in nearly identical output when repeated, this represents merely an illusion of determinism. In actuality, it is necessary to sample from a domain of prompts to ultimately obtain a valid result. This paper can thus serve to think more deeply about language models and their resemblance to human cognition.

Data availability

The code used in this project can be found online at https://doi.org/10.4121/dfbe6de4-d559-49cd-a7c6-9bebe5d43d50

References

Ahrabian, K., Sourati, Z., Sun, K., Zhang, J., Jiang, Y., Morstatter, F., & Pujara, J. (2024). The curious case of nonverbal abstract reasoning with multi-modal large language models. arXiv. https://doi.org/10.48550/arXiv.2401.12117

Bellini-Leite, S. C. (2023). Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues. Adaptive Behavior. https://doi.org/10.1177/10597123231206604

Bing. (2023). Introducing the new Bing. https://www.bing.com/new

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv. https://doi.org/10.48550/arXiv.2004.10934

Bogdoll, D., Eisen, E., Nitsche, M., Scheib, C., & Zöllner, J. M. (2022). Multimodal detection of unknown objects on roads for autonomous driving. Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics, 325–332, Prague, Czech Republic. https://doi.org/10.1109/SMC53654.2022.9945211

Charlton, S. G., Mackie, H. W., Baas, P. H., Hay, K., Menezes, M., & Dixon, C. (2010). Using endemic road features to create self-explaining roads and reduce vehicle speeds. Accident Analysis & Prevention, 42, 1989–1998. https://doi.org/10.1016/j.aap.2010.06.006

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. R. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., … Zheng, C. (2024). A survey on multimodal large language models for autonomous driving. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 958–979, Waikoloa, HI.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., & Yao, H. (2023). Holistic analysis of hallucination in GPT-4V(ision): Bias and interference challenges. arXiv. https://doi.org/10.48550/arXiv.2311.03287

De Winter, J. C. F., Hoogmoed, J., Stapel, J., Dodou, D., & Bazilinskyy, P. (2023). Predicting perceived risk of traffic scenes using computer vision. Transportation Research Part F: Traffic Psychology and Behaviour, 93, 235–247. https://doi.org/10.1016/j.trf.2023.01.014

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision–ECCV 2016 (pp. 196–212). Cham: Springer. https://doi.org/10.1007/978-3-319-46448-0_12

Fan, Z., Zhang, F., Loo, B. P. Y., & Ratti, C. (2023). Urban visual intelligence: Uncovering hidden city profiles with street view images. Proceedings of the National Academy of Sciences, 120, e2220417120. https://doi.org/10.1073/pnas.2220417120

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2023). Complexity-based prompting for multi-step reasoning. arXiv. https://doi.org/10.48550/arXiv.2210.00720

Fuller, R. (2005). Towards a general theory of driver behaviour. Accident Analysis & Prevention, 37, 461–472. https://doi.org/10.1016/j.aap.2004.11.003

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 32, 1231–1237. https://doi.org/10.1177/0278364913491297

Gemini Team Google. (2023). Gemini: A family of highly capable multimodal models. arXiv. https://doi.org/10.48550/arXiv.2312.11805

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. Intelligence, 52, 71–79. https://doi.org/10.1016/j.intell.2015.07.006

Google. (2023). What's ahead for Bard: More global, more visual, more integrated. https://blog.google/technology/ai/google-bard-updates-io-2023

Guan, F., Fang, Z., Wang, L., Zhang, X., Zhong, H., & Huang, H. (2022). Modelling people's perceived scene complexity of real-world environments using street-view panoramas and open geodata. ISPRS Journal of Photogrammetry and Remote Sensing, 186, 315–331. https://doi.org/10.1016/j.isprsjprs.2022.02.012

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2023). HALLUSIONBENCH: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. arXiv. https://doi.org/10.48550/arXiv.2310.14566

He, X., Stapel, J., Wang, M., & Happee, R. (2022). Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles. Transportation Research Part F: Traffic Psychology and Behaviour, 86, 178–195. https://doi.org/10.1016/j.trf.2022.02.016

Hou, I., Man, O., Mettille, S., Gutierrez, S., Angelikas, K., & MacNeil, S. (2024). More robots are coming: Large multimodal models (ChatGPT) can solve visually diverse images of Parsons problems. Proceedings of the 26th Australasian Computing Education Conference, 29–38, Sydney, Australia. https://doi.org/10.1145/3636243.3636247

Huang, J., Jiang, P., Gautam, A., & Saripalli, S. (2023). GPT-4V takes the wheel: Evaluating promise and challenges for pedestrian behavior prediction. arXiv. https://doi.org/10.48550/arXiv.2311.14786

Hwang, H., Kwon, S., Kim, Y., & Kim, D. (2024). Is it safe to cross? Interpretable risk assessment with GPT-4V for safety-aware street crossing. arXiv. https://doi.org/10.48550/arXiv.2402.06794

Jain, A., Del Pero, L., Grimmett, H., & Ondruska, P. (2021). Autonomy 2.0: Why is self-driving always 5 years away? arXiv. https://doi.org/10.48550/arXiv.2107.08142

Johnson, W., Bouchard, T. J., Jr., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. Intelligence, 32, 95–107. https://doi.org/10.1016/S0160-2896(03)00062-X

Kolekar, S., De Winter, J., & Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. Nature Communications, 11, 4850. https://doi.org/10.1038/s41467-020-18353-4

Kolekar, S., Petermeijer, B., Boer, E., De Winter, J. C. F., & Abbink, D. A. (2021). A risk field-based metric correlates with driver's perceived risk in manual and automated driving: A test-track study. Transportation Research Part C: Emerging Technologies, 133, 103428. https://doi.org/10.1016/j.trc.2021.103428

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of the International Conference on Machine Learning, 12888–12900, Baltimore, MD.

Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). More agents is all you need. arXiv. https://doi.org/10.48550/arXiv.2402.05120

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., & Chen, W. (2023). Making language models better reasoners with step-aware verifier. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 5315–5333, Toronto, Canada. https://doi.org/10.18653/v1/2023.acl-long.291

Li, Y., Wang, L., Hu, B., Chen, X., Zhong, W., Lyu, C., & Zhang, M. (2024). A comprehensive evaluation of GPT-4V on knowledge-intensive visual question answering. arXiv. https://doi.org/10.48550/arXiv.2311.07536

Liao, H., Shen, H., Li, Z., Wang, C., Li, G., Bie, Y., & Xu, C. (2024). GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. Communications in Transportation Research, 4, 100116. https://doi.org/10.1016/j.commtr.2023.100116

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. Psychological Methods, 18, 285–300. https://doi.org/10.1037/a0033266

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Improved baselines with visual instruction tuning. arXiv. https://doi.org/10.48550/arXiv.2310.03744

Liu, M., Chen, C., & Gurari, D. (2024). An evaluation of GPT-4V and Gemini in online VQA. arXiv. https://doi.org/10.48550/arXiv.2312.10637

Liu, Y., Wang, Y., Sun, L., & Yu, P. S. (2024). Rec-GPT4V: Multimodal recommendation with large vision-language models. arXiv. https://doi.org/10.48550/arXiv.2402.08670

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2023). MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv. https://doi.org/10.48550/arXiv.2310.02255

Lu, X., Liusie, A., Raina, V., Zhang, Y., & Beauchamp, W. (2024). Blending is all you need: Cheaper, better alternative to trillion-parameters LLM. arXiv. https://doi.org/10.48550/arXiv.2401.02994

Markus, K. A., & Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. New Ideas in Psychology, 31, 54–64. https://doi.org/10.1016/j.newideapsych.2011.02.008

McDonald, R. P. (2003). Behavior domains in theory and in practice. Alberta Journal of Educational Research, 49, 212–230. https://doi.org/10.11575/ajer.v49i3.54980

Melman, T., Abbink, D. A., Van Paassen, M. M., Boer, E. R., & De Winter, J. C. F. (2018). What determines drivers' speed? A replication of three behavioural adaptation experiments in a single driving simulator study. Ergonomics, 61, 966–987. https://doi.org/10.1080/00140139.2018.1426790

Näätänen, R., & Summala, H. (1974). A model for the role of motivational factors in drivers' decision-making. Accident Analysis & Prevention, 6, 243–261. https://doi.org/10.1016/0001-4575(74)90003-7

Nagle, F., & Lavie, N. (2020). Predicting human complexity perception of real-world scenes. Royal Society Open Science, 7, 191487. https://doi.org/10.1098/rsos.191487

Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. Proceedings of the National Academy of Sciences, 114, 7571–7576. https://doi.org/10.1073/pnas.1619003114

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. New York. NY: McGraw-Hill.

OpenAI. (2023). GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Qi, Z., Fang, Y., Zhang, M., Sun, Z., Wu, T., Liu, Z., Lin, D., Wang, J., & Zhao, H. (2023). Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases. arXiv. https://doi.org/10.48550/arXiv.2312.15011

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv. https://doi.org/10.48550/arXiv.1804.02767

Salinas, A., & Morstatter, F. (2024). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. arXiv. https://doi.org/10.48550/arXiv.2401.03729

Sawaki, Y. (2010). Generalizability theory. In N. J. Salkind (Ed.), Encyclopedia of research design. Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781412961288

Senkaiahliyan, S., Toma, A., Ma, J., Chan, A.-W., Ha, A., An, K. R., Suresh, H., Rubin, B., & Wang, B. (2023). GPT-4V(ision) unsuitable for clinical care and education: A clinician-evaluated assessment. medRxiv. https://doi.org/10.1101/2023.11.15.23298575

Tabone, W., & De Winter, J. C. F. (2023). Using ChatGPT for human-computer interaction: A primer. Royal Society Open Science, 10, 231053. https://doi.org/10.1098/rsos.231053

Tang, R., Zhang, X., Ma, X., Lin, J., & Ture, F. (2023). Found in the middle: Permutation self-consistency improves listwise ranking in large language models. arXiv. https://doi.org/10.48550/arXiv.2310.07712

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. arXiv. https://doi.org/10.48550/arXiv.2401.06209

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. arXiv. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in neural information processing systems, Vol. 35 (pp. 24824–24837). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2201.11903

Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., Ma, T., Li, Y., Xu, L., Shang, D., Zhu, Z., Sun, S., Bai, Y., Cai, X., Dou, M., Hu, S., Shi, B., & Qiao, Y. (2023). On the road with GPT-4V(ision): Early explorations of visual-language model on autonomous driving. arXiv. https://doi.org/10.48550/arXiv.2311.05332

Wilde, G. J. S. (1982). The theory of risk homeostasis: implications for safety and health. Risk Analysis, 2, 209–225. https://doi.org/10.1111/j.1539-6924.1982.tb01384.x

Wilde, G. J. S. (2013). Homeostasis drives behavioural adaptation. In C. M. Rudin-Brown & S. L. Jamson (Eds.), Behavioural adaptation and road safety: Theory, evidence and action (pp. 61–86). Boca Raton, FL: CRC Press.

Wu, C., Lei, J., Zheng, Q., Zhao, W., Lin, W., Zhang, X., Zhou, X., Zhao, Z., Zhang, Y., Wang, Y., & Xie, W. (2023). Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. arXiv. https://doi.org/10.48550/arXiv.2310.09909

Yan, A., Yang, Z., Zhu, W., Lin, K., Li, L., Wang, J., Yang, J., Zhong, Y., McAuley, J., Gao, J., Liu, Z., & Wang, L. (2023). GPT-4V in wonderland: Large multimodal models for zero-shot smartphone GUI navigation. arXiv. https://doi.org/10.48550/arXiv.2311.07562

Yan, X., Zhang, H., Cai, Y., Guo, J., Qiu, W., Gao, B., Zhou, K., Zhao, Y., Jin, H., Gao, J., Li, Z., Jiang, L., Zhang, W., Zhang, H., Dai, D., & Liu, B. (2024). Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. arXiv. https://doi.org/10.48550/arXiv.2401.08045

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., & Gao, J. (2023). Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv. https://doi.org/10.48550/arXiv.2310.11441

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of LLMs: Preliminary explorations with GPT-4V(ision). arXiv. https://doi.org/10.48550/arXiv.2309.17421

Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., & Zhou, J. (2023). mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv. https://doi.org/10.48550/arXiv.2311.04257

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2023). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. arXiv. https://doi.org/10.48550/arXiv.2311.16502

Zhang, C., & Wang, S. (2024). Good at captioning, bad at counting: Benchmarking GPT-4V on Earth observation data. arXiv. https://doi.org/10.48550/arXiv.2401.17600

Zhang, D., Yang, J., Lyu, H., Jin, Z., Yao, Y., Chen, M., & Luo, J. (2024). CoCoT: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. arXiv. https://doi.org/10.48550/arXiv.2401.02582

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning, 180, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W. Y., & Petzold, L. R. (2023). GPT-4V(ision) as a generalist evaluator for vision-language tasks. arXiv. https://doi.org/10.48550/arXiv.2311.01361

Zhou, X., & Knoll, A. C. (2024). GPT-4V as traffic assistant: An in-depth look at vision language model on complex traffic events. arXiv. https://doi.org/10.48550/arXiv.2402.02205

Appendix

Figure 8.6: Results of YOLOv4 for 2 of the 210 images.

9

Discussion and Conclusions

Technological advancements such as the increasing integration of sensors in modern vehicles, availability of computation power, and new algorithms to process naturalistic data have created new opportunities for improving human driving through data-driven analysis and personalized feedback. The primary aim of this dissertation was to develop and test algorithms capable of detecting patterns in human driving, with a focus on car and truck drivers. This chapter presents a chapter-by-chapter summary of the findings, followed by a synthesis of overarching conclusions. The chapter concludes with a reflection on the findings in the dissertation as a whole.

9.1. Summary of findings per chapter

Part I: Perspectives on data use and technology in driver testing and the trucking industry

Chapter 2: Data needs of driving examiners for assessing candidates

Chapter 2 presented the results of semi-structured interviews with 37 driving examiners about their perspectives on data-driven assessment of driving candidates. Examiners supported using data to explain their pass/fail decisions to candidates and preferred accessible formats like graphs of eye movements, headway, speed, and braking behavior. They believed this approach could clarify decisions and resolve candidate disagreements. However, they were cautious about using data for higher-level decision support, arguing that context-specific information and obtaining an overall image of the candidate are important for their verdict and that they could not rely solely on data for this. The interviews also revealed potential applications of data beyond assessment, such as optimizing test routes, improving standardization, and preparing candidates better before taking the driving test.

Chapter 3: Perspectives of truck drivers on their profession and safety technologies

Chapter 3 presented survey findings from 3,708 Dutch truck drivers, focusing on their views about their profession and the impact of safety-related technology on their work. Drivers expressed dissatisfaction with their public image and appeared hesitant to recommend their profession to others. While they reported moderate work pressure, they faced challenges such as low wages, poor work-life balance, and insufficient support from transport organizations. Factor analysis indicated that national drivers experienced more work pressure, which could be explained by frequent trips and dense traffic, whereas international drivers were more concerned with safety issues like transport crime during night shifts.

Regarding in-vehicle technology, drivers had mixed views on Advanced Driver Assistance Systems (ADAS). They appreciated features such as Adaptive Cruise Control (ACC) for improving safety and fuel efficiency, but criticized systems like Lane Keeping Assist (LKA) and Automatic Emergency Braking (AEB) for generating false alerts, leading some drivers to disable these features. While on-board computers were valued for logistical efficiency, they also contributed to feelings of being constantly monitored and increased work pressure due to real-time tracking. Overall, drivers favored technology that genuinely reduced their workload and improved safety but resisted systems perceived as un-reliable or intrusive.

Part II: Measuring human driving behavior: practical approaches

Chapter 4: Detecting lane change maneuvers from GPS data

Chapter 4 evaluated the accuracy of a lane change detection method using GPS data. Lane changes were identified using lateral movement thresholds and road geometry information, achieving an overall accuracy of 90%. The true positive rate was 89%, with a false positive rate of 8%. The false positives varied by road type, being lower (3%) on straight highways and higher (10%) on complex roads with curves and exits.

The method shows potential for applications in traffic research and road design when aggregated statistics are required, such as identifying lane change hotspots. However, the method may not be suitable for real-time systems like lane drift warnings due to the high frequency of false positives, which could result in drivers turning off such alerts.

Chapter 5: Detecting driving style from accelerometer and GPS data

Chapter 5 explored the use of accelerometer and GPS data to differentiate between driving styles. Experienced driving examiners acted out cautious, normal, and aggressive driving styles, and the analysis focused on metrics such as harsh accelerations, jerk, and speeding. The results showed that driving style, vehicle type, and route had a significant impact on these metrics. Despite the small sample size and other limitations, the findings demonstrated the potential of sensor-based data to support more objective assessments in driver training and testing, and indicated that relatively cheap sensor setups can already prove to be a valuable addition.

Chapter 6: Predicting damage incidents, fines and fuel consumption from telematics data

Chapter 6 analyzed telematics data from 180 truck drivers over a two-year period to understand the factors contributing to their damage incidents, traffic fines, and fuel consumption. The analysis revealed that the frequency of harsh braking events was a predictor of both damage incidents and traffic fines, while fuel consumption was more closely associated with engine torque exceedances. However, these outcomes were not solely determined by driver actions. For example, driving in more complex environments, such as urban routes, required more braking in general and was thus characterized by more harsh braking events. Furthermore, larger trucks were associated with fewer harsh braking incidents. These findings suggest that the impact of driver behavior on costs must be understood in the context of the driving conditions and the vehicle characteristics.

Part III: Measuring human driving behavior: towards context-aware methods

Chapter 7: Using AI from automated driving systems to assess human driving behavior

Chapter 7 introduced a new approach to assessing human driving performance using AI capabilities from modern automated driving systems. Traditional assessment methods such as counting harsh braking events lack the context behind these actions. In this study, an automated driving system (Openpilot) was adapted to replay recordings of human driving and compare it with the internal predictions of the AI. Discrepancies between the AI's plan and the human's actions can give indications of the context under which the event occurred.

Specifically, this chapter demonstrated a scenario where the driver chose to brake late and hard (aggressively) in response to an early observable oncoming obstruction, and showed that the AI suggested an early deceleration action. Conversely, in a surprise scenario, in which an oncoming danger could not be predicted until the last moment, both the human and the AI braked late, indicating that in this case, the hard brake was appropriate or inevitable. This analysis demonstrated how using AI, we can distinguish between justified and unjustified braking in post-hoc analysis of driving events, providing a more accurate picture of driver behavior. Thus, this method provides a more context-aware evaluation of driving behavior than traditional methods.

Chapter 8: Using multimodal large language models to predict driving risk

The final chapter demonstrated that GPT-4V, a Large Language Model with image recognition, can accurately predict the perceived risk in traffic images, achieving a strong correlation with human assessments. Using a dataset of 210 traffic images previously rated by human participants, the model's predictions reached a validity coefficient of $r = 0.83$. The study explored various prompting techniques and the integration of object detection features to refine the model's output and showed that (1) repeating prompts under the same conditions improves output consistency, (2) varying prompts and averaging results increases accuracy, and (3) combining object detection data with GPT-4V's risk scores increases predictive power. These findings demonstrate the potential of general-purpose AI models in understanding complex traffic contexts.

9.2. Overarching conclusions

Conclusion 1: Examiners and truck drivers want data-driven tools that support professional autonomy rather than constrain their judgment

Part I showed what driving professionals think about integrating data-driven tools in their work. Both driving examiners (*Chapter 2*) and truck drivers (*Chapter 3*) appreciated technologies that would support their professional judgment and reliably improve their work, while preserving their autonomy in decision-making. This overarching conclusion can be drawn for both groups, though with distinct details for the different professions.

Driving examiners were specifically interviewed about potential future applications. They saw value in tools supporting their existing assessment process, rather than tools that prescribe decisions. They particularly recommended designing interfaces that would help them communicate their judgment to candidates and argued that establishing an overall image of the candidate's ability to drive likely remains a human task.

Professional truck drivers commented on experiences they had in the past with technology. They accepted tools that supported their professional capabilities without imposing unnecessary restrictions. They often expressed frustration with ADAS due to frequent false alerts that often led drivers to turn off these features (for similar observations, see

Dedhia et al., 2023; Dreger et al., 2020). Furthermore, planning and monitoring systems can come across as intrusive and give the feeling of being monitored. Whether this is a wanted or unwanted effect may be up for debate. On the one hand, the sense of being monitored may contribute to road safety. On the other hand, it may decrease the sense of freedom and autonomy that has led many truck drivers to choose the job (Bhoopalam et al., 2023) and stand in the way of gaining acceptance for new technologies, particularly if introduced in a way that is not transparent, beyond the control of the drivers, and without the associated safety benefit (Camden et al., 2022).

Conclusion 2: Readily available sensors enable population-level insights but require contextual information for individual assessment

Part II demonstrated that data from modern vehicle sensors and mobile devices can capture driving patterns. This is shown by contributions such as 90% accuracy in lane change detection while using GPS only (*Chapter 4*), successful differentiation between re-enacted driving styles using accelerometers (*Chapter 5*), and predicting damage incidents, fines, and fuel consumption through telematics data in truck fleets (*Chapter 6*).

The primary value of this sensor data lies in its application to population-level insights. Analyses based on aggregated data can allow for the identification of lane change hotspots (*Chapter 4*), the identification of typical driving styles among novice drivers (*Chapter 5*), or the understanding of fleet-wide risk and fuel consumption patterns (*Chapter 6*). Population-level insights work because they average out individual and contextual variations under typically large sample sizes; when studying lane change hotspots or fleet-wide risk patterns, individual and contextual variations become less important.

However, when evaluating individual driver performance, both statistical and contextual limitations emerge. From a statistical perspective, correlations between individual driving metrics and outcomes have been low (*Chapter 6*) and detection algorithms regularly produce false positives (*Chapter 5*). Such statistical limitations stem partly from the relative rarity of certain events, such as damage incidents, making it difficult to establish reliable patterns with sufficient statistical power. Although presenting the driver with performance data may raise awareness and encourage improvement (for example, reducing harsh braking), caution should be taken when attributing low scores to driver's individual traits or behavior. In other words, patterns identified at the group level should not be mistaken for patterns that apply to individual drivers.

From a contextual perspective, we have seen that the same recorded driving event could represent either good or poor driving depending on the circumstances. A hard brake might represent dangerous driving; yet the same braking intensity could indicate excellent defensive driving to a sudden emergency (*Chapter 7*). This principle was also noted in our interviews with driving examiners (*Chapter 2*), who pointed out that proper assessment requires establishing an overall image of a driver's capabilities rather than a focus on individual events (see also Souman et al., 2021). Another contextual factor to take into account is that drivers may simply live in areas that require more frequent harsh braking. Though the techniques presented in *Part II* can capture context in some way, for example, GPS may provide location and road type information, which may be enriched with information about weather, traffic, or location-specific irregularities, the lack of visual information meant that the analyses remained relatively agnostic to the specific scene. However, camera images arguably provide the most context-rich source of information. These image-based approaches were explored in *Part III*, in combination with AI techniques.

Conclusion 3. Modern AI-driven techniques form promising avenues for context-aware driving evaluation

The limitations of traditional metrics and the need for contextual understanding have led to exploring the potential of modern AI systems in driving assessment. In the final chapters of this dissertation, it was found that the AI-driven techniques designed and explored in Part III offer promising first steps towards incorporating the rich contextual awareness that humans naturally possess into the automated assessment of driving behavior.

Chapter 7 demonstrated a novel assessment method by comparing the behavior of an automated driving system with that of a human operator. The actions suggested by the AI of the automated driving system were characterized by a degree of contextuality that was not found in the metrics from Part II and could thus be used as a form of reference driving behavior. In the demonstration, this idea was used to distinguish between justified and unjustified driving maneuvers. Instead of simply flagging a harsh braking event, the system could determine whether that brake was an appropriate response to the situation. For instance, when a driver braked aggressively in response to an early observable obstacle, the AI system indicated an earlier desire to brake than the human did. However, when faced with a surprise scenario, the lack of early suggested action by the AI system indicated that the harsh braking scenario was unforeseen and that the driver's sharp braking was an appropriate response.

It is noteworthy that comparing AI predictions to human driving is a workflow that is already used by automated vehicle developers to evaluate the performance of the automated driving system, sometimes referred to as *shadow mode* (Golson, 2016; Harris, 2022; Kuipers, 2022). In essence, the approach we used turned this concept around by drawing conclusions about the human instead of the machine, similar to the "reference driver model" by Michon (1993, p. 12).

The potential of novel AI methods in driving assessment was further confirmed by the findings in *Chapter 8*. The strong correlation between GPT-4V's risk assessment and human judgments demonstrates that modern AI systems can process and interpret traffic scenarios in ways that are similar to human observations. The fact that GPT-4V, unlike the AI used in *Chapter 7*, was not specifically designed for driving situations but rather for general-purpose tasks, makes the performance perhaps more impressive.

Though the two explorations in Part III were relatively preliminary in nature, the results showed remarkable promise. The rapid development of AI suggests even greater potential ahead. Both approaches, using specialized driving AI as a reference driver and using advancements in (visual) LLMs for context understanding, demonstrated progress towards addressing the challenges of context-aware driving assessment, and an increasing number of studies are surfacing on the use of visual and/or multimodal LLMs on the topic (e.g., Hwang et al., 2024; Zhang et al., 2024; Zhou & Knoll, 2024). Looking forward, these AI-driven approaches could open new possibilities for large-scale, automated evaluation systems that can provide multiple perspectives on real-time driving performance. As AI technology continues to evolve, these systems may not only fundamentally transform how we assess and understand human driving behavior, but may eventually even take driving out of our hands completely, ironically removing the need for assessing human drivers altogether.

9.3. Ethical considerations and limitations

9.3.1 Proxy discrimination, fairness, and explainable AI

The inability to fully separate driving behavior from environmental factors might result in proxy discrimination. Predictive models could unintentionally discriminate based on sensitive traits that are correlated with characteristics like race, socio-economic status, or other demographic factors (Barocas & Selbst, 2016). Even without explicit bias, certain patterns in driving behavior linked to environmental or demographic factors could result in unfair outcomes. For example, models trained to optimize safety might unfairly penalize drivers in regions with more unpredictable traffic conditions or less-developed infrastructure, indirectly reflecting socio-economic disparities.

Another concern regarding proxy variables, particularly relating to insurance, is that they can create compounding penalties for the same underlying risk factors. For example, urban residents already pay higher premiums due to increased accident risk in cities, usually based on the insured's registration address (Verbelen et al., 2018). If the same drivers are penalized for more frequent hard braking events—a natural consequence of driving in dense urban traffic—they are essentially being charged twice for the same environmental risk factor. Similarly, novice drivers traditionally pay higher insurance premiums due to their limited experience (Isotupa et al., 2019; Verbelen et al., 2018). When behavioral monitoring systems flag events like engine stalling, over-revving, or jerky driving, these drivers might face additional penalties for patterns that are to be expected from novice drivers.

The extent to which certain variables such as policyholder's location or telematics measures should be used in risk assessment remains an ongoing discussion. In many cases, absolute fairness is unattainable, and efforts to mitigate biases or phenomena such as proxy discrimination can decrease other measures of group fairness (Lindholm et al., 2024). For example, a straightforward but crude mitigation strategy against this proxy discrimination would be to exclude policyholder location from premium calculation. While this removes location-based proxy discrimination, it would introduce other group unfairness issues; policyholders in low-risk, less congested areas might end up subsidizing those in high-risk regions.

In vehicle insurance, it is therefore an accepted practice to request ZIP code data when determining premiums, since it is a valid and easy-to-explain predictor of risk. However, modern algorithms described in this dissertation and discount strategies applied in pay-how-you-drive insurance models are much more difficult to explain to users, as they are often based on more opaque algorithms and sensors. Improving the explainability of such model outcomes is often categorized under the term "Explainable AI", and strategies include "rationale explanation"–providing understandable reasons for why a model reached a particular decision–and offering estimates of model and measurement uncertainties such as confidence intervals when presenting performance scores (Barredo Arrieta et al., 2020). This can improve the transparency of decisions made by these algorithms.

9.3.2 Loss of freedom, normalization of surveillance

While there may be a safety benefit, the normalization of surveillance may lead to a reduction in personal freedom. Driving systems that constantly monitor behavior might shift driving from a personal skill to a highly regulated and supervised activity. The risk is that systems will be designed in an overly strict manner, where every slight misstep or deviation from 'optimal' driving is logged and evaluated.

9.3.3 Risk of predicting risk: the insurance model

The current dissertation contributed to algorithms that can be used to predict risk. In *Chapter 6*, for example, we found that damage incidents could be predicted by tracking harsh braking events. In the case of the insurance company we collaborated with, these scores were used by a third-party service solely for the purpose of providing coaching and feedback to drivers. However, an increasing number of insurance companies (e.g., Allianz, 2024; ANWB, 2024; see *Chapter 7*) offer financial rewards or lower premiums for drivers who demonstrate safer driving behaviors.

While this approach likely increases awareness and promotes safer driving habits, it also introduces challenges to the traditional insurance model, which is based on the principle of distributing risk and costs across a group. Even though accidents will likely remain events that are difficult to predict with high certainty, the increasing ability to assess individual risk may seduce insurance providers to become more aggressive in the way they implement reward (or punishment—depending on the angle) strategies.

This raises ethical and philosophical questions. For instance, in health insurance, it is illegal to discriminate based on certain factors such as pre-existing conditions, weight, or whether an individual is a smoker, in order to maintain fairness and collective risk-sharing. If driving risk can be predicted with greater accuracy, should insurers be allowed to penalize or reward individuals based on these predictions? Such a shift could lead to preselection of low-risk individuals. It also undermines the basic principle of insurance as a form of collective pooling of uncertain outcomes (Cevolini & Esposito, 2020).

9.4. Reflections

9.4.1 The future of driving: more automation, more assessment, or more of both?

This thesis has demonstrated techniques to measure human driving behavior by using increasingly advanced algorithms. For example, *Chapter 8* has demonstrated that risk in traffic images can be predicted using one of the first publicly available multimodal LLMs, and *Chapter 7* offers a demonstration of how existing driving models may be used to compare human driving with an AI "reference" driver. The pace at which AI is developing is also reflected in the impressive demonstrations by vehicle manufacturers such as Waymo (2024), who now offer automated taxis in select places. The rate of innovation can also make one think that perhaps automated driving is around the corner. In this light, a fair question arises: if vehicles will soon be driving themselves, why do we still need to focus on improving human driving skills? Besides issues in public trust and legal challenges that were briefly addressed in the introduction section of this dissertation, in the following sections I offer other arguments as to why achieving fully autonomous vehicles for mass use is complex and may be further away than expected.

Accessibility and socioeconomic barriers

While fully autonomous vehicles (AVs) may soon become available to the public, this "soon" has been continuously postponed. Even when AVs do reach the market, they may not be affordable or accessible to everyone. According to the World Health Organization (2023), 92% of road fatalities occur in low- and middle-income countries, in which market penetration of advanced safety features is delayed. In high-income countries, individuals with lower socioeconomic status are also more likely to be involved in crashes. New technologies tend to be introduced in electric vehicles (EVs), but in many regions, including lower-income countries, EV infrastructure may not be sufficiently developed for mass adoption. The spread of safety features also depends on the rate at which modern vehicles replace older ones within the fleet. The lag in adoption is illustrated by the average age of passenger cars on the road, which was over 12 years in the EU and 14 years in the US as of 2024 (ACEA, 2024; Bureau of Transportation Statistics, 2024). This means that most vehicles were developed with safety standards and features from previous decades. For the foreseeable future, quicker solutions to improving road safety will likely come from mobile applications and external hardware (e.g., smartphones, telematics devices), as explored in *Chapters 4* and *6*.

Challenging environments and willingness to take risks

The development of vehicle automation has largely been concentrated in the United States, where most of the testing and deployment occurs in relatively controlled environments, such as on highways or in urban areas with established road infrastructure.

In many places, the driving environment poses greater challenges. For example, countries like the Netherlands are known for their dense network of cyclists, often sharing roads with cars in complex situations. Moreover, in some situations, like merging in heavy traffic or claiming the right of way at intersections, human drivers often make split-second decisions that involve balancing safety with assertiveness in combination with social signals like gestures, approaching speed, vehicle placement et cetera.

It remains unclear whether manufacturers of automated driving systems are willing to program their vehicles to take similar risks. Autonomous vehicles might be overly cautious in situations that require more assertive maneuvering. It is possible that in certain areas—even though automated vehicles theoretically would be able to drive safer there—these vehicles will not become available for this reason.

From assessing humans to assessing machines

In conclusion, while the above reflections have argued that human driving assessment will likely remain relevant for years to come, it is worth considering how the assessment techniques developed in this dissertation could evolve in a future where automated vehicles become the norm. The context-aware evaluation methods explored in this work in particular show promise for validating automated systems. The AI-as-a-reference approach demonstrated in *Chapter 7* could develop into a multi-agent validation framework, where a committee of certified automated systems serves as a benchmark for evaluating new systems. Moreover, the ability of general-purpose AI models to perceive risk in traffic contexts (*Chapter 8*) suggests promising directions for automated evaluation of AI driving decisions. Thus, some of the techniques developed for human driver assessment from this dissertation may find new future applications in validating the very technology that could eventually replace human drivers.

9.4.2 A reflection on using generative AI in academic writing

The period in which this dissertation was written (late 2020–early 2025) spans a unique time period regarding the general adoption of generative AI. Approximately in the middle, November 2022, ChatGPT was released to a wide audience. By the end of the period, such models outperform humans in many tasks (De Winter et al., 2024a; Latif et al., 2024; Mittelstädt et al., 2024; Zhai et al., 2024).

Although earlier LLM models were already available before this release, they did not yet have the typical chat interaction now common in many applications. From my early interactions with those models, I remember being quite amazed by the fact that computers could come up with such sophisticated human-like completions. Yet, I did not find much practical use for it in my day-to-day activities. The release of ChatGPT changed this.

LLMs as a productivity aid

When ChatGPT first was released, I immediately recognized its user-friendliness and its potential. A current walk through the hallways of our faculty building shows that many students typically have a window of ChatGPT or other popular LLMs opened. How quickly ChatGPT was adopted was also noticeable as it became common to recognize AI-generated text in papers, emails, and even academic peer reviews that I and colleagues received. Texts often had a similar tone to them and sounded somewhat cliché, molded into the same sentence structure or using similar specific words. Many texts were now "delving" into "comprehensive" analyses, and their "key" and "significant" insights were suddenly leading to "crucial" contributions (see De Winter et al., 2023; Kobak et al., 2024, for more examples of such terms).

Increased productivity or inflation of words?

I noticed that while individual passages encountered were clearer and less repetitive, the overall text, when compared to other AI-assisted writings, had become more uniform and homogenous. It is as if we were all using the same proofreader. This is especially noticeable in unedited LLM output such as the GPT-4 generated summaries of *Chapter 3* (in that case, used as part of the research method rather than as a writing assistant), that contain, for example, "key issues", "key points", "key themes", "significant concern(s)" (2x), "a significant contributor", "significant roles" over the span of two pages (p. 91-93).

The ability to transform scattered drafts and bullet points into polished texts within seconds also came with other effects. In a way, the value of a piece of text decreases when the perceived effort is lower, which can create a certain sense of meaninglessness to the reader. De Winter et al. (2025) used the term *inflation of words* to describe how AI-generated text, while technically proficient, often lacks the inherent value and meaning of human-written content due to large quantities of words becoming available at low cost, or without struggle. Before LLMs, any text offered a window into the author's thought process and understanding. AI-generated text, despite being technically correct, often feels hollow.

However, the appeal of using LLMs for writing seems to outweigh some of the negative aspects. There are widely documented positive effects on productivity in both text and code writing. Besides that, I believe an important positive aspect of current LLMs is that they have the ability to make participation in academia more accessible to a larger group of people.

Accessibility of academia

Some of the smartest minds I have encountered were fellow students who excelled at math, coding, prototyping, or other hands-on technical skills. Yet many of these same individuals, including some with dyslexia, dreaded the documentation process. They often struggled through their master's thesis writing, despite doing exceptional work. If I were to ask them about pursuing a PhD, they would likely dismiss the idea due to the associated writing demands. And though there are valid reasons to question pursuing an academic career, I believe that adversity towards writing does not need to be one anymore.

Writing research papers has evolved into the primary way to compress and disseminate academic ideas efficiently. I believe that a written traditional research paper remains the format for storing the outcome of empirical research work. It is universal and forces careful consideration of what information to include. However, the path to creating these papers is transforming. Instead of processing several months of work by sitting down and carefully constructing a research paper, authors can now engage in organic, conversation-based workflows with LLMs.

Through interactions with LLMs, especially longer format conversations with counter-questions for clarity, authors can achieve compression of their ideas with fewer concerns about grammar, style, or writing flow. These interactions can use mixed modalities, such as a mix of raw drafts of paragraphs and bullet points about the author's explanations and ideas, combined with a turn-taking auditory conversation where the LLM asks clarifying questions about the work. The LLM can then help structure these various inputs into coherent academic writing. The researcher can take their turn in verifying the output, and prepare a next iteration. Ideally, this improves the overall rigor and depth expected from research papers, while removing the barriers of writing.

If writing becomes more approachable and even enjoyable for those who traditionally struggle with it, science as a whole could benefit tremendously. I am personally experiencing more joy in writing, and even more so in coding, when there is always an assistant at hand to have a back-and-forth with, one that displays a certain understanding of the topic and occasionally comes up with creative suggestions to integrate the ideas into a consistent whole.

I wonder how many potential discoveries have been lost because good thinkers have felt intimidated by the writing process. If the future of academia can be more about the quality of ideas rather than writing ability, by allowing researchers to focus on their talents while using AI to help translate their thoughts into well-structured papers, the gains for scientific progress could be big.

The future of generative AI

The experiences and concerns discussed above have mostly been shaped by my interactions with generative AI thus far. However, concerns such as the homogenization of written language are specifically focused on the topic of writing using LLMs, and they pale in comparison to the broader discussion surrounding the consequences of generative AI in the future. Though a full exploration of these potential consequences could fill an entire dissertation, several pressing issues deserve mention.

First, there is the concentration of power in the hands of a few tech companies that own and control these AI systems. They are not just providing a service; they are becoming the gatekeepers of an infrastructure that powers how we write, think, and create. When your work needs to flow through someone else's servers, questions of privacy and control come into play. However, the rapid development of open-source models like Llama (Grattafiori et al., 2024) and Deepseek (Deepseek-AI, 2025), which offer performance comparable to commercial models and can be run locally, suggests advanced models may in the future not always remain proprietary.

A second and perhaps most worrying issue is how cheap and easy it is becoming to deploy armies of intelligent bots. This is not only spam;these are sophisticated systems that can engage in online conversations, shape public opinion, and influence discourse at a scale we have never seen before (Harari, 2024). When artificial voices become indistinguishable from real ones, how do we maintain authentic public dialogue?

Autonomous researchers

I have discussed several ways I have used LLMs in this dissertation; as productivity aids for writing and coding, as a productivity aid within research methods (*Chapter 3*), and as research subject (*Chapter 8*). However, a fourth potential role looms on the horizon: LLMs as independent researchers. While I have had no success using current LLMs for generating truly novel ideas or executing research independently, the pace at which the

Box 1: How LLMs were used in this dissertation

As LLM capabilities progressed over the past years, my usage changed accordingly. While most chapters primarily involved simple language checks, more advanced applications such as code completion, automatic summarization and vision-based analysis were also used. Specifically, here I describe how LLMs were used in the chapters:

- Chapter 2 and 4 were written without any assistance from LLMs. In fact, ChatGPT became available only in November 2022 (i.e., after the publication of these chapters). Also, at that time, ChatGPT was still of limited quality. Current models, such as o1 and o3-mini are considerably more powerful and allow for reasoning, tabulation, sophisticated computer coding, and other advanced tasks.
- In Chapters 1, 3, 5, 6 and 7 LLMs were used for retroactive language checks but not for content generation. This is not much different from the use of Grammarly, for example.
- For the present thesis summary and section 9.1, I prompted an LLM (specifically: Claude Sonnet 3.5; a competitor to OpenAI's ChatGPT) to create automatic summaries of my work, which I subsequently rewrote and edited. The summary has also been translated from English to Dutch by an LLM; I subsequently manually verified and edited this translation.
- Chapter 7 made use of Github Copilot for coding assistance (the code is available on GitHub; Driessen, 2024).
- In Chapter 3, LLMs were used as a research method by systematically using OpenAI's API to generate automated summaries and for sentiment analysis of questionnaire textbox responses. I firmly believe that this type of research approach will become more and more popular in the coming years.
- Besides as a research method, LLMs were used as a research subject. In Chapter 8, it was tested if multimodal LLMs are capable of recognizing risk in driving situations. At the time, the content of Chapter 8 was one of the first in the world to evaluate a multimodal LLM (also referred to as a vision-language model), that is, an LLM that can process text and images in combination.
- And, not included in this dissertation but nonetheless noteworthy, LLMs were used as virtual research subjects in De Winter et al. (2024b), where we investigated the potential for using LLMs for the simulation of human questionnaire participants. The use of 'virtual humans' to pre-test questionnaires and product designs without involving actual humans is something I foresee will become more and more prevalent in the years to come.

models are better at executing tasks is striking.

Having observed the evolution of these models over the past few years, their trajectory of improvement is remarkable. The costs to generate words are dropping very fast. For example, the model used in *Chapter 3* (GPT-4-0125-preview), which was the most capable model at the time, cost $30 per million tokens (approximately 1.3 million words) on March 1, 2024 (OpenAI, 2024). At the time of writing, less than a year later, on January 27, 2025, a model with superior performance (GPT-4o-mini, according to LMSYS ratings (Chiang et al., 2024)) was available for just $0.60 per million output tokens (OpenAI, 2025). And with the release of the aforementioned open-source public models, advanced models can be run locally or on private servers at cost price.

Alongside dropping costs, there is increasing competence of these models. Dario Amodei, CEO of Anthropic–the AI company that released Claude, my model of choice for coding assistance over the past months, predicts that based on his observations within the company, AI could surpass human intelligence within several years (WSJ News, 2025). Although the history of futuristic predictions by big tech leaders, such as those in automated driving, has taught us to be wary of timeline estimates, these predictions do not seem very shocking in light of the rapid progress of the past several years.

This potential future raises fundamental questions about the nature of academic research itself. What defines academia when machines can potentially conduct research autonomously? Should certain types of research remain exclusively human domains? These questions might seem premature, but given the pace of AI development, they deserve serious consideration now.

9.5. Practical applications and impact

The methods and findings presented in this dissertation have implications for various domains within road transport and traffic safety.

The survey results from *Chapter 3* may directly inform policymakers regarding the needs and pain points of truck drivers, and their relation with technology (in addition, see de Winter et al., 2024c, regarding truck drivers' preference of mirrors over smart cameras).

Three related domains that may find practical use in the findings from this dissertation are vehicle insurance companies, fleet management providers, and driver coaching platforms such as NEXTdriver (2024). Specifically, the findings from *Chapter 6* (correlation between telematics measures and damage incidents, fines, and fuel consumption) can be used to optimize feedback mechanisms, focusing on the measures that have the strongest correlation with negative outcomes and ignoring measures that are less relevant.

While the sensor-based detection methods from *Chapter 4* and *Chapter 5* and AI-driven methods from *Chapter 7* and *Chapter 8* are still in early stages of development, they point toward valuable future applications. For example, the methods presented for detecting lane change patterns (*Chapter 4*) could support evidence-based policy making. By identifying high-risk locations or common behavioral patterns associated with lane changes, policymakers and road authorities can make more informed decisions about infrastructure improvements and traffic management strategies. The AI-driven methods of *Chapter 7* and *Chapter 8* show promise for insurance companies and driver coaching platforms, potentially enabling more nuanced evaluation of driving performance by considering environmental and situational factors.

Perhaps the most immediate practical impact of the research in this dissertation can be seen in driver testing and education, particularly in its implementation by the Dutch Central Office of Driving Certification (CBR). Box 2 describes how CBR is incorporating results from this research (specifically from *Chapter 2* and *Chapter 5*) into their driver testing procedures.

Box 2: Impact at the Dutch Central Office of Driving Certification (CBR)

The findings from this dissertation have directly influenced the development of data-driven assessment tools at the Dutch Central Office of Driving Certification (CBR). Building on the examiner interviews (*Chapter 2*) and sensor processing methods from *Chapter 5*, CBR has developed a prototype system that supplements traditional examination procedures with objective data collection and analysis.

The prototype equips examination vehicles with sensors (GPS, OBD, and accelerometers) to collect real-time driving data during tests. This data is presented through a dashboard interface that examiners can access within minutes after test completion (Figure 9.1). The system transforms subjective observations into quantifiable measurements—for example, changing statements like "you drove too fast" into specific feedback such as "you exceeded the speed limit by 10 km/h at the fourth intersection." (Schippers & Stefan, 2024a)

This implementation directly built on findings from *Chapter 2*, where examiners expressed interest in tools that would help them communicate their judgments more effectively while preserving the freedom to form their own judgments. The prototype has been successfully deployed in internal training sessions, with examiners reporting improved ability to explain their verdicts and help candidates understand their driving behavior in relation to traffic safety (Schippers & Stefan, 2024a).

The success of this initial implementation has led CBR to develop these capabilities further through their recently introduced "Driving Data" project, with plans to deploy the dashboard across their nationwide driver testing program. The impact of this research is further illustrated by CBR's recruitment of David Stefan, the joint first author of *Chapter 5*.

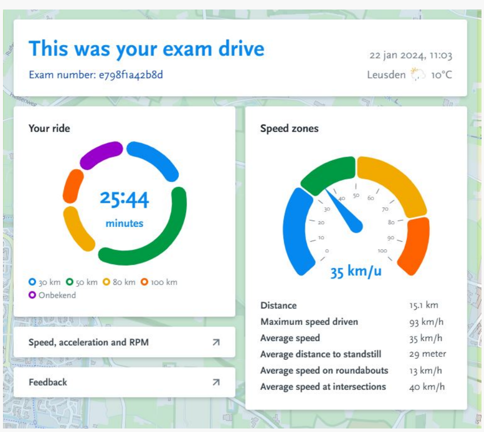


Figure 9.1: Prototype of dashboard interface by CBR (from Schippers & Stefan, 2024b).

References

ACEA (2024, March 28). Average vehicle age on the rise: Policy makers must do more to incentivise the switch. https://www.acea.auto/news/average-vehicle-age-on-the-rise-policy-makers-must-do-more-to-incentivise-the-switch

Allianz. (2024). BonusDrive. https://www.allianz.de/auto/kfz-versicherung/telematik-versicherung

ANWB. (2024). Veilig rijden [Safe driving]. https://www.anwb.nl/verzekeringen/autoverzekering/veilig-rijden/hoe-werkt-het

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. https://doi.org/10.2139/ssrn.2477899

Bhoopalam, A. K., Van den Berg, R., Agatz, N., & Chorus, C. G. (2023). The long road to automated trucking: Insights from driver focus groups. Transportation Research Part C: Emerging Technologies, 156, 104351. https://doi.org/10.1016/j.trc.2023.104351

Bureau of Transportation Statistics. (2024). Average age of automobiles and trucks in operation in the United States. https://www.bts.gov/content/average-age-automobiles-and-trucks-operation-united-states

Camden, M. C., Glenn, T. L., Manke, A., & Hanowski, R. J. (2022). Fleet-based driver monitoring systems: Accelerating commercial motor vehicle and occupational driver acceptance of driver-facing cameras (Report No. 22-UI-113). Virginia Tech Transportation Institute: National Surface Transportation Safety Center for Excellence.

Cevolini, A., & Esposito, E. (2020). From pool to profile: Social consequences of algorithmic prediction in insurance. Big Data & Society, 7, 2053951720939228. https://doi.org/10.1177/2053951720939228

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. arXiv. https://doi.org/10.48550/arXiv.2403.04132

Dedhia, Y., Singh, A., Tomar, V. S., Rangaswamy, N., & Thakur, D. S. (2023). Building AI and human capital for road safety. arXiv. https://doi.org/10.48550/arXiv.2312.10319

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.2501.12948

De Winter, J. C. F., Dodou, D., & Eisma, Y. B. (2024a). System 2 thinking in OpenAI's o1-preview model: Near-perfect performance on a mathematics exam. Computers, 13, 278. https://doi.org/10.3390/computers13110278

De Winter, J. C. F., Driessen, T., & Dodou, D. (2024b). The use of ChatGPT for personality research: Administering questionnaires using generated personas. Personality and Individual Differences, 228, 112729. https://doi.org/10.1016/j.paid.2024.112729

De Winter, J. C. F., Driessen, T., Dodou, D., Cannoo, A. (2024c). Truck drivers prefer traditional mirrors over camera systems: A questionnaire study. *Proceedings of the International Co-operation on Theories and Concepts in Traffic Safety*, The Hague, The Netherlands. https://www.ictct.net/wp-content/uploads/36-Hague-2024/ICTCT2024_paper_119.pdf

De Winter, J. C. F., Dodou, D., & Stienen, A. H. A. (2023). ChatGPT in education: Empowering educators through methods for recognition and assessment. Informatics, 10, 87. https://doi.org/10.3390/informatics10040087

de Winter, J. C. F., Hancock, P. A. & Eisma, Y. B. (2025). ChatGPT and academic work: new psychological phenomena. AI & Society. https://doi.org/10.1007/s00146-025-02241-w

Dreger, F. A., De Winter, J. C. F., & Happee, R. (2020). How do drivers merge heavy goods vehicles onto freeways? A semi-structured interview unveiling needs for communication and support. Cognition, Technology & Work, 22, 825–842. https://doi.org/10.1007/s10111-019-00601-3

Golson, J. (2016, October 20). Tesla's new Autopilot will run in 'shadow mode' to prove that it's safer than human driving. The Verge. https://www.theverge.com/2016/10/19/13341194/tesla-autopilot-shadow-mode-autonomous-regulations

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). The Llama 3 herd of models. arXiv. https://doi.org/10.48550/arxiv.2407.21783

Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI (First edition). Random House

Harris, M. (2022, August 4). Tesla's Autopilot depends on a deluge of data: But can a fire-hose approach solve self-driving's biggest problems? IEEE Spectrum. https://spectrum.ieee.org/tesla-autopilot-data-deluge

Hwang, H., Kwon, S., Kim, Y., & Kim, D. (2024). Is it safe to cross? Interpretable risk assessment with GPT-4V for safety-aware street crossing. arXiv. https://doi.org/10.48550/arXiv.2402.06794

Isotupa, K. S., Kelly, M., & Kleffner, A. (2019). Experience-rating mechanisms in auto insurance: Implications for high-risk, low-risk, and novice drivers. North American Actuarial Journal, 23, 395–411. https://doi.org/10.1080/10920277.2019.1572524

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2024). Delving into ChatGPT usage in academic

writing through excess vocabulary. arXiv. https://doi.org/10.48550/arxiv.2406.07016
Kuipers, J. (2022). Automated driving safety data protocol – Practical demonstrators (Technical Report No. ITU-T FGAI4AD-03). International Telecommunication Union.
Latif, E., Zhou, Y., Guo, S., Shi, L., Gao, Y., Nyaaba, M., Bewerdorff, A., Yang, X., & Zhai, X. (2024). Can OpenAI o1 outperform humans in higher-order cognitive thinking? arXiv. https://doi.org/10.48550/arxiv.2412.05753
Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024). What is fair? Proxy discrimination vs. demographic disparities in insurance pricing. Scandinavian Actuarial Journal, 2024(9), 935–970. https://doi.org/10.1080/03461238.2024.2364741
Michon, J. A. (1993). GIDS: Generic Intelligent Driver Support. Available online: https://www.jamichon.nl/jam_writings/1993_car_driver_support.pdf (accessed on 30 January 2025).
Mittelstädt, J. M., Maier, J., Goerke, P., Zinn, F., & Hermes, M. (2024). Large language models can outperform humans in social situational judgments. Scientific Reports, 14, 27449. https://doi.org/10.1038/s41598-024-79048-0
NEXTdriver (2024). Personal assistance for Safe driving behaviour. https://nextdriver.nl/en/
OpenAI. (2024, March 4). Pricing. Wayback Machine. https://web.archive.org/web/20240304150812/https://openai.com/pricing
OpenAI. (2025, January 27). Pricing. Wayback Machine. https://web.archive.org/web/20250127090905/https://openai.com/api/pricing/
Schippers, D., & Stefan, D. (2024a). Can driving data objectify the examiner's verdict? Proceedings of the International Co-operation on Theories and Concepts in Traffic Safety, The Hague, The Netherlands. https://www.ictct.net/wp-content/uploads/36-Hague-2024/ICTCT2024_paper_24.pdf
Schippers, D., & Stefan, D. (2024b). Can driving data objectify the examiner's verdict? [Presentation], 56th CIECA Congress, Dubai, UAE. https://www.cieca.eu/sites/default/files/members-area/General-Assembly-Members-Forum/2024-Congress-Dubai/Congress-7-3-2024/SCHIPPERS%20CIECA_2024.pdf
Souman, J., Adjenughwure, K., Van Dam, E., Van Weperen, M., & Tejada, A. (2021). Quantification of safe driving (TNO 2021 R12632 | 0.1). Helmond, The Netherlands: TNO.
Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society Series C: Applied Statistics, 67, 1275–1304. https://doi.org/10.1111/rssc.12283
Waymo. (2024). Waymo One. https://waymo.com/waymo-one
World Health Organization. (2023). Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517
WSJ News. (2025, January 21). Inside Anthropic's race to build a smarter Claude and human-level AI | WSJ [Video]. YouTube. https://www.youtube.com/watch?v=snkOMOjiVOk?t=1197
Zhang, K., Wang, S., Jia, N., Zhao, L., Han, C., & Li, L. (2024). Integrating visual large language model and reasoning chain for driver behavior analysis and risk assessment. Accident Analysis & Prevention, 198, 107497. https://doi.org/10.1016/j.aap.2024.107497
Zhai, X., Nyaaba, M., & Ma, W. (2024). Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science? Science & Education. https://doi.org/10.1007/s11191-024-00496-1
Zhou, X., & Knoll, A. C. (2024). GPT-4V as traffic assistant: An in-depth look at vision language model on complex traffic events. arXiv. https://doi.org/10.48550/arXiv.2402.02205

Dankwoord

De tijd vliegt! Dit slotwoord markeert het einde van mijn proefschrift en van mijn tijd als promovendus aan de TU Delft. Ik kijk terug op deze periode met veel positiviteit en ik zie het als zeer leerzame jaren, maar vooral ook als een tijd waar ik waardevolle vriendschappen aan over heb gehouden. Ik zou graag de mensen bedanken die hier de afgelopen jaren aan hebben bijgedragen.

Mijn bijzondere dank gaat uit naar mijn promotoren **Joost de Winter**, **Dimitra Dodou** en **Dick de Waard**.

Joost en Dimitra, ik heb onze samenwerking enorm gewaardeerd. Jullie bereikbaarheid, openheid en ontvankelijkheid voor nieuwe ideeën waren essentieel. Ik wil jullie bedanken voor het bewaken van de structuur, maar ook voor de flexibiliteit die jullie toelieten, in plaats van te veel rigiditeit en planning; iets dat wellicht voor meer zekerheid had gezorgd, maar zeker niet zo'n leuk en gevarieerd eindresultaat had opgeleverd.

Dick, zeer veel dank voor al je hulp en feedback. Je uitgezoomde perspectief en nuchtere blik waren voor mij steeds erg nuttig.

I would like to thank **Angèle Picco** for the excellent collaboration, especially during the early days of the project. Angèle, it has been a pleasure working with you and I wish you the best for your own defense.

I would also like to express gratitude to **David, Iva, Lauren** and **Lokin**, for their clever and hard work, and for their dedication while writing their MSc. thesis. I am proud to have supervised them.

Voor de totstandkoming van het onderzoeksproject en beschikbaarheidsstelling van diverse voorzieningen wil ik het **CBR**, **NEXTdriver**, **ABW Verzekeringen** en **TLN bedanken**.

Van het CBR bedank ik alle deelnemende examinatoren en instructeurs aan het interview en de praktijkstudie en daarnaast **Harold Bekhuis, Emy van den Broek-Müller, René Claesen, Daniël Heikoop, Sjoerd Houwing, Eef Jonkers** en **Diana Schippers** voor hun enthousiaste medewerking aan deze projecten.

Van NEXTdriver dank ik in het bijzonder **Jasper Pauwelussen** en **Guido Sluijsmans** voor hun hulp en nauwe betrokkenheid gedurende het project.

Van ABW Verzekeringen gaat in het bijzonder mijn dank uit naar **Jan van Gurp, Ger Lampen, Martin Sterrenburg** en **Marius van der Valk**.

Daarnaast gaat mijn dank uit naar **Gerrit Bosman** van **Bentvelzen Transport & Logistiek** en **Aschwin Cannoo** van **TransportEffect BV & Chauffeursnieuws**. Dank voor de hulp bij het doorgronden van de belevingswereld van Nederlandse vrachtwagenchauffeurs.

Ik wil ook graag een aantal andere mensen bedanken die deze tijd zo bijzonder hebben gemaakt.

Ik dank **Pavlo Bazilinskyy** en **Thomas de Boer** voor hun ondersteuning gedurende verschillende periodes in het project.

Olger, dankjewel voor je gezelschap tijdens onze vele lunches, en richting het einde van mijn PhD-traject voor je hulp en uitstekende samenwerking.

Vishal, dankjewel voor de vriendschap en avonturen die we bijvoorbeeld in Zweden hebben beleefd. Ik ben erg onder de indruk van je performance en storytelling skills, die zelfs indrukwekkend blijven als ik de story volledig kwijt ben.

Wilbert, thank you for the fond memories of New York, where you've shown me the most interesting bars and nearly made me a fan of musicals.

Ook mijn andere collega's van de afdeling Cognitive Robotics met wie ik de kamer heb gedeeld wil ik bedanken voor de fijne sfeer.

Ten slotte wil ik mijn dierbaren bedanken. Mijn vriendin **Julie** bedank ik voor alle steun en liefde. En, tot slot, **Mam**, **Pap**, **Jan** en **Noor**, bedankt dat jullie er altijd voor mij zijn. Jullie zijn een geweldige familie en ik houd erg veel van jullie.

Den Haag, mei 2025

Curriculum Vitae

Tom Driessen

23-09-1993	Born in Leiden, the Netherlands

Education

2020-2024	PhD, Cognitive Robotics, Delft University of Technology
2015-2019	MSc Mechanical Engineering, Delft University of Technology
2011-2015	BSc Mechanical Engineering, Delft University of Technology
2005-2011	VWO, Teylingen College Leeuwenhorst, Noordwijkerhout

Experience

2018	Research Internship, Honda Research Institute, Offenbach, Germany

List of Publications

Journal Publications

1. **Driessen, T.**, Siebinga, O., De Boer, T., Dodou, D., De Waard, D., & De Winter, J. (2024). How AI from automated driving systems can contribute to the assessment of human driving behavior. Robotics, 13(12), 169. https://doi.org/10.3390/robotics13120169 **(Chapter 7)**
2. **Driessen, T.**, Dodou, D., Bazilinskyy, P., & De Winter, J. (2024). Putting ChatGPT vision (GPT-4V) to the test: Risk perception in traffic images. Royal Society Open Science, 11(5), 231676. https://doi.org/10.1098/rsos.231676 **(Chapter 8)**
3. De Winter, **J., Driessen, T.**, & Dodou, D. (2024). The use of ChatGPT for personality research: Administering questionnaires using generated personas. Personality and Individual Differences, 228, 112729. https://doi.org/10.1016/j.paid.2024.112729
4. **Driessen, T.***, Stefan, D.*, Heikoop, D., Dodou, D., & De Winter, J. (2024). Using mobile devices for driving test assessment: A study of acceleration and GPS data. Transportation Letters. https://doi.org/10.1080/19427867.2024.2352198 **(Chapter 5)**
5. **Driessen, T**., Dodou, D., De Waard, D., & De Winter, J. (2024). Predicting Damage Incidents, Fines, and Fuel Consumption from Truck Driver Data: A Study from the Netherlands. Transportation Research Record: Journal of the Transportation Research Board, 2678(7), 1026–1042. https://doi.org/10.1177/03611981231211897 **(Chapter 6)**
6. De Winter, J*., **Driessen, T.***, Dodou, D., & Cannoo, A. (2024). Exploring the challenges faced by Dutch truck drivers in the era of technological advancement. Frontiers in Public Health, 12, 1352979. https://doi.org/10.3389/fpubh.2024.1352979 **(Chapter 3)**
7. **Driessen, T.***, Picco, A.*, Dodou, D., De Waard, D., & De Winter, J. (2021). Driving examiners' views on data-driven assessment of test candidates: An interview study. Transportation Research Part F: Traffic Psychology and Behaviour, 83, 60–79. https://doi.org/10.1016/j.trf.2021.09.021 **(Chapter 2)**
8. Krüger, M.*, **Driessen, T.***, Wiebel-Herboth, C. B., De Winter, J., & Wersing, H. (2020). Feeling uncertain—Effects of a vibrotactile belt that communicates vehicle sensor uncertainty. Information, 11(7), 353. https://doi.org/10.3390/info11070353

Conference Proceedings

9. De Winter, J., **Driessen, T**., Dodou, D., Cannoo, A. (2024). Truck drivers prefer traditional mirrors over camera systems: A questionnaire study. *Proceedings of the International Co-operation on Theories and Concepts in Traffic Safety*, The Hague, The Netherlands. https://www.ictct.net/wp-content/uploads/36-Hague-2024/ICTCT2024_paper_119.pdf
10. **Driessen, T.**, Prasad, L., Bazilinskyy, P., & De Winter, J. (2022). Identifying Lane Changes Automatically using the GPS Sensors of Portable Devices. *13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022).* New York, USA. https://doi.org/10.54941/ahfe1002433 **(Chapter 4)**

Patents

11. **Driessen, T.**, Krüger, M., Wiebel-Herboth, C. (2020). *Method for assisting a person acting in a dynamic environment and corresponding system.* European Patent no. EP3723066A1. Honda Research Institute Europe GmbH.

Datasets and Repositories

12. **Driessen, T.** (2024). Code repository for the paper: How AI from Automated Vehicles Can Contribute to the Assessment of Human Driving Behavior. GitHub. https://github.com/tomdries/AI-driving-assessment
13. De Winter, J. C. F., **Driessen, T.**, & Dodou, D. (2024). The use of ChatGPT for personality research: Administering questionnaires using generated personas [Dataset]. https://doi.org/10.4121/6e0f2f2b-f1fc-4300-b8ca-eb9031a7b257.v1
14. **Driessen, T.***, Stefan, D*., Heikoop, D., Dodou, D., & De Winter, J. C. F. (2024). Using mobile devices for driving test assessment: A study of acceleration and GPS data [Dataset]. https://doi.org/10.4121/3bb2f535-59ec-426c-b69a-e113810543b2.v1
15. **Driessen, T.**, Dodou, D., Bazilinskyy, P., & De Winter, J. C. F. (2024). Putting ChatGPT Vision (GPT-4V) to the test: Risk perception in traffic images. [Dataset]. https://doi.org/10.4121/dfbe6de4-d559-49cd-a7c6-9bebe5d43d50
16. De Winter, J. C. F.*, **Driessen, T.***, Dodou, D., & Cannoo, A. (2024). Exploring the challenges faced by Dutch truck drivers in the era of technological advancement. [Supplementary data]. https://doi.org/10.4121/577c120a-b5bb-4ba5-93b8-6143759d0249
17. **Driessen, T.**, Bindu Prasad, L. L., Bazilinskyy, P., & De Winter, J. (2022). Identifying lane changes automatically using the GPS sensors of portable devices. [Dataset]. https://doi.org/10.4121/19170302

* joint first authors

Propositions

These propositions are regarded as opposable and defendable, and have been approved as such by the promotors Prof. dr. ir. J. C. F. de Winter, Dr. D. Dodou, and Prof. dr. D. de Waard.

1. (this thesis) The biggest drawback of current automated driver assessment is that context awareness is lacking.
2. (this thesis) The planned actions from automated driving systems are suitable reference values for driver assessment.
3. (this thesis) Pay-how-you-drive insurance schemes perpetuate inequality by penalizing driver patterns that are necessitated by environmental conditions.
4. (this thesis) Vision-language models improve driving assessment by providing context-aware risk assessments.
5. No car should be able to drive faster than the local speed limit.
6. The evolution of written human language, once primarily influenced by cultural forces, is now primarily influenced by the design of large language models.
7. The day-to-day motivation of most technically trained people is driven by simple desires, namely diving into technical stuff, trying out fancy new technologies, and satisfying primitive curiosities.
8. Social media creates phantom social pressures, where individuals feel compelled to take positions on issues they rarely encounter in their daily lives.
9. As in basketball (Gilovich et al., 1985), perceived flow states in computer coding are illusions created by variance in outcomes rather than by psychological momentum.
10. Future PhD dissertations will include an ethics statement that no humans were unnecessarily bothered in the data generation or writing process of this research.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6