How should drivers' use of Automated Lane Keeping Systems (ALKS) be assessed? A study with experienced driving assessors in a Wizard-of-Oz vehicle

P. Bazilinskyy, D. Heikoop, R. Verstegen, M. H. Martens, J. C. F. de Winter

Abstract

This study aims to contribute to guidelines for driver licensing organizations on assessing driver competence in using Level 3 Automated Lane Keeping Systems (ALKS), based on an on-road experiment with eight professional driving assessors (i.e., expert driving examiners who train examiner candidates) in a Wizard-of-Oz vehicle. Using a think-aloud protocol, we captured cognitive processes during system supervision and take-over requests (TORs) in real-world traffic jams. An LLM-based thematic analysis of transcripts revealed five themes: (1) Requirement for immediate environmental assessment, (2) Requirement for causal understanding, (3) Requirement for proactive intervention to maintain traffic flow, (4) Requirement for continuous "supervisor" engagement, and (5) Physical ergonomics and mode awareness. These findings indicate that, at least during short-duration usage, drivers do not simply rely on the system to disengage from driving; instead, they maintain active monitoring, physical readiness, and anticipatory skills. These observations blur the distinction between Level 2 and Level 3 automation, as the expert participants in our study generally remained attentive rather than adopting the 'mind-off' state that Level 3 theoretically allows. In conclusion, assessing ALKS usage involves not only evaluating a driver's reaction to a TOR but also judging their performance as a systems manager responsible for anticipating conflicts and smoothly executing control transitions.

Introduction

Car driving may undergo a transformation with the introduction of conditionally automated driving systems. Codified as Level 3 in the SAE International standard J3016 (SAE International, 2021), these systems represent a departure from Level 2 (Partial Driving Automation). While Level 2 systems entail that the driver remains responsible for the safe operation of the driving task throughout, Level 3 systems, such as the Automated Lane Keeping System (ALKS) regulated under UN Regulation 157 (United Nations, 2022), allow the driver to disengage from the driving task and engage in non-driving-related tasks (NDRTs) under predefined circumstances known as the Operational Design Domain (ODD).

Under UN Regulation 157, ALKS operation is restricted to a specific operational design domain (ODD), such as access-controlled roads with physical separation between opposing traffic. While the original regulation limited ALKS operation to 60 km/h, an amendment adopted in 2022 permits operation up to 130 km/h, but only for systems that can perform a lane change during a Minimum Risk Manoeuvre. Although an ALKS is considered a Level 3 system allowing driver disengagement, this does not mean the driver can remain continuously disengaged while the system is in operation (i.e., on a highway during a traffic jam). Regular transitions of control are to be expected; for example, poor weather conditions, road works, tunnels, merges, and night-time all fall outside the ODD, consequently triggering a take-over request (TOR). Additionally, automated systems such as the ALKS are prone to system errors, and human factors such as mode error and complacency increase the frequency of take-over requests or contribute to potential hazards. For example, a hazard analysis of an ALKS identified 87 potential unsafe control actions, stressing the limitations of automated systems (Elizebeth et al., 2025). In the present study, we used a Wizard-of-Oz, ALKS-inspired implementation to study supervision and transitions of control in real congested traffic; the simulation did not enforce all regulatory ODD constraints.

This tendency for automation to shift the operator's role from one of continuous physical control to that of a higher-level supervisor focused on monitoring and decision-making is a long-studied concept in the field of human-vehicle systems (e.g., Johannsen, 1976). This change of role is becoming increasingly pertinent with Level 3 automation in which drivers are granted the freedom to divert their attention away from the road, yet are requested to intervene safely when prompted (e.g., Louw et al., 2015; Saffarian et al., 2012; Heikoop et al., 2019). This shift in the driver's role may require new competencies, such as comprehending system performance and managing control transitions, that are not covered by traditional

training and examination. For a national licensing authority such as the Dutch Centraal Bureau Rijvaardigheidsbewijzen (CBR), in charge of assessing driver skill and fitness to drive, this presents a challenge. Establishing a framework of normative behaviors and assessment criteria for managing the critical aspects of Level 3 automation is required to ensure these new competencies are properly evaluated.

As mentioned above, TORs represent one such critical aspect of Level 3 automation, where drivers must resume manual control from the automation when they are warned by the system that it reaches its operational limits. A substantial number of studies, primarily conducted in simulators, have examined factors influencing driver performance after TORs, including factors such as driver state, type of NDRT, and type of human-machine interaction. A meta-analysis of 129 studies found that shorter take-over response times are associated with higher situational urgency, absence of handheld devices, and multimodal TORs combining auditory or vibrotactile cues with visual ones, while driver age showed no consistent effect (Zhang et al., 2019). Similarly, a systematic review of 36 simulator-based experiments showed that situational urgency, NDRT engagement, and scenario complexity impact performance measures such as take-over response times and lateral/longitudinal control quality (Soares et al., 2021).

A limitation of the existing studies is that most evidence is based on simulator-based environments rather than real-world conditions. The use of simulators is due to the advantages they offer: simulators provide a safe environment to test potentially dangerous scenarios with novel (or non-existing) technology, allow for clear experimental control over variables such as traffic density and TOR timing, and facilitate accurate data collection. However, a limitation of simulators is their limited realism, particularly regarding the variability of naturally occurring road conditions and traffic situations, which cannot easily be programmed a priori into a virtual environment.

A growing body of on-road research is emerging to understand driver behavior when interacting with automation in real-world conditions. Such research reveals various human-machine interaction challenges, including mode confusion after a transition from Level 3 automation (Kim et al., 2025). Other studies have shown that TORs consist of a sequence of actions, influenced by the driver's initial gaze and motoric demands of their NDRT (Berghöfer et al., 2018; Naujoks et al., 2019; Pipkorn et al., 2023). Wizard-of-Oz studies (i.e., studies where a system, in this case a vehicle, appears autonomous but is covertly controlled by a human) have also found that on-road response times were faster than those in simulators (Eriksson et al., 2017) and became faster across multiple rides (Dillmann et al., 2023; Rydström et al., 2023), and that it can take a considerable time of at least 15 seconds for the driver's visual attention to the forward road to return to the levels of normal manual driving (Pipkorn et al., 2024). In addition to Wizard-of-Oz studies, naturalistic driving studies, which observe drivers during their everyday travel (e.g., using Level 2 automation systems), are providing valuable data on the frequency and reasons behind driver-initiated take-overs (Gershon et al., 2021; Schwager et al., 2024; Yang et al., 2023).

The present experiment engaged professional driving assessors as expert participants. These participants possess a deep understanding of safe driving practices and are highly skilled in assessing driver behavior, which makes their insights particularly valuable when evaluating novel in-vehicle technologies (Balassa et al., 2024; Driessen et al., 2021; Driessen et al., 2025). This study used a Wizard-of-Oz approach, where the automated system was simulated to expose participants to realistic take-over scenarios on public roads. In conjunction, a think-aloud protocol was used, which required the participants to verbalize their thoughts. This approach was chosen to gain insight into the cognitive processes and decision-making strategies of these expert drivers.

Methods

Participants

The participant group consisted of eight professional driving assessors employed by the Dutch Central Office of Driving Certification (CBR). All participants possessed extensive experience in assessing driving behavior and training new examiners. They were recruited via professional networks, and informed consent was obtained from each individual prior to the study. Although ten participants originally took part

in the experiment, two participants (1 and 6) were excluded from the analysis due to technical issues with the think-aloud audio recordings. The final sample of 8 participants consisted of 2 females and 6 males. Four of the eight participants reported driving more than 30,000 km per year, and the other four reported driving between 20,000 and 30,000 km per year. Seven participants reported driving daily, and one reported driving 4 to 6 days a week.

Ethical approval was obtained from the institutional review board at Eindhoven University of Technology (TU/e), Reference ERB2024ID465. Participants were debriefed at the end of the session about the details of the Wizard-of-Oz implementation.

Apparatus

The experiment was conducted in a modified Renault Espace configured as a Wizard-of-Oz setup, where a hidden human driver (the real driver) controlled the car while creating the illusion for the participant that the car was operating automatically by means of an ALKS. The participant was located in the rear-left seat behind tinted windows, designed to replicate the driver's position. The participant was isolated from the front cabin by a plywood board and TV screen presenting a live windshield view as if from the driver's position to prevent visibility of the real driver (Detjen et al., 2020; Karjanto et al., 2018; Wang, 2023).

This specific setup of the present experiment included:

- A non-functional "secondary" steering wheel, clutch, and pedals, synchronized via Raspberry Pi and ESP modules to mimic movements from the real steering wheel. Participants engaged/disengaged ALKS or responded to TORs by pressing two white buttons on the steering wheel.
- A TV screen, positioned to simulate the windshield view, displaying a live forward feed from a GoPro HERO 12 camera mounted on the windshield.
- Two 7-inch Full HD monitors simulated side mirrors. These monitors presented the camera input fed by GoPro HERO 13 cameras, which were mounted on the car's exterior mirrors.
- Two more GoPro cameras were mounted inside the vehicle and were used to capture participant behavior and audio.
- A central Surface Pro tablet serving as the human-machine interface. It could display five system states: (1) manual mode (white screen), (2) ALKS ready for activation (purple screen), (3) ALKS active (blue screen), (4) non-urgent take-over request (TOR; orange screen with auditory cue), (5) urgent TOR (red screen with auditory cue), see Figure 1. The design followed ISO standards for automated driving systems (e.g., ISO 2575:2021; International Organization for Standardization, 2021). A Bluetooth keyboard was used by the experimenter to trigger state changes.
- Navigation: A phone with Google Maps in a rear holder for participant visibility, guiding the route. It also displayed vehicle speed and served as a reminder that ALKS does not follow navigation, requiring manual intervention for exits.
- The dashboard of the car was equipped with a display which mirrored the ALKS display available to the participant. This way, the real driver was informed of the ALKS system status.

Figure 1. The five possible human-machine interface states

Figure 2 depicts the experimental setup during a TOR.

Figure 2. The Wizard-of-Oz setup. The figure shows an urgent take-over request (TOR).

Instructions for Participants

Participants were informed via a paper form that the experiment would take place in a Wizard-of-Oz vehicle and were seated in a simulated driver's station in the left-rear of the car. They were told they would be interacting with a simulated ALKS capable of automated driving in traffic jams. Participants were also informed that while the ALKS was active, they were not required to keep their hands on the wheel, feet on the pedals, or their eyes on the road. Their responsibility was to be ready to take over control when prompted by the system. When the vehicle speed dropped below 60 km/h in a traffic jam, the "ALKS ready" screen appeared. Participants were informed that they could choose to activate the ALKS at that point if they wished, by placing both hands on the steering wheel and pressing the top two buttons with their thumbs. A blue screen on a tablet confirmed when the ALKS was active. Participants were also told that they could deactivate the ALKS at any time using the same button press, which returned the system to "manual mode" as indicated by a white screen. The system could issue a TOR indicated by the tablet screen turning orange (non-urgent) or red (urgent) together with an auditory alert. For urgent (red) TORs, participants were instructed to take control immediately. Participants were also told they were responsible for taking over when the navigation system indicated an upcoming highway exit, as the ALKS would only follow the main roadway. They could choose to take control at any other moment they deemed appropriate. A key component of the experiment involved participants verbalizing their thoughts, observations, and decision-making processes, especially before, during, and after taking control.

Experimenter's Role

The experimenter, using the concealed keyboard, controlled the ALKS status according to a predefined set of rules. The study was conducted by four experimenters. When the vehicle entered a traffic jam and its speed dropped to 60 km/h or below, the experimenter activated the ALKS-ready screen, and the participant could activate the ALKS if they wished. If the participant chose to do so by pressing the two top steering wheel buttons with their thumbs, the experimenter switched the display to the blue "ALKS active" mode. The experimenter triggered TORs in specific situations. An urgent TOR (red screen) was issued if the time gap to the lead vehicle increased to over 5 seconds, as estimated by the experimenter. Urgent or non-urgent TORs could also be initiated in response to scenarios that would be challenging for

a real system, such as unusual shapes of nearby vehicles, erratic or dangerous behavior from other drivers, the presence of motorcycles, upcoming or ongoing road works, or approaching tunnels. In the case of tunnels, a non-urgent TOR (orange screen) was given first, escalating to an urgent TOR if the participant did not take control. Additionally, the experimenter could issue urgent or non-urgent TORs at random to simulate unexpected system events. When a take-over occurred, whether initiated by the participant or prompted by a TOR, the experimenter pressed “A” to confirm the switch to manual mode (white screen). The experimenter provided reminders to think-aloud as needed throughout the drive.

Real Driver’s Role

For the purpose of this experiment, a professional driver was hired. This ‘real driver’ was instructed to drive in a manner that mimicked a real ALKS when the system was active. This involved maintaining a speed of no more than 60 km/h, staying within the lane, and keeping an appropriate distance from the car ahead, even if surrounding traffic was faster. Safety was the highest priority, with the driver instructed to deviate from the protocol if any situation became risky. In manual mode, the driver would drive normally to navigate to the next scenario. As mentioned, the real driver was equipped with a special display which mirrored the participant’s ALKS screen so that the real driver could see the ALKS status and act accordingly. To further limit the need for verbal communication with the real driver during the experiment, as much of the route as possible was set before the experiment started, based on the presence of traffic jams.

Procedure

The experiment was conducted from Monday, November 25, 2024, to Friday, November 29, 2024, on roads in the Netherlands with frequent congestion, including highways and, when needed to reach and remain in congestion, selected provincial roads (N-roads). Participants were assigned to either a morning or afternoon session and informed that the entire process would take approximately two to three hours. Before each session, traffic congestion was checked using Google Maps, and the nearest, most suitable highway or provincial road with a traffic jam was selected for the experiment. Upon arrival, participants received an information letter and provided written informed consent before the experiment started. They were then briefed about the overall aim of the experiment.

Before the main experiment began, the drive to the designated road served as a training period. During this phase, the experimenter, seated next to the participant, explained the system’s functions and the different modes shown on the central screen. The experimenter guided the participant through the process of activating and deactivating the ALKS using the two top white buttons on the steering wheel. Participants were also instructed to practice the think-aloud protocol, vocally communicating what they were doing and observing during these simulated transfers of control. Participants were occasionally nudged to speak, especially regarding their thoughts and behaviors regarding a transition of control. This training phase allowed participants to become familiar with the digital displays, and the physical actions required for a safe take-over before encountering a live traffic jam scenario.

The main experiment commenced once the vehicle entered a traffic jam on the highway, with the simulated ODD limited to speeds of 60 km/h or less; the procedures and roles described above were then followed.

Upon completion of the driving session, a debrief was conducted with the participant to discuss their experience.

Analysis

This research focused on the analysis of the think-aloud data. The audio files were transcribed using WhisperV3 (Radford et al., 2023), a state-of-the-art transcription tool that ran locally on a Dell XPS laptop. The transcripts were then automatically stripped of hallucinations (i.e., repetitive, meaningless text) that could occur when neither the experimenter nor the participant was speaking. Subsequently, all recordings were manually reviewed and corrected where necessary. All data falling outside the experimental phase, including the training phase, were removed. Furthermore, names, private details, and conversations about non-driving-related topics were removed, as well as logistical interactions (such as about the route) with the real driver. Only the utterances of the participant and the experimenter were retained.

Next, a thematic analysis was performed on these anonymous transcripts, using a large language model (LLM; Gemini 3.0 Pro Preview; Google DeepMind, 2025) for theme generation and quote selection. The quotes selected by the LLM have all been manually checked against the recordings and adjusted for accuracy where necessary. The following prompt was used:

```
This experiment used a Wizard-of-Oz implementation of the Automated Lane-Keeping System (ALKS,
UNECE Regulation R157). The participant, an experienced driving examiner, sat in the rear
seat, while a confederate driver in the front seat actually controlled the vehicle. A large
forward-view screen gave the participant the illusion of being behind the wheel. A secondary
display showed the ALKS status: Manual mode (white), ALKS ready (purple), ALKS active (blue),
Non-urgent take-over request (TOR) with chime (orange), Urgent TOR with chime (red).

The experimenter encouraged the participant to think aloud and at times discussed ALKS or
broader topics such as the driving test. The participant could engage ALKS or retake control
by pressing two white buttons on the steering wheel.

Using the attached transcripts of eight participants, perform a thematic analysis of "human
requirements during transitions of control". Focus on recurring themes reported by
participants, supported by quotations. Make sure that the quotes are from diverse
participants.
```

Results

Across the eight participants, the mean duration of the training phase (from the start of the drive to the first "ALKS ready" screen) was 11.5 minutes (*SD* = 10.7). The experimental phase, defined as the time from the "ALKS ready" screen to the last moment the ALKS was switched to manual or the end of the recording, averaged 72.6 minutes (*SD* = 32.9). This average includes one trial (Participant 4) that ended prematurely at 11 minutes because of equipment failure.

In total, the eight participants experienced 54 non-urgent (orange) and 34 urgent (red) TORs, four of which were escalations from non-urgent TORs. Additionally, the participants performed 43 discretionary automated-to-manual transitions, i.e., without a TOR. Differences were observed across experimenters and driving conditions: one experimenter (who ran Participants 2 and 3) offered a non-urgent TOR relatively often, while the three other experimenters (who ran Participants 7, 8, 9, & 10) more often escalated the situation (e.g., a large following distance), which led to a high number of discretionary take-overs.

The thematic analysis of their think-aloud reports reveals which cognitive processes, evaluation criteria, and action strategies these experts discussed. Each theme is illustrated with quotations[1].

Theme 1: Requirement for Immediate Environmental Assessment (Visual Scanning)
A dominant requirement reported by participants is the need to instantly re-establish a complete mental model of the vehicle's surroundings upon receiving a take-over request (TOR). Participants emphasized that checking mirrors and blind spots is the immediate physical reaction required before or whilst accepting control, to ensure a "space cushion" exists.

Participant 04: "... yes, I checked both mirrors but I know what is coming because I keep looking in those mirrors... I keep an eye on it so that if [the system] doesn't do it, I do it."

[1]Participants were labeled 1–10 during data collection; after excluding Participants 1 and 6, the analyzed sample comprised Participants 2, 3, 4, 5, 7, 8, 9, and 10. One quote was removed because it pertained to the experimenter rather than the participant. As a robustness check, the same prompt was also tried for other large language models.
Although the level of granularity differed , the main results were similar to those obtained with Gemini 3.0 Pro Preview:
Gemini 2.5 Pro (September 2025): 1) Continuous Situational Awareness as a Prerequisite for Trust and Action, (2) The Need for System Transparency and Predictability, (3) Proactive Manual Control to Ensure Safety and Traffic Flow, (4) The Take-Over Action as an Immediate, Reflexive Response
SuperGrok Expert (September 2025): 1) Sustained Vigilance and Environmental Scanning, 2) Anticipation and Proactive Decision-Making, 3) Physical Readiness and Actions, 4) Trust Calibration with the System.
ChatGPT 5 Thinking (September 2025): 1) Continuous 360° situational awareness—especially at (and just before) TOR, 2) Clear, timely, and trustworthy handover cues + responsive controls, 3) A need to understand why a TOR happens, 4) Agency to preserve traffic flow and visibility (proactive takeovers), 5) Managing merges and "space-cushion courtesy" 6) Early TORs for special actors/situations (motorbikes, emergency lanes, tapers, tunnels), 7) Maintain readiness: posture, hands, glance strategy, 8) Calibrated trust without complacency; secondary-task boundaries.

Participant 05: "It's almost an automatism now, because it is immediately taking over and immediately checking the mirrors again." / "Suppose a takeover comes, then you must of course... know your space cushion, your free spaces and your escape possibilities."
Experimenter: "And during a takeover moment, where do you look?" **Participant 07:** "...quickly the 360-degree check, hey. Your mirror left and right. Suppose I have to brake abruptly, isn't there someone on my bumper...?"
Participant 03: "But, this situation, the simulation system, tempts me to not look around... but I should continue to watch around the car."

Theme 2: Requirement for Causal Understanding (The "Why?")
Participants frequently expressed a requirement to understand *why* the system was requesting a takeover. When the cause was not visible (e.g., no immediate danger), it caused confusion. The human requirement here is for the system to communicate the rationale for the transition so the driver can prioritize their attention (e.g., looking for a hazard vs. just resuming manual driving due to a geofence ending).

Participant 04: "He asked to take it over, that I take it over again, but I do not know why, because we are just staying in the line. We are also not yet driving 60 plus, so that would not be the reason either. So, anyway, it had to be taken over, so fine."
Participant 05: "I was surprised that he said "takeover", while there was actually not much going on. So maybe I missed something... It could be that he reacts to that [motorcyclist]."
Participant 07: "During a takeover moment then you also quickly look for the cause. Why does he ask that?"
Participant 07: "If that red image of 'turn off ALKS', then you immediately start looking: hey, how, what, what is the cause? What am I doing wrong?" / "In my experience there was nothing wrong, why should I have to take over? But probably it was the traffic lights, I think."

Theme 3: Requirement for Proactive Intervention (Traffic Flow & Etiquette)
A distinct theme among these expert drivers was the requirement to take over control *voluntarily* (driver-initiated transition) to maintain social traffic flow. The system was often viewed as too defensive, slow, or socially awkward (e.g., leaving large gaps), requiring the human to step in to prevent irritation from other road users or to facilitate merging.

Participant 08: "If I were driving here with candidates, I would hope they anticipate in time to move to the left lane. That they take over control themselves."
Participant 09: "I take over... , because otherwise you don't flow well with the other traffic" / "Let's see, I would overtake here for the flow of traffic, so I would move one lane to the left to get back up to the speed limit."
Participant 10: "It is the question, though, if the gap with the front is going to be closed in on fast enough... I take it over myself, that is undesirable. It is also not always understood [by other drivers], why does he keep driving so slowly."

Theme 4: Requirement for Continuous "Supervisor" Engagement
Despite the system being automated, the participants rejected the idea of "eyes off/mind off." They highlighted a human requirement to remain mentally "in the loop" to facilitate a safe transition. They viewed the transition not as a cold start, but as shifting from a supervisor role to an active operator role.

Participant 05: "Somehow I am still observing what is happening around the car... I trust that he stays within his lane of course... [but] somehow you keep watching."
Experimenter: "And you still like to check the mirrors every once in a while, even when the system is on?" **Participant 02:** "Yes."
Participant 07: "So that ensures for me that I still remain alert." / "I feel myself on the edge of the takeover. ...."
Participant 10: "Hands-off, brain-off; for me that doesn't work... Still trying to get that information you would need as a driver yourself."

Theme 5: Physical Ergonomics and Mode Awareness
There was a functional requirement regarding the physical act of the transition—knowing the system state (colors) and the physical interaction with the buttons/wheel. Participants needed to confirm the mode change was successful.

Participant 02: "I pushed the buttons too soon, when the screen is changing, then it seems the system doesn't react... Next time I will hold a second."
Participant 05: "I saw it light up red, so I went actually automatically to those two white buttons. So yes. It went actually by itself, despite that I maybe was not 1, 2, 3 prepared for it, the little beep signal in red still triggers a sort of alarm system in your body, huh, because you must now really take action."
Participant 10: "Just a quick check like: am I driving myself or is it on? It happens unconsciously. Sort of like: oh right, I don't need to take action."

Taken together, these five themes depict the participants as active supervisors rather than passive occupants when overseeing a Level 3 system on public roads. Participants prioritized immediate visual scanning (Theme 1) to ensure a safety buffer and sought a causal understanding of the system's behavior (Theme 2). They also engaged in proactive manual control (Theme 3) to maintain traffic flow and social etiquette. Furthermore, participants maintained continuous mental engagement (Theme 4) and emphasized the importance of physical ergonomics and clear mode awareness (Theme 5) during the transition. Together, these patterns reveal that safe ALKS supervision by our participants was not a matter of “waiting for instructions”, but a process of monitoring, interpreting, and managing both the system and the driving environment.

Discussion

The aim of this study was to examine how professional driver assessors manage the transition from automated to manual control in a Wizard-of-Oz setup of an ALKS. By hiding a human confederate driver and offering participants a relatively realistic driving experience, we could measure the cognitive and behavioral strategies that experts use when they need to resume driving. This allowed us to move beyond earlier simulator-based and test track studies that typically used scripted scenarios, and provide ecologically-valid insights from professionals who routinely judge safety on public roads.

Our LLM-based thematic analysis of the think-aloud data showed that the participants did not treat the take-over as a mere button-press but rather as a rapid yet multi-faceted activity. This activity comprises immediate environmental scanning, a search for causal understanding of the system’s behavior, proactive interventions to maintain social traffic flow, and a continuous mental engagement as a supervisor. Furthermore, the analysis highlighted the critical role of physical ergonomics and mode awareness during the transition. In other words, safe ALKS supervision hinges not just on raw reaction time but also on the driver’s ability to combine attentive monitoring, causal reasoning, physical readiness, and decisive action into one routine, i.e., different information-processing stages are involved (cf. Parasuraman et al., 2000). These are skills that current training regimes rarely teach and that future certification tests have to explicitly assess.

An implication of these findings is that, in this short-duration on-road Wizard-of-Oz setting with a concurrent think-aloud task, the expert drivers behaved in a manner more consistent with Level 2-style supervision (i.e., sustained monitoring and readiness) than with the “mind-off” posture that Level 3 systems theoretically permit within their ODD. Although there were a small number of exceptions, such as a participant briefly checking his mobile phone when ALKS was active, overall, the proactive monitoring and discretionary interventions show that participants did not adopt a true ‘mind-off’ state but rather remained cognitively engaged, even if their hands and feet were off the controls. This behavior challenges the practical distinction between Level 2 and Level 3 automation when it comes to short-term driving sessions in which drivers are assessed. This finding resonates with another on-road study where the ambiguity between automation levels led to mode confusion, such as drivers unnecessarily keeping their hands on the wheel during Level 3 operation or, conversely, incorrectly removing them during Level 2 (Kim et al., 2025).

This study has several limitations. One limitation is the small sample size of eight participants. However, this may be less concerning than in quantitative research for two reasons. First, thematic saturation (i.e., the point at which new data no longer yield new themes) can be reached with small and homogeneous samples (Hennink & Kaiser, 2022). Second, we recruited experienced driving assessors as a specific sample of experts; their professional training likely results in less variable behavioral patterns compared to the general population. A second limitation is that the Wizard-of-Oz method meant participants were aware they were in an experimental simulation, and the think-aloud protocol itself may have prompted participants to be more analytical than they would be in a naturalistic driving situation (Dahlbäck et al., 1993). Additionally, simulating a real ALKS in a consistent manner proved difficult in this regard, for example, regarding the criteria for triggering a non-urgent TOR. Because the real driver had to search for congestion, he sometimes deviated from the planned route and encountered features such as traffic lights, which characterize a road type where ALKS would normally not operate. However, these interruptions reflect the reality of current Level 3 systems, which rarely permit prolonged ‘mind-off’ periods due to frequent warnings and disengagement. Future research should validate our findings with larger

populations of expert and non-expert drivers. Ideally, commercially available Level 3 vehicles should be used to confirm the behaviors in a non-experimental context and over longer periods of time.

In conclusion, this study found that expert examiners supervising a simulated Level 3 automation system (ALKS) maintained situational awareness, akin to supervising a Level 2 system. What do these findings imply for the examination of candidates on the driving test? The experts' cautious approach (essentially treating the ALKS as a Level 2 system) may not reflect the behavior of average drivers prone to over-trust (Wintersberger & Riener, 2016). We argue that the experts' behavior reveals a principle for driver assessment: even if the ALKS technology permits driver disengagement, the licensing of the driver should require a demonstration of more foundational skills first. That is, although a Level 3 system is engineered to be responsible for the driving task within its ODD, the assessment of a driver's competence should follow a tiered approach. An effective way to verify that a driving test candidate possesses the situational awareness and vehicle control skills for this transition is to require them to demonstrate these skills during the test. A parallel can be drawn from aviation: pilots must first demonstrate mastery of manual flight to earn their foundational pilot certificate, which is a prerequisite for being trained to safely manage the advanced autopilot systems in more complex aircraft (Federal Aviation Administration, 2021). Following this logic, demonstrating the competence of an *active* supervisor should be a prerequisite for earning the privilege to use a Level 3 system that allows for disengagement. Further research is needed to establish how, and to what degree, testing agencies should assess a driver's skills in continuous monitoring, causal reasoning, physical interaction with the controls, and decisive action.

Acknowledgements

This research has been made possible by Eigenchauffeur.nl, which provided the expert driver for the field tests. Also, thanks go out to Ilse Harms and Boris van Waterschoot, who assisted in the inception of the idea of this experiment. We express gratitude to Aloysia Prakoso for creating the interface. A final gratitude is towards all the assessors who volunteered in participating in this study, as their expert input provided invaluable insights into the take-over process in an automated vehicle.

Supplementary material

The Unity project used to run the human-machine interface is available at https://www.dropbox.com/scl/fo/cojlmto3wlzpiih0omtp8/ALEX3oPYuRPkftE1of8AvY4?rlkey=boggm3hvva7o23g1cql86aqn7.

References

Balassa, B. E., & Koteczki, R. (2024). Driving examiners' perceptions and awareness of Advanced Driver Assistance Systems: A survey-based analysis. *Engineering Proceedings*, *79*(1), Article 21. https://doi.org/10.3390/engproc2024079021

Berghöfer, F. L., Purucker, C., Naujoks, F., Wiedemann, K., & Marberger, C. (2018). Prediction of take-over time demand in conditionally automated driving - results of a real world driving study. In D. de Waard, K. Brookhuis, D. Coelho, S. Fairclough, D. Manzey, A. Naumann, L. Onnasch, S. Rottger, A. Toffetti, & R. Wiczorek (Eds.), *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference* (pp. 69–81). https://www.hfes-europe.org/wp-content/uploads/2018/10/Purucker2018.pdf

Dahlbäck, N., Jönssen, A., & Ahrenberg, L. (1993). Wizard of Oz studies - why and how. *Proceedings of the 1st International Conference on Intelligent User Interfaces '93,* 193–200. https://doi.org/10.1145/169891.169968

De Winter, J. C. F., Van Leeuwen, P. M., & Happee, R. (2012). Advantages and disadvantages of driving simulators: A discussion. *Proceedings of the Measuring Behavior Conference*, Utrecht, The Netherlands, 47–50. https://repository.tudelft.nl/file/File_f60633b3-f70c-4154-85dd-be119b5c4617

Detjen, H., Pfleging, B., & Schneegass, S. (2020). A Wizard of Oz field study to understand non-driving-related activities, trust, and acceptance of automated vehicles. *Proceedings AutomotiveUI*

‘20: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 19–29. https://doi.org/10.1145/3409120.3410662

Dillmann, J., Den Hartigh, R. J. R., Kurpiers, C. M., Raisch, F. K., Kadrileev, N., Cox, R. F. A., & De Waard, D. (2023). Repeated conditionally automated driving on the road: How do drivers leave the loop over time? *Accident Analysis & Prevention*, *181*, Article 106927. https://doi.org/10.1016/j.aap.2022.106927

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

Driessen, T., Stefan, D., Heikoop, D., Dodou, D., & de Winter, J. (2025). Using mobile devices for driving test assessment: a study of acceleration and GPS data. *Transportation Letters, 17*, 384–394. https://doi.org/10.1080/19427867.2024.2352198

Elizebeth, M. J., Khastgir, S., & Jennings, P. (2025). Hazard analysis of an Automated Lane Keeping System using Systems-Theoretic Process Analysis. *Accident Analysis & Prevention, 221,* Article 108171, https://doi.org/10.1016/j.aap.2025.108171

Eriksson, A., Banks, V. A., & Stanton, N. A. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident Analysis & Prevention*, *102*, 227–234. https://doi.org/10.1016/j.aap.2017.03.011

Federal Aviation Administration. (2021). Airplane flying handbook (Report No. FAA-H-8083-3C). U.S. Department of Transportation. https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/airplane_handbook/

Gershon, P., Seaman, S., Mehler, B., Reimer, B., & Coughlin, J. (2021). Driver behavior and the use of automation in real-world driving. *Accident Analysis & Prevention*, *158*, Article 106217. https://doi.org/10.1016/j.aap.2021.106217

Google DeepMind. (2025). Gemini 3 Pro. https://deepmind.google/models/gemini/pro

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & Van Arem, B. (2019). Human behaviour with automated driving systems: a quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science, 20*, 711–730. https://doi.org/10.1080/1463922X.2019.1574931

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, *292*, 114523. https://doi.org/10.1016/j.socscimed.2021.114523

International Organization for Standardization. (2021). *Road vehicles — Symbols for controls, indicators and tell-tales* (ISO 2575:2021). https://www.iso.org/standard/68409.html

Johannsen, G. (1976). Preview of man-vehicle control session. In T. B. Sheridan & G. Johannsen (Eds.), *Monitoring behavior and supervisory control* (pp. 3–12). Plenum Press.

Karjanto, J., Yusof, N. M., Terken, J., Delbressine, F., Rauterberg, M., & Hassan, M. Z. (2018). Development of on-road automated vehicle simulator for motion sickness studies. *International Journal of Driving Science*, *1*(1), Article 2. https://doi.org/10.5334/ijds.8

Kim, S., Novakazi, F., & Karlsson, I. C. M. (2025). Is conditionally automated driving a bad idea? Observations from an on-road study in automated vehicles with multiple levels of driving automation. *Applied Ergonomics, 129*, 104617. https://doi.org/10.1016/j.apergo.2025.104617

Louw, T. L., Merat, N., & Jamson, A. H. (2015). Engaging with highly automated driving: To be or not to be in the loop? *8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* Salt Lake City, UT. https://doi.org/10.17077/drivingassessment.1570

Naujoks, F., Purucker, C., Wiedemann, K., & Marberger, C. (2019). Noncritical state transitions during conditionally automated driving on German freeways: Effects of non–driving related tasks on takeover time and takeover quality. *Human Factors*, *61*(4), 596–613. https://doi.org/10.1177/0018720818824002

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30*, 286–297. https://doi.org/10.1109/3468.844354

Pipkorn, L., Dozza, M., & Tivesten, E. (2024). Driver visual attention before and after take-over requests during automated driving on public roads. *Human Factors*, *66*(2), 336–347. https://doi.org/10.1177/00187208221093863

Pipkorn, L., Tivesten, E., Flannagan, C., & Dozza, M. (2023). Driver response to take-over requests in real traffic. *IEEE Transactions on Human-Machine Systems*, *53*(5), 823–833. https://doi.org/10.1109/THMS.2023.3304003

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518. https://proceedings.mlr.press/v202/radford23a/radford23a.pdf

Rydström, A., Mullaart, M. S., Novakazi, F., Johansson, M., & Eriksson, A. (2023). Drivers' performance in non-critical take-overs from an automated driving system—An on-road study. *Human Factors*, 65(8), 1841–1857. https://doi.org/10.1177/00187208211053460

Saffarian, M., De Winter, J. C., & Happee, R. (2012). Automated driving: human-factors issues and design solutions. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56, 2296–2300. https://doi.org/10.1177/1071181312561483

SAE International. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (SAE Standard J3016_202104, Rev. Apr. 2021). https://www.sae.org/standards/content/j3016_202104

Schwager, R., Grimm, M., Ewecker, L., Bruehl, T., Sohn, T. S., & Hohmann, S. (2024). An analysis of driver-initiated takeovers during assisted driving and their effect on driver satisfaction. *Proceedings of the 2024 IEEE Intelligent Vehicles Symposium*, Jeju Island, Republic of Korea, 1907–1914. https://doi.org/10.1109/IV55156.2024.10588585

Soares, S., Lobo, A., Ferreira, S., Cunha, L., Couto, A., & Coelho, M. C. (2021). Takeover performance evaluation using driving simulation: A systematic review and meta-analysis. *European Transport Research Review*, *13*(1), Article 47. https://doi.org/10.1186/s12544-021-00505-2

United Nations. (2022). Proposal for the 01 series of amendments to UN Regulation No. 157 (Automated Lane Keeping Systems). (ECE/TRANS/WP.29/2022/59/Rev.1). Retrieved from https://unece.org/sites/default/files/2025-03/ECE_TRANS_WP.29_2022_59_Rev.1e.pdf

Wang, Y. (2023). *A real road study of automated driving. The influence of the non-critical cue on automation surprise* (MSc thesis, Eindhoven University of Technology). https://pure.tue.nl/ws/portalfiles/portal/282437684/1560034_Master_Thesis_Report.pdf

Wintersberger, P., & Riener, A. (2016). Trust in technology as a safety aspect in highly automated driving. *i-com*, *15*(3), 297-310. https://doi.org/10.1515/icom-2016-0034

Yang, S., McKerral, A., Mulhall, M. D., Lenné, M. G., Gershon, P., & Reimer, B. (2023). Takeover context matters: Characterising context of takeover in naturalistic driving using Super Cruise and Autopilot. *Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Ingolstadt, Germany, 112–122. https://doi.org/10.1145/3580585.3606459

Zhang, B., De Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 285–307. https://doi.org/10.1016/j.trf.2019.04.020