



Utrecht University
Artificial Intelligence Master Thesis

How accurate can we simulate human multitasking?
*Modeling texting while driving using a computational
cognitive architecture*

Florian Fikkert - 6976298

Daily Supervisor: Dr. Chao Zhang

First Examiner: Dr. Shihan Wang

Second Examiner: Dr. Leendert van Maanen

Abstract

This thesis aimed to investigate whether it is possible to model a relatively new and complex phenomenon in distracted driving, namely texting while driving. This with the goal to ultimately be able to generate simulated human distracted driving data for the development of more advanced driving assistance systems. Two computational cognitive models were combined, one to mimic texting and another one to simulate human driving behavior. Four different interleaving strategies were implemented and the resulting secondary task times together with the simulated driving performance in terms of lateral deviation were finally compared with empirical data, which was gathered under the same distracted driving circumstances. By qualitatively analyzing the results, we could make the case that at least two interleaving strategies could be omitted, as these were furthest away from the empirical data. For the other two remaining interleaving strategies, we found that they did overlap with the empirical data reasonably well, which is promising for future research in simulating more complex secondary tasks using cognitive architectures for data generation.

Table of Contents

ABSTRACT	2
1: INTRODUCTION	5
1.1 Theory.....	6
1.2 Related Work.....	7
1.2.1 Integrated Driving Model.....	7
1.2.2 Driving and Distraction Models.....	8
1.3 Objective.....	9
2: METHODS	11
2.1 Modeling and Simulation.....	11
2.1.1 Model implementation.....	12
2.1.2 Environment implementation.....	17
2.2 Empirical Study.....	19
2.2.1 Participants.....	19
2.2.2 Materials.....	20
2.2.3 Design.....	23
2.2.4 Procedure.....	24
3: RESULTS	26
3.1 Modeling and Simulation.....	26
3.1.1 Task time analysis.....	27
3.1.2 Lateral deviation analysis.....	29
3.2 Empirical Study.....	32
3.2.1 Task time analysis.....	32
3.2.2 Lateral deviation analysis.....	35
3.3 Data Comparison.....	38
3.3.1 Task time.....	38
3.3.2 Lateral deviation.....	40
4: GENERAL DISCUSSION	42
4.1 Task time.....	42
4.2 Driving performance.....	43

4.3 Practical Implications.....	44
4.4 Limitations and Future Work	45
4.5 Remarks.....	45
REFERENCES	47
APPENDIX	56
Informed Consent.....	56
Information Letter.....	57

1: Introduction

The exercise of driving a car is a complex task which consists of multiple higher-level interacting cognitive processes. It involves perception, attention, learning, memory, decision making, and action control (Groeger, 2000). Even though specific tasks can be automated by implementing sophisticated systems within the field of artificial intelligence, fully autonomous vehicles co-existing with human-controlled vehicles remain largely unsolved. This is mainly constrained due to the fact that driving itself requires a fairly deep level of understanding of the world and its complex interactions. Furthermore, if we were to ignore the technical limitations, there are ethical and regulatory issues to be solved prior to autonomous vehicles sharing the road with conventional vehicles (Martínez-Díaz & Soriguera, 2018). However, this does not imply that no autonomy-related technological advances within the automotive sector found their way into implemented applications.

Over time, to improve safety and driving comfort, cars have been progressively equipped with different forms of driver assistance tools, also known as advanced driver assistance systems (ADAS). These systems technically surpass the more conventional assistance systems (e.g., cruise control) and come with their own set of technical challenges. An earlier extension to the more conventional systems is known as adaptive/active cruise control (ACC), which provides automatic control of the longitudinal position of the car. A more modern implementation of this system combines both longitudinal and lateral control of the vehicle, which enables (semi-) autonomous driving in certain scenarios (Khodayari et al., 2010).

Especially nowadays, with potentially distracting electronic devices being present both in the car and in the vicinity of the driver, the chances of the driver engaging in risky multitasking behavior are greatly increased. From these electronic device interactions, cell phone usage remains one of the largest causes of accidents (CDC, 2020). In these dangerous situations, an advanced driver assistance system could provide a safety net, however, as is explained in the next paragraph this causes a new set of problems.

When implemented in a hybrid setting, advanced driver assistance systems are commonly activated when either the lane deviation or the distance to the car ahead is exceeding a certain threshold. Implicitly, the assumption is made that this will need to happen when the driver is distracted, which in turn causes the system to intervene. However, these techniques can be perceived as limiting to the user's perceived autonomy when the user is situational aware, thus being able to predict and track the attention of the driver is of importance (Pohl et al., 2007). There have been different techniques proposed in combination with ADAS that try to infer the driver's attention through sensors (Minoiu Enache et al., 2009). Another method of determining distraction is by measuring the car's absolute lane deviation (Choudhary & Velaga, 2017). What these methods have in common is that they are mostly based on a given set of rules. For instance, the absolute lane deviation is taken into account whether or not the system should intervene, without knowledge of the driver's situational awareness. Therefore, it would be beneficial to be able to

predict the situational awareness of the driver in a non-intrusive manner, which could provide a higher feeling of autonomy while driving a vehicle. In this way, the driver still has full control of the vehicle, while having a safety net available when the driver happens to be not fully situationally aware. In essence, this means that we aim to have a better understanding of when secondary tasks are being performed and what impact this has on the driving performance to predict the possibility of a potentially dangerous situation.

One way of achieving this is through having access to human driving data. For such systems to be able to learn how humans drive, often large quantities of real human driving data are needed (Bhattacharyya et al., 2020; Z. Huang et al., 2021). There are datasets available for basic driving scenarios, such as the NGSIM highway driving dataset that for example is being used for training on lane change behaviors for advanced driver assistance systems (L. Huang et al., 2018). However, when we consider the wide variety of possible different distractions; if we want a system to learn how a particular specific task interferes with the driving performance, often that data is not readily available. For example in a study on the effect of smartphone usage, this data still has to be empirically collected (Khan et al., 2021).

A possible solution for generating large quantities of human driving data, where the driver is engaging in a particular secondary task, would be through the utilization of cognitive models to simulate this data; as cognitive models aim to capture human behavior. In this thesis, we will model one of these potentially dangerous distractions and try to derive whether the resulting driving behavior data would potentially be useful for the eventual training of systems, such as advanced driver assistance systems.

1.1 Theory

Cognitive modeling is an area of artificial intelligence that aims to simulate human mental processes in order to predict or understand human decision-making and performance during certain tasks. Within cognitive modeling, three different subclasses of cognitive models can be distinguished; computational, mathematical and verbal-conceptual models (Bechtel & Graham, 1999). Since computational cognitive models recently have been considered the most promising and flexible category for embedding cognitive theories for practical applications (Sun, 2008), the focus will be solely on computational cognitive models.

Since its inception, there has been an intent to simulate the human mind within an all-encompassing unified theory (Newell, 1990). Rather than only being able to simulate narrow processes, there was the need to have a more generalized theory of cognition. This eventually resulted in the introduction of cognitive architectures such as SOAR, ACT-R, EPIC and others (Oulasvirta et al., 2018). These cognitive architectures embody a theory of mind (e.g., memory, vision, motor action), which is, in turn, built upon theories from cognitive psychology. Cognitive architectures aim to provide an algorithmic basis for simulating a broad range of different processes within the human mind. They offer a framework that has built-in constraints to closely emulate real-world tasks, to either examine the processes involved or predict certain outcomes.

An active area of research has been the implementation of multitasking within said architectures. While singular tasks are easier to model, it has been proven more difficult to model the switching between two separate tasks. Within the context of the ACT-R architecture, however, a general theory of multitasking is implemented, also known as threaded cognition (Salvucci & Taatgen, 2008, 2011). Threaded cognition is an extension to the architecture, such that it is possible to have concurrent tasks (or goals) active at once, while still being constrained within the predefined processing capabilities of the cognitive architecture. We are interested in one particular application of such models, which is the modeling of human car driving with secondary tasks that lead to the driver being distracted, which will be elaborated on in the next section.

1.2 Related Work

1.2.1 Integrated Driving Model

As mentioned in the previous section, since driving is a complex task, it has been proven challenging to accurately model human driving behavior. Early attempts mostly focused on one particular part of driving itself, but there was a need for a more fully-fledged integrated driving model (Salvucci et al., 2001). This soon resulted in a theory on how human steering occurs, the so-called two-point steering model (Salvucci & Gray, 2004). With these advancements, it did not take long before a fully integrated driving model was implemented. Unsurprisingly, the first implementations were achieved within the Adaptive Control of Thought-Rational (ACT-R) cognitive architecture (Anderson et al., 2004). Since the first integrated driving models were implemented, there has been steady progress in extending the computational driving models within ACT-R and its various iterations.

One of the first rigorous computational models of driving behavior is by the works of Salvucci (2006). This model accounts for steering profiles, lateral position profiles and gaze distributions of human drivers during lane keeping and lane changing. Follow-up work shows that predictions considering lane change can be made using the aforementioned models. An example of this is shown in the study: “Lane-Change detection Using a Computational Driver” (Salvucci et al., 2007). In this study, the intention to lane-change in human participants could be predicted with significant accuracy, by using the trained data from the cognitive driving model. As the paper points out, this method is promising for applications within the smart (i.e., adaptive) driving assistance systems.

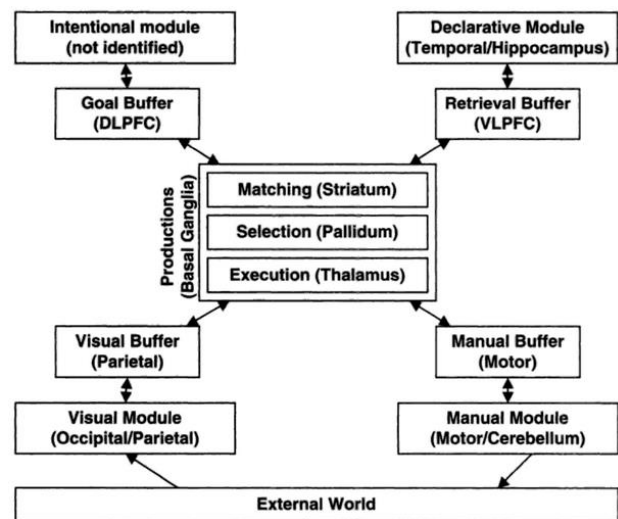


Figure 1.1 ACT-R 6.0 (Sun, 2006)

Moreover, different additions have been implemented within the integrated driving model. One of those studies extended the model with an attention model for simple driving tasks, such as identifying crossroads and traffic signs (K. Haring et al., 2012). Other extensions include a variety of other more complex tasks, such as reading signs and handling different situations that emerge at crossroads (Deng et al., 2017; K. S. Haring et al., 2012).

1.2.2 Driving and Distraction Models

Throughout the years, numerous studies have been conducted to gain deeper insight into the attention and awareness of human drivers (Chakraborty & Nakano, 2016; Choudhary & Velaga, 2019; Oviedo-Trespalacios et al., 2016; Pekkanen et al., 2018; Stothart et al., 2015). Especially in a world where distractions in the form of smartphones, navigational guidance and car interfaces have become ubiquitous, these areas of research have become increasingly important. An insightful way of conducting research within the field has been to develop cognitive models, which enables researchers to better understand the underlying processes that take place whilst controlling a vehicle. Since the topic itself is fairly broad, we will solely focus on models that focus on modeling either multitasking or attention in a driving setting.

For example, the paper “Modeling Driver Distraction from Cognitive Tasks” (Salvucci, 2002) where the first attempt is being made at predicting driving performance by using a computational cognitive model within the ACT-R architecture. In this study, a secondary task is modeled, known as the “Sentence-Span Task”, which involves the processing of sentences and the recall of words in these sentences (Alm & Nilsson, 1995). In this study, it was shown that the model was able to qualitatively predict multiple measures of driver performance, reported in an earlier empirical study.

In similar research, the effect of cell-phone dialing was modeled, to again predict the resulting driving behavior and performance (Salvucci & Macuga, 2002). This study focused more on employing a secondary motoric task in a naturalistic driving setting. In this case, four different models and strategies were used to validate the model. The baseline driving performance itself was also compared to the empirical data. This study showed that using an integrated driving model and secondary tasks within ACT-R enables revealing behavioral patterns and the resulting driving performance.

Follow-up research was also conducted to further generalize these findings to other interfaces rather than cell phones. A rapid prototyping application was created to evaluate in-vehicle interfaces, known as Distract-R (Salvucci et al., 2005). With this study, we see the first attempt at creating less-distracting interfaces, solely based on the predicted driving performance that results from a computational cognitive model. This application made it possible to use the whole range of modalities within ACT-R for the rapid prototypes, such as speech and sound, which shows the versatility of using a cognitive architecture.

As distractions have become more complex than merely entering a phone number or pressing a button, research within the cognitive models has been striving to model these more complex tasks as well. In one of those studies, searching for a song in an interface while driving was implemented (Kujala & Salvucci, 2015). It is interesting to see that in this study, assumptions are built into the model as in where and when people are expected to interleave. These so-called natural breakpoints are of importance when secondary tasks are performed while driving (Janssen et al., 2012).

Recently, research on multitasking and distractions while driving has received a new influx since the conception of the Queuing Network ACT-R (QN-ACT-R) framework, which greatly improves usability and is more versatile than the older LISP-based ACT-R frameworks. Multi-tasking performance is also improved with the addition of the queuing network (Cao, 2013; Cao & Liu, 2011; Liu, 2009). The framework itself does not change the definition of the models but is more sophisticated in its implementation.

Other influences on driving behavior have been modeled using this framework, such as the effect of limited sight distance through fog on car-following performance (Deng et al., 2018). Also, the take-over reaction was modeled for a semi-autonomous car driving scenario where an emergency occurs, with its resulting driving behavior (Deng et al., 2019).

1.3 Objective

We could argue that the techniques for detecting distractions, as described in the previous section, are based upon the *resulting* driver behavior whilst being distracted. The mental processes of *being* distracted are not directly taken into account, solely the resulting behavior from it. Therefore, one could argue that cognitive models, at least in theory, could better explain the inner workings of the distraction itself. This could, in turn, lead to a better understanding of the resulting behavior from *different* kinds of distractions, as these can be individually modeled. As we will see in the related works section, this idea is not new but remains fairly complicated especially with the more complex secondary tasks we encounter today.

These requirements open the door for creating a cognitive model that aims to simulate distracted driving behavior, which as earlier stated, could potentially reduce the need for human participants in the process of gathering such data. Consequentially, this could be beneficial to potentially train an artificial intelligence system in the future, to classify whether an actual human is being distracted and what task is being performed.

It would also be interesting to see whether computational cognitive models can actually approximate the resulting driving behavior with a not previously modeled more complex distraction, which perhaps could eventually be easily altered to include a wider range of different cognitive tasks.

With these two objectives in mind, we can formulate the following two research questions:

RQ 1: Can we realistically simulate driving performance with a periodically occurring complex secondary task within an existing cognitive framework?

RQ 2: Is the resulting simulated driving behavior usable for approximating real human distractions within the same task environment?

Considering earlier work on multitasking while driving, we would preferably want to model a secondary task that is relevant to modern-day distracted driving. Since texting while driving is one of the most dangerous types of distracted driving nowadays (Foreman et al., 2021), it would be logical to model this rather complex task. Another reason for modeling this particular secondary task would be the challenge of having an unknown interleaving strategy, thus assumptions have to be made on where task-switching occurs such as with the song searching task (Kujala & Salvucci, 2015).

To model this secondary texting task, we will be using the mobile phone touchscreen transcription typing model (Cao et al., 2018), which would be closest to mimicking a situation when someone is sending messages on their phone. As the authors state in their paper as a recommendation for future research; it would be interesting to examine whether the combination of the transcription typing model and a driving model would result in realistic driving behavior.

As a common driving performance measurement in the related work, we will focus on the lateral deviation of the car. Often the longitudinal deviation is included as well as a performance measurement, but since this task is rather complex, the goal is for the driving scenario is to be as simple as possible to exclude external factors such as distractions in the form of other cars.

2: Methods

The aim of this research is to test whether a computational cognitive driving model and a secondary computational model can be combined to simulate the resulting driving behavior. For the primary model, we are using the previously mentioned well-established cognitive driving model from Salvucci (Salvucci, 2006). For the secondary task, we will be using the mobile phone touchscreen transcription typing model (Cao et al., 2018). Both of these models are created and validated within the computational cognitive architecture (QN-)ACT-R. To achieve an accurate representation of the combination of the two models, we have to establish certain assumptions on how we would expect humans to multitask between these two tasks in a general sense, as these assumptions have to be built into the complete combined model itself in order to be able to interleave (Salvucci et al., 2005). The simulated driving behavior will then be compared against the driving behavior that resulted from the empirical study.

2.1 Modeling and Simulation

Within ACT-R production rules are the basic operations. For example, in a typing model, there are production rules that describe the finding and pressing of a certain key on a keyboard. The action that follows is described within the model itself, which depends on the successive states of the simulated mental model, which are stored in buffers. This is what makes multitasking somewhat difficult to accurately describe within a formal cognitive production rule system such as ACT-R, since the model is based on certain assumptions regarding human behavior. Due to this limitation, we have to hypothesize about where the task switching (i.e., interleaving) would most likely occur.

By consulting literature on earlier studies, it can be stated that it would be beneficial to establish different approaches to model the different strategies for the interleaving between two tasks (Brumby et al., 2007). This is necessary so that we can create simulated data for different strategies, to determine if we can capture the true human behavior with one of those strategies. Thus, eventually, these simulated results will be compared with the empirical data that we find in the experiment. Hereby we can test whether one of the interleaving strategies fits the data and reason whether ACT-R would be suitable for modeling more complex multitasking scenarios.

Four different interleaving strategies are implemented.

Strategy 0: We have one naïve approach, where after every find and press key production rule the attention is brought back to the road and to continue typing when the car is perceived to be in a stable state.

Strategy 1: A more reasonable interleaving strategy would be that instead of after every letter, the interleaving occurs after every word. This interleaving strategy is chosen as an adaption to the natural breakpoints that have been shown to arise when phone numbers are entered during driving (Janssen et al., 2012). Especially when a given sentence has to be typed during a driving task, we could hypothesize that the interleaving occurs in similar chunks as they do with phone numbers,

to have at least some situational awareness on the road periodically, which is also known as the task performance (or speed-accuracy) trade-off (Janssen, 2012).

Strategy 2: The third strategy is to interleave after every two words, which is based on the theory that the “word sentence span”, i.e., for transcription typing is usually between 2 and 8 words (Salthouse, 1986). Based on this earlier work, the transcription model takes two words as its base-level “word sentence span”, as they found that this best fitted the human data (Cao et al., 2018).

Strategy 3: Finally, we have the last strategy on task switching, where the full secondary task (e.g., typing a full sentence) is completed before the attention is brought back to the driving itself. In other words, no interleaving is taking place during the secondary task.

Besides determining the different strategies, we also have to provide the text that will be typed by the models. Since we want to have a somewhat realistic scenario considering mobile phone usage as a secondary task, we will be using short simple sentences. We will create three distinct categories of different numbers of words per sentence; four-, five- and six-word long sentences. For every category, three different sentences are predefined which adds up to nine different sentences in total (see *section 2.2.3* for the exact sentences).

2.1.1 Model implementation

Since both models are combined, extra production rules are necessary to describe the imminent change of attention/focus back and forth. Fortunately, research has been done on this topic with empirically validated models that have been tested for smaller distractions during driving (Cao & Liu, 2013; Salvucci et al., 2005). We will go through both models and discuss how they are implemented and ultimately interwoven with each other. All models are implemented in the QN-ACT-R framework (*QN-ACTR-Release, 2017/2019*).

We will be using an adaptation of the driving model by Salvucci (Salvucci, 2006) with two-point visual-motor control implemented in (QN-)ACT-R. This driving model uses the same values as the original model as these have proven to be the best-known fit for human driving behavior. The model, however, is missing the stability assessment procedures of the original model so these had to be added in. The original model also had a scaling parameter for the stability assessment of the car, which is used for adjustment to the environment in which the simulation is run. Since these values have not been tested in *TORCS* before, we will have to estimate the best scaling parameters for stability. At the end of this section, we will examine these values further.

Parameter	Value	Method
k_{far}	16.0	Estimated
k_{near}	4.0	Estimated
k_i	3.0	Estimated
θ_{nmax}	0.07 rad	Estimated
k_{car}	3.0	Informal
k_{follow}	1.0	Informal
thw_{follow}	1.0 s	Informal
thw_{pass}	2.0 s	Informal
$p_{monitor}$.20	Informal
d_{safe}	40 m	Informal
θ_{stable}	0.07 rad ($\approx 1/4$ lane)	Informal
$\dot{\theta}_{stable}$	0.035 rad/s ($\approx 1/8$ lane/s)	Informal

Figure 2.1 Parameter values used in the driving model (D. D. Salvucci, 2006)

Besides the driving model, we will be using a transcription typing model that has been implemented within the QN-ACTR framework as well (Cao et al., 2018). For every run of the simulation, the sentences are set within the model accompanied with its corresponding interleaving strategy. The model has been adjusted to be able to use different interleaving strategies. The two-word sentence span is only altered for the one-word and single-letter interleaving strategies. In those cases, it means that the imaginal chunk for the temporary storage of the words is re-initiated after every interleaving occurrence. With two-word interleaving and full-sentence strategies, this happens every two words as originally modeled.

To interleave between the two models, extra productions rules have to be added to handle the on and offloading of chunks within the buffers. In *figure 2.2* below, the flowchart is illustrated showing the transitions between the two models.

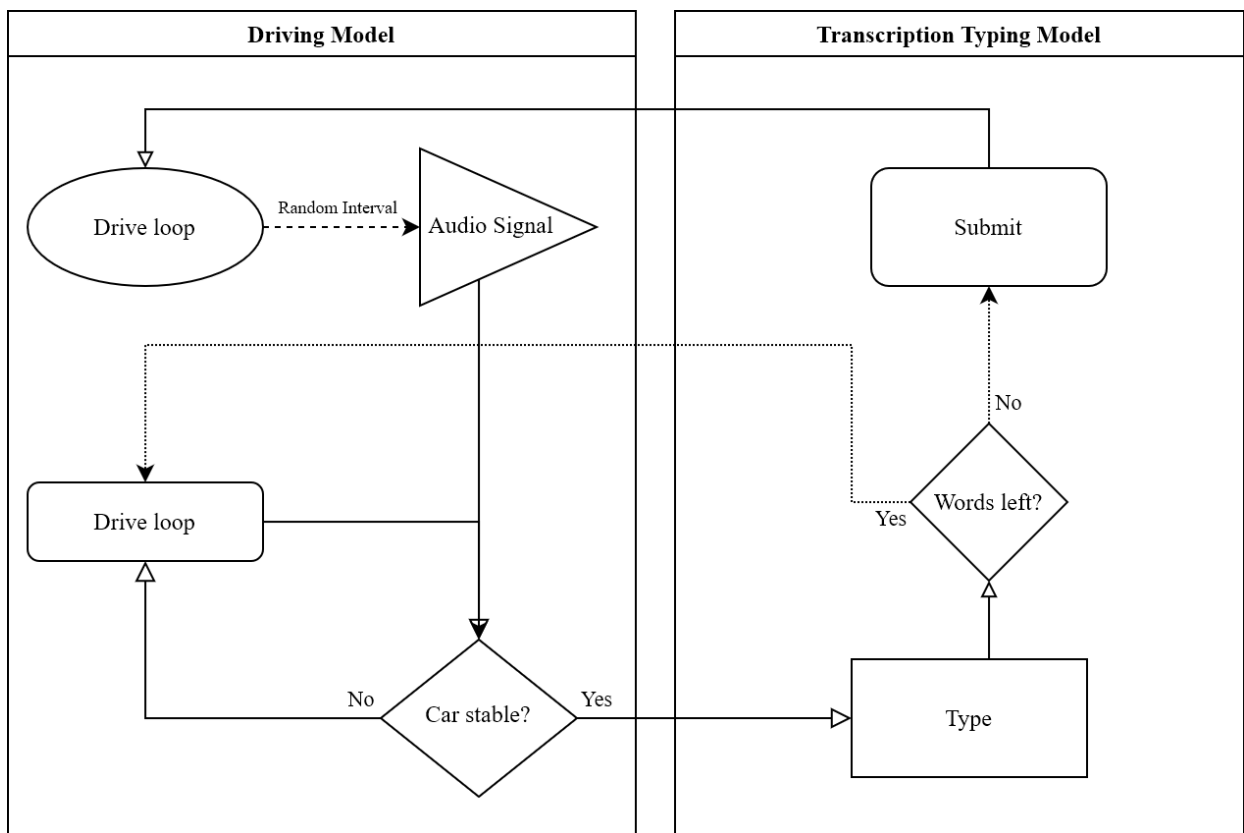


Figure 2.2 Flow chart between the two models

The initiation of attending to and preparing for the secondary task can be divided into three steps; *orienting* to the sound of the phone, *locating* the phone and *handling* the phone (Haddington & Rauniomaa, 2011). The first step is modeled as the processing of the audio signal by using the “Aural Module” that is available in ACT-R. Locating and handling of the phone is simulated only when the car is assessed as being stable after the processing of the audio signal. As ACT-R is limited in more complex operations within its “Motor Module”, this delay has been set as a fixed variable at 1370 ms, which is based on variables used in earlier studies (Salvucci et al., 2005). The total time between the audio cue and the first letter typed ideally should be around 4 seconds (Fitch et al., 2013), as these operations closely resemble the initiation steps for hand-held calling.

The task switching itself is implemented similarly to the threaded cognition theory that we discussed in the introduction. (Salvucci & Taatgen, 2008). In practice, this means that multiple-goal chunks can coexist in the goal buffer. When the interleaving occurs between typing and driving, all ACT-R modules are re-initiated that were cleared for the typing task, except for the second goal buffer chunk which remains on the typing task if there are still remaining characters/words to be typed. This results in a delay before the attention is shifted back to driving. When all buffers and chunks are back in the original car driving task state, the stability assessment production rule is performed. When the car is deemed to be in a stable enough state, the switching to the typing task occurs again when the second goal buffer chunk still holds the typing task state. Except for the goal buffer, the other buffers are constrained to their basic functionality. In practice, this means that, as is defined within ACT-R; the visual buffer, the manual buffer, the retrieval buffer, the imaginal buffer, and the aural buffer can practically only hold one chunk at a time.

As mentioned before, this stability assessment parameter has been set as our only free parameter for our model as this was originally implemented with a scaling parameter for a different environment. Due to this, we will establish three different scaling values. The stability function itself takes into account the stability of the near and far points and the lane position (Salvucci & Macuga, 2002). To achieve this, we test the model with the simplest typing task sentences of four words and gradually scale up the stability parameter. In *figure 2.3* below we can see that the lateral deviation while performing the secondary task goes up in a close to parabolic trend.

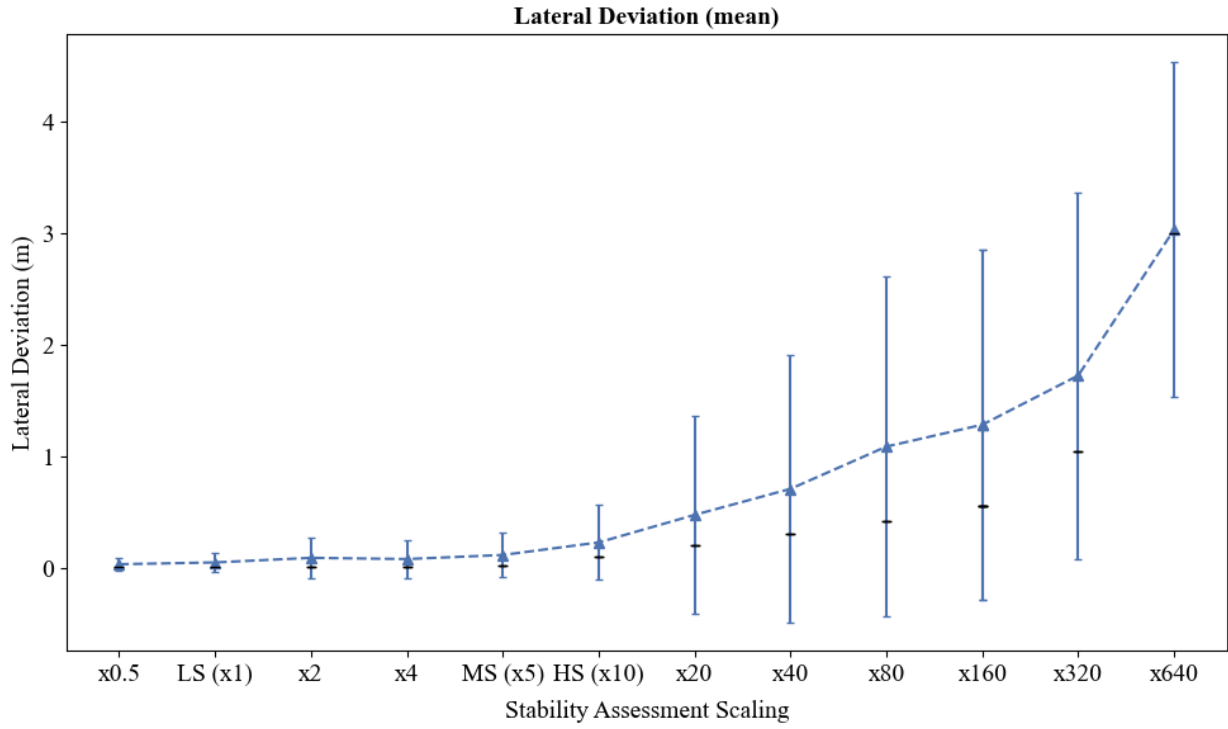


Figure 2.3 Mean lateral deviation for different stability scaling values

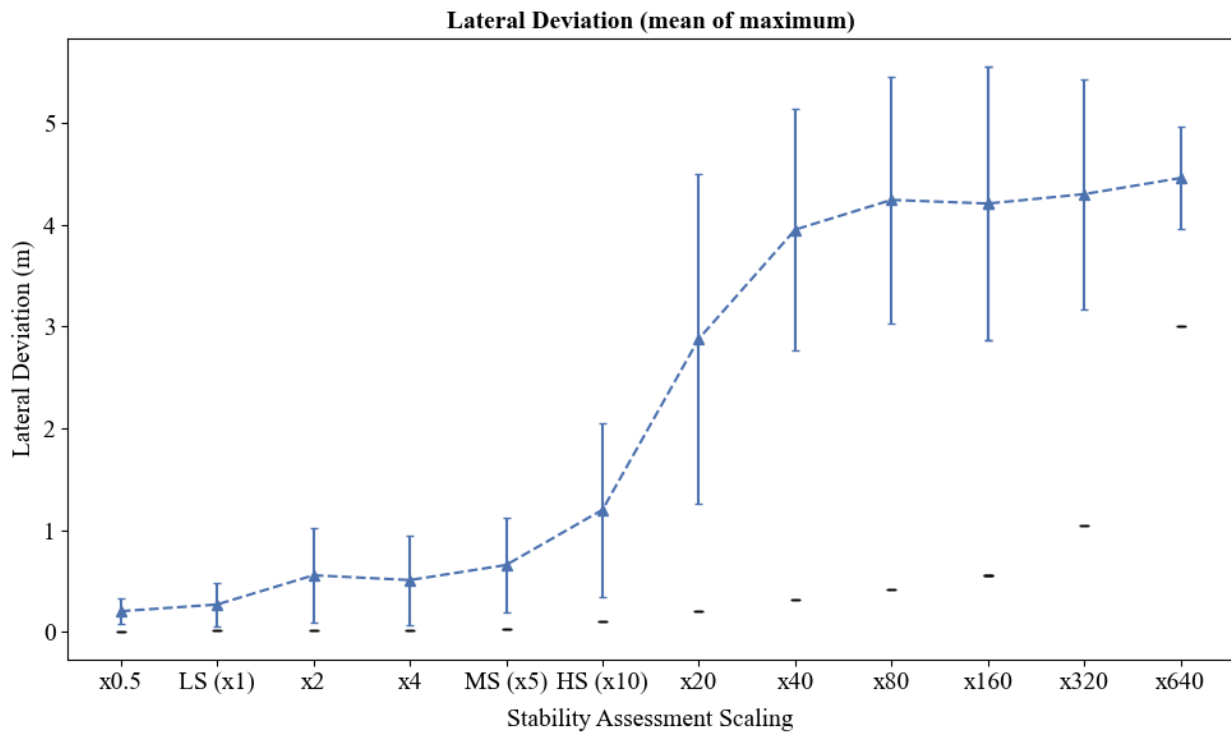


Figure 2.4 Mean of the maximum lateral deviation for different stability scaling values

Even though this is an expected result, this does not tell us enough if we want to find a good value for the scaling. Therefore, we can also examine the mean of the maximum lateral deviation per task. In *figure 2.4*, we see that we soon hit a ceiling, indicating that the driving performance is becoming too atrocious at a certain level. As a matter of fact, the $x640$ scaling is not able to finish the track at all. In the task times below, we get an unsurprising result of lower task times accounting for higher scaling values.

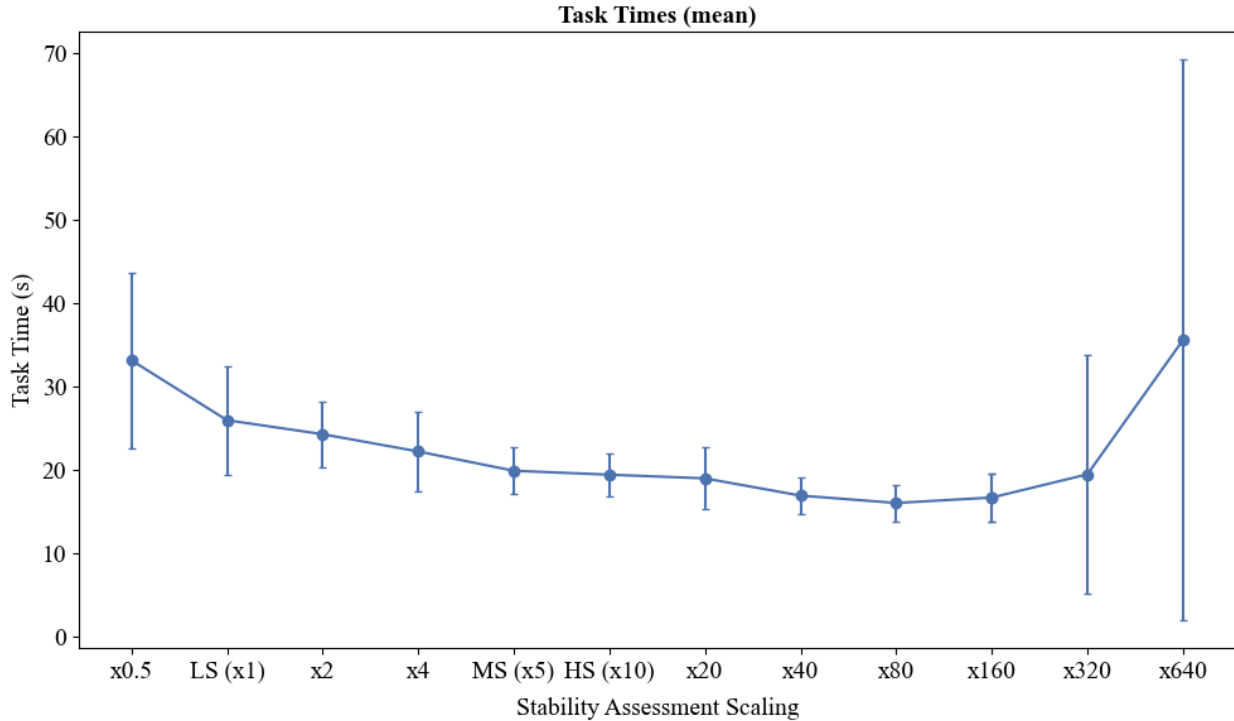


Figure 2.5 Mean task times for different stability scaling values

For a good balance between task time differences and lateral deviation differences, the three scaling parameters that have been selected are: $1x$, $5x$ and $10x$. The decision was made to not go higher than a $10x$ scaling parameter, as that would result in unrealistic bad driving behavior. So, every interleaving strategy model will be run three times with these three individual scaling parameter values.

Concluding, for the final simulations in total this sums up to 4 different interleaving strategy models, times 3 for the adjusted stability assessment value, which means that in total we have 12 different models. Each individual model will perform $N=180$ tasks for every sentence group, which means that for every individual sentence there are $N=60$ performed tasks. This sums up to each different model, using a different interleaving strategy, having conducted $N=540$ tasks in total.

2.1.2 Environment implementation

To run the model a task environment is needed. In this case, we use an external environment that ACT-R can control. For the environment and vehicle physics, a modified version of the open-source driving simulator *TORCS* (Wymann et al., 2015) is used. The driving model is able to interact with the driving simulator using the UDP network protocol. In practice, this means that all relevant data is being sent from *TORCS*, whereafter it is controlled by the model through raw input based on the last received values.

The driving scenario itself consists of a highway of 17km with two lanes going in both directions (*figure 2.6*). The two lanes themselves are 3.75m wide and the emergency lane on the right measures 3.5m in width. There are 17 straight sections connected by an equal number of left and right curved sections. No other cars or distractions are present on the track itself, to ensure that the only distraction will come from the typing task in both the model and the experiment.

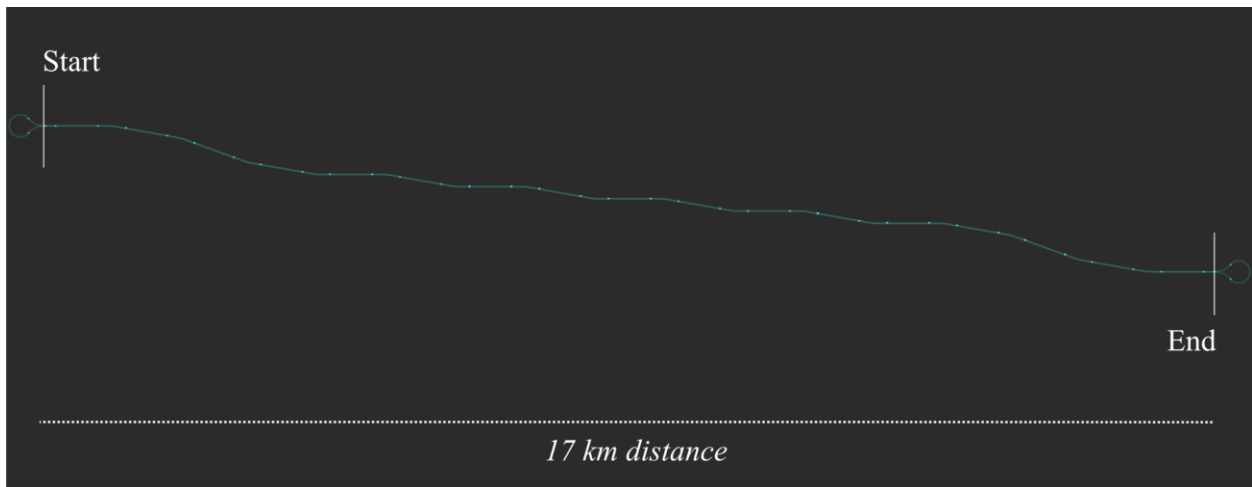


Figure 2.6 The track for both the model and the experiment

The speed of the car itself has been limited to 20 m/s, based on previous research employing a similar scenario (Rehman et al., 2019). Since we are interested in the lateral deviation of the car, the speed remains constant over the duration of the track. After every run of the simulation (i.e., having driven 17km), the scenario is reset with either a different sentence/interleaving strategy or left unchanged depending on the number of desired secondary task trials. All variables of interest are logged over the duration of the course, such as; time, lane position, distance, distraction and task times. An excerpt of variables of interest that were logged during one of the simulations can be seen below in *figure 2.7*.

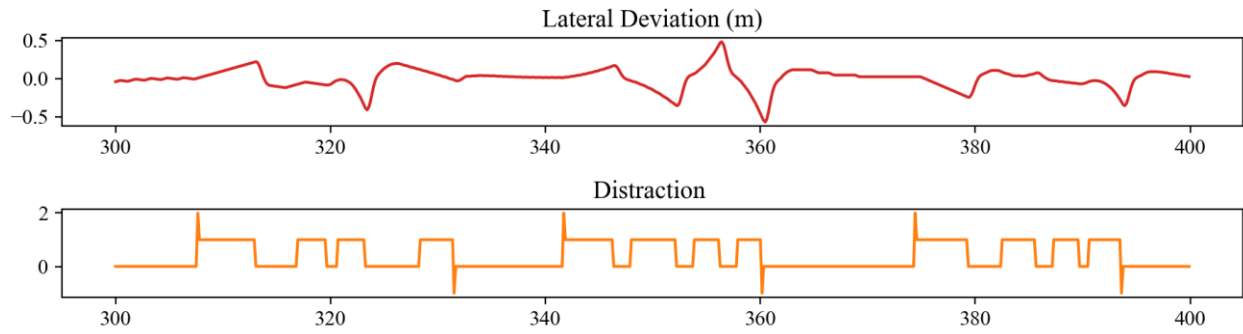


Figure 2.7 Lateral deviation (lane deviation) and distraction plotted over time in seconds. The spikes in distraction indicate the start or end of an individual typing task.

2.2 Empirical Study

Since both individual cognitive models have been validated in the past (Cao et al., 2018; Salvucci, 2006), the experiment will not be focused on validating both models individually. Instead, the experiment has two different goals. Since we have acquired the simulated driving performance from four different interleaving strategies within the model, the main objective is to acquire the same data on the human participants to be able to compare these. The data concerns two measures of driving and task performance, the mean lateral deviation of the car as a function of time during the tasks and the time it took to complete each individual task. The secondary objective is to record all the user inputs on the smartphone as well as the controls for driving the car, to possibly derive interleaving data from the inputs themselves, as we are not using an eye tracker for this particular experimental setup. The study was approved by the ethical committee at the *University of Utrecht*.

It is of major importance that the data gathered during the empirical research resembles the experiment conditions of the simulation phase as closely as possible. By doing so we can minimize the bias in the data and make a better comparison between the two data sets. In this section, we will examine the experiment design and setup.

2.2.1 Participants

The requirements for the participants have been adopted from a comparable study on driving behavior during distractions in a simulator (He et al., 2014). In order to participate, the participants were screened to be in possession of a driver's license and have at least 2 years of driving experience. Furthermore, they had to have at least good experience with touch screen typing, preferably in terms of daily usage. Considering the demography, the aim was to assemble a predominately homogeneous group, which should result in the most uniform results. The final group consisted of twenty (N=20) subjects within the age range of 22 to 33 years old that participated in this study. The distribution of male and female participants was exactly 50/50 and all of them were right-handed. All participants were fluent in the English language.

2.2.2 Materials

The experiment itself was conducted in one of the labs located in the *University of Utrecht* that are dedicated to simulated driving studies. The room itself was void of any other distractions and the participants were left alone after the instructions, to minimize any external distractions or influences during the experiment. The participants had a place next to them where they could place their phone when needed.

For the primary task (the driving task) of the experiment, a portable driving simulator was used which consisted of a *Logitech G27* steering wheel and pedals. The wheel was set to its' 900 degrees input mode, to mimic the steering of a real car as closely as possible. The included H-shifter was not used as the car was set to automatic transmission. Besides the driving control setup, a full HD (1080p) 27" monitor was connected to a system capable of running the aforementioned driving simulator environment *TORCS*. In the driving simulator, the same highway scenario and layout as in the simulation phase were presented to the user, where the participant was able to manually control their virtual car.

For the secondary task, a *Samsung Galaxy S6* smartphone running *Android* was used which ran a custom app that prompted the user to write a particular sentence at random intervals. Similar to the implementation within the models, the app provides an auditory signal to attend the user to perform the typing task. It is designed in such a way that this will feel familiar and comparable to receiving a text message through an instant messaging app. During the experiment, all the user input on the touchscreen and all the steering wheel input are recorded and synchronized for later analysis.



Figure 2.8 The final setup of the experiment

The app itself is created with *Ionic* using the *Angular* framework, which is a powerful cross-platform mobile app development toolkit. The app communicates with a custom *NodeJS* server that is able to directly read the *TORCS* clock timer from the process memory by using memory injection. By doing so, every keystroke within the app can be precisely synchronized and logged with all the other user input (*Figure 2.9*).

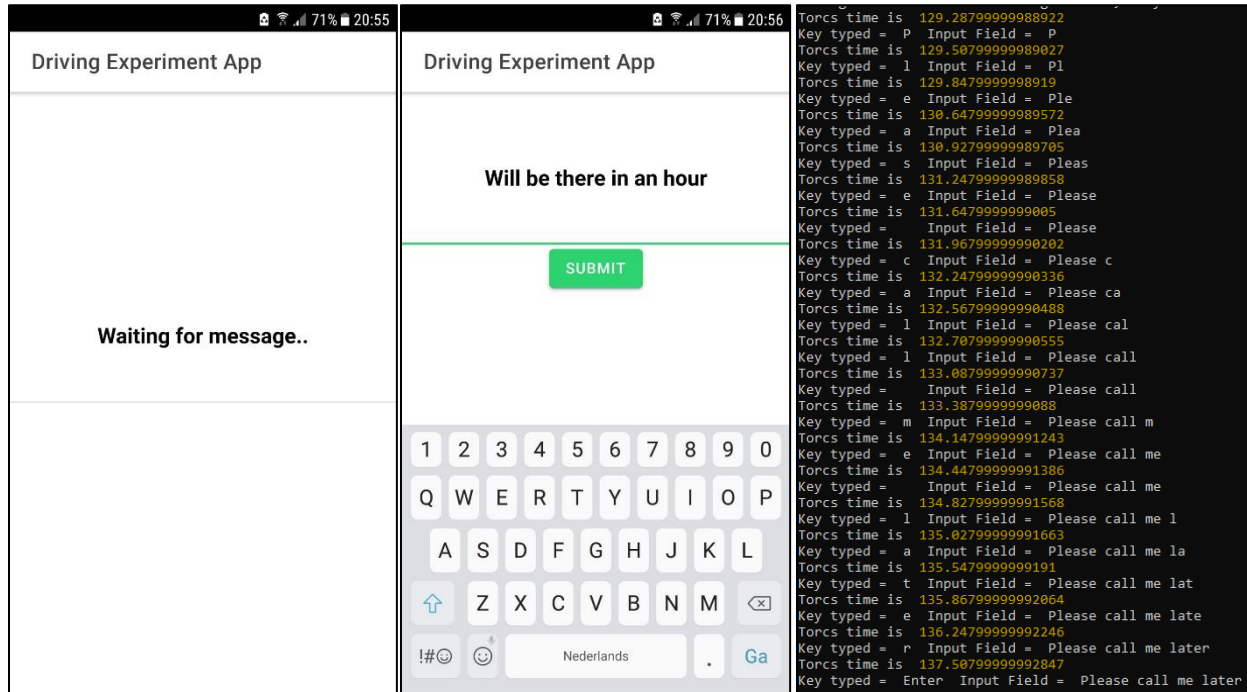


Figure 2.9 Left and center illustrations: simple app interface. On the right: receiving the input from the app

The input from the steering wheel is logged as well in the *NodeJS* server. An example of a full session from one of the participants can be seen in *Figure 2.10*. In the top two graphs, we can see the lateral deviation in meters from the center of the lane and the corresponding absolute steering wheel angle. The graph ‘*KeyPressed*’ shows all the individual recorded keystrokes and the graph ‘*Space Typed*’ shows when a ‘space’ character was entered on the touch keyboard. Finally, the ‘*Words*’ graph shows the length of the particular sentence that had to be entered.

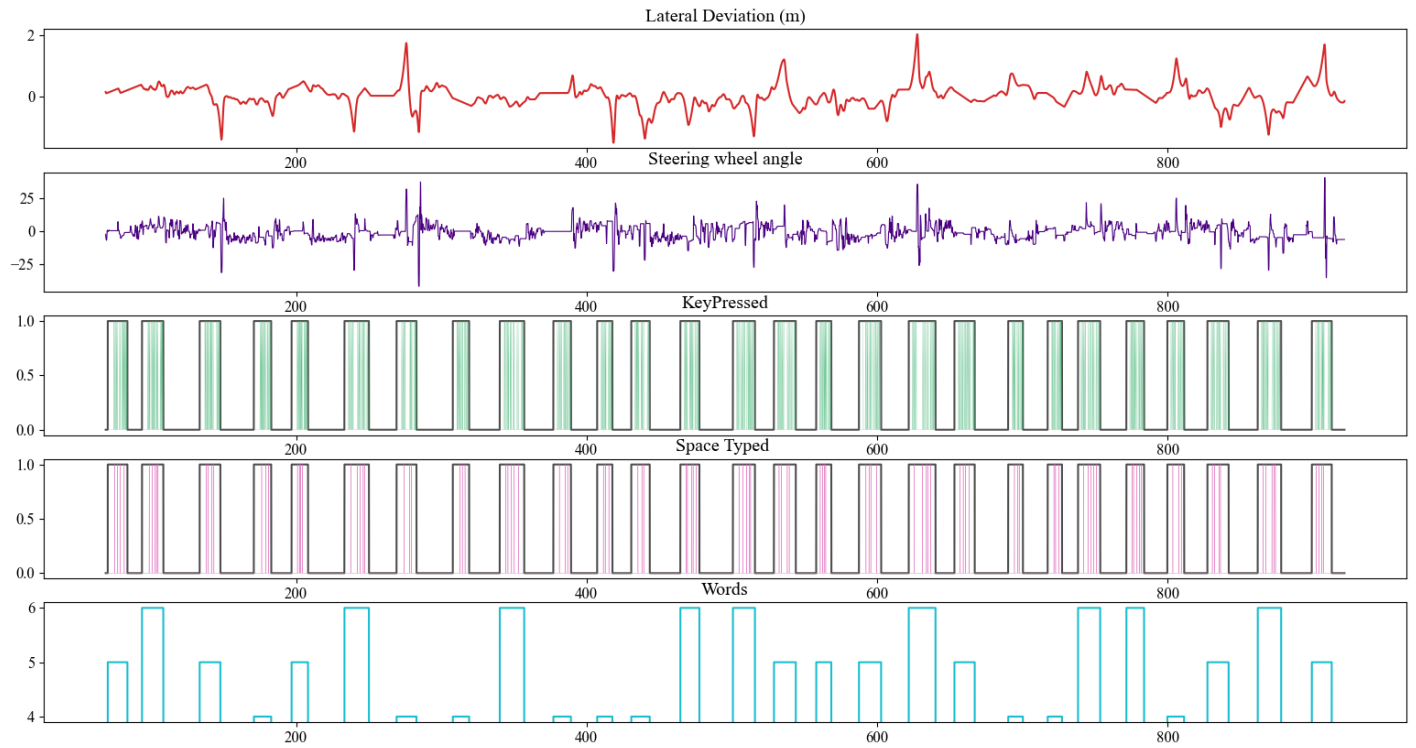


Figure 2.10 The resulting data from one of the participants

2.2.3 Design

Similar to the simulation phase, the main experiment can be divided into two different states where the participant might find themselves and where the relevant data is measured over:

State 1: The participant is driving the car in the rightmost lane to the best of their abilities. Baseline driving performance is assessed by monitoring the mean lateral deviation from the center of the rightmost lane.

State 2: The participant is performing the typing task while concurrently driving the car in the rightmost lane to the best of their abilities. The driving performance **while** having to perform the typing task is assessed by monitoring the user’s input on the touchscreen as well as monitoring the mean lateral deviation from the center of the rightmost lane.

As can be seen in *Figure 2.10*, over the course of the experiment the participant will be in either one of the two states or transitioning between the two. The main state is the first state where the participant will be focused solely on driving. At random intervals, sampled from a uniform distribution between 10 and 25 seconds, the participant will be prompted to perform the secondary task which will transition them to the secondary state. The switching itself, between state 1 and state 2 (after state 2 was initially activated) and the resulting driving performance is measured in mean lateral deviation. Besides the driving performance, we are also measuring the time it took for the complete task to complete, which amounts to the time between “Phone Alert” and “Submit”.

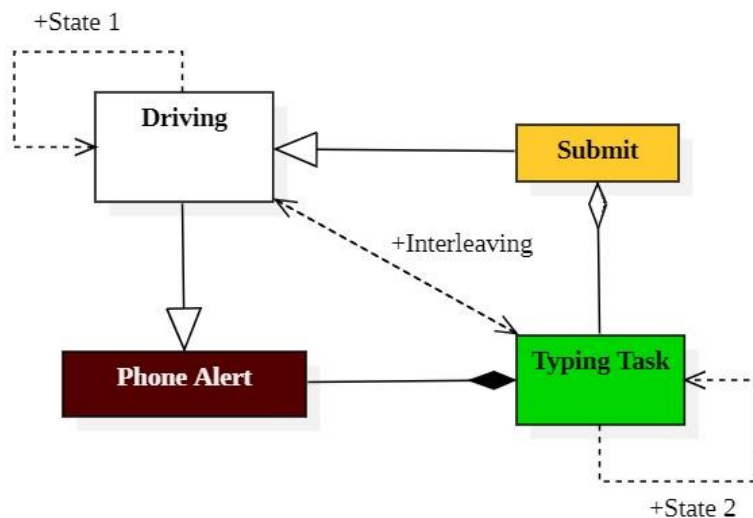


Figure 2.11 The experiment states and transitions

The experiment uses the same nine sentences as the models and is divided into three different categories depending on the number of words per sentence. The words have been chosen in such a way to make sure that the total characters per sentence category are not fluctuating too much. Individual words have been chosen to have a high likelihood of familiarity in order to reduce potential confusion, as this is something that the transcription typing model does not take into account as well.

4-word sentences	5-word sentences	6-word sentences
<i>“Will contact you soon”</i>	<i>“See you later no problem”</i>	<i>“Hey call later I am driving”</i>
<i>“Please call me later”</i>	<i>“The road is pretty boring”</i>	<i>“Will be there in an hour”</i>
<i>“What is the address”</i>	<i>“I will do the groceries”</i>	<i>“What time will you be home”</i>

Figure 2.12 Sentences categorized in their respective categories

To make sure that all categories are performed multiple times, there are three different runs in which all nine sentences are shown in random order. This results in 27 typing tasks being performed per participant over the course of the experiment. By doing so we can aggregate all the data from each category.

2.2.4 Procedure

After the participant has read the general information of the purpose and the goal of the experiment and has given their consent, the instructions of the experiment were presented to the participant. During the instructions, it was emphasized that the participant should try to feel as immersed as possible and that the experiment is not about achieving the best driving or typing performance. They were asked to engage as much in the setting as possible and to do what felt most natural to them; as if they were driving a real car and had to respond to an important message, even though that would obviously be questionable behavior in a non-simulated environment due to this being a violation of traffic rules.

The participant was also informed that the main objective is to keep the vehicle within its designated lane and that the throttle could be kept at a maximum at all times until the completion of the experiment. To prevent participants from keeping the phone in their hands at all times, they were instructed to place the phone back next to them or on their lap when they were not using it, depending on what they felt was more comfortable. Also, they were instructed that the typing task had to be conducted with one hand rather than with two. Furthermore, they were also instructed to not correct typing mistakes, as this is also something the model does not account for.

The instructions were first followed by a practice trial to get familiar with the typing task on the smartphone. After that, the primary task started which is the car driving simulator. After the participant had familiarized themselves with the car driving for 2 minutes, the main experiment began. This meant that from that point on, the auditory cues started to alert the participant to perform the secondary task. As soon as it was clear that the participant had understood the task and performed the first transcription typing successfully, the participant was left alone for the further duration of the trial.

After 27 completions of the secondary task, the full trial is finished. After a short debriefing, the experiment itself was concluded. If the participant was particularly interested in their driving and typing performance, the results were optionally shown to the participant. The complete duration of the experiment took around 25 to 30 minutes per participant.

3: Results

In this section, we will examine the results that were collected from the models and the empirical data from the study. First, we will present the data generated by the models and in the subsequent section, we will continue with the empirical data yielded from the participants that participated in the experiment. Finally, we will end with comparing both acquired datasets. All the data processing and graphs have been conducted using *Python 3.7* (Rossum & Drake, 2009) and *Pandas* (Reback et al., 2021). The statistical analysis was performed in *R* (R Core Team, 2020) using *RStudio* (RStudio Team, 2020).

3.1 Modeling and Simulation

As described in the methods section, four different interleaving strategies have been defined. These consist of every letter interleaving, every word interleaving, every two-word interleaving and no interleaving; which have been all simulated over the groups of 4, 5 and 6 letter sentences. Since all parameters of the driving model itself have been empirically validated, these are not changed over the course of the simulations. As a reminder, the stability scaling factor can be interpreted as low being the most conservative in the stability assessment, while the higher values result in a more liberal assessment. We are interested in two distinct measurements: the lateral deviation during the tasks and the time it took for individual secondary tasks to be completed.

3.1.1 Task time analysis

The first step was to analyze the resulting task times from all the individual models categorized in their separate sentence groups, which is shown in *Figure 3.1*. All the times are shown as mean times with the error bars respectively showing the standard deviations. The individual task times were measured from the onset of the audio stimulus up until the final text is submitted.

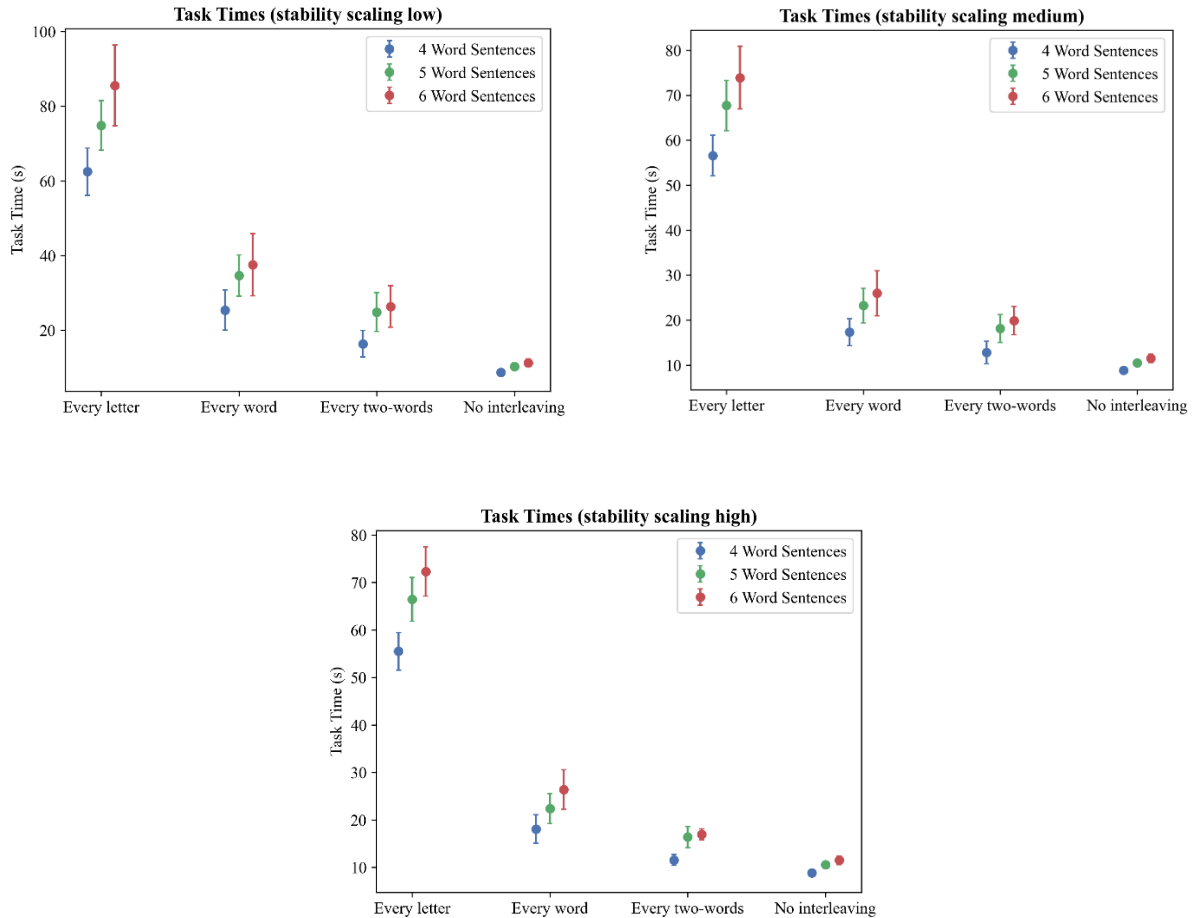


Figure 3.1 The mean task times per strategy, sentence group and stability factor

As is to be expected, we can see that for every individual strategy the sentences with the most words take the most time. Note that the difference when using the “Every two-words” interleaving strategy is less between the 5- and 6-word sentences compared to the 4-word sentences. This could be logically explained as the 5- and 6-word sentences are performing one extra interleave compared to the 4-word sentences. The difference in task times between these two is also lower as it apparently does not take that much more time to type an extra word uninterrupted successively.

It is also evident that when the “No interleaving” strategy is utilized, the standard deviation is very small. This is caused by the secondary task being deterministic, which results in very similar task times for the same sentences, as no interleaving occurs. Considering that when the audio stimulus is given, the car often finds itself in an already stable position, the secondary task can immediately start which results in very similar task times.

By aggregating the previous data over their respective sentence categories, we are better able to compare the effect of the stability assessments on the task times. In *Figure 3.2*, this aggregated data is shown. We can see that the error bars are becoming less extended for higher stability scaling values. This is a direct result of the time it takes between every interleaving action to assess if the car is sufficiently stable to move back to the secondary task. That also explains why the “No interleaving” strategy does not show these characteristics.

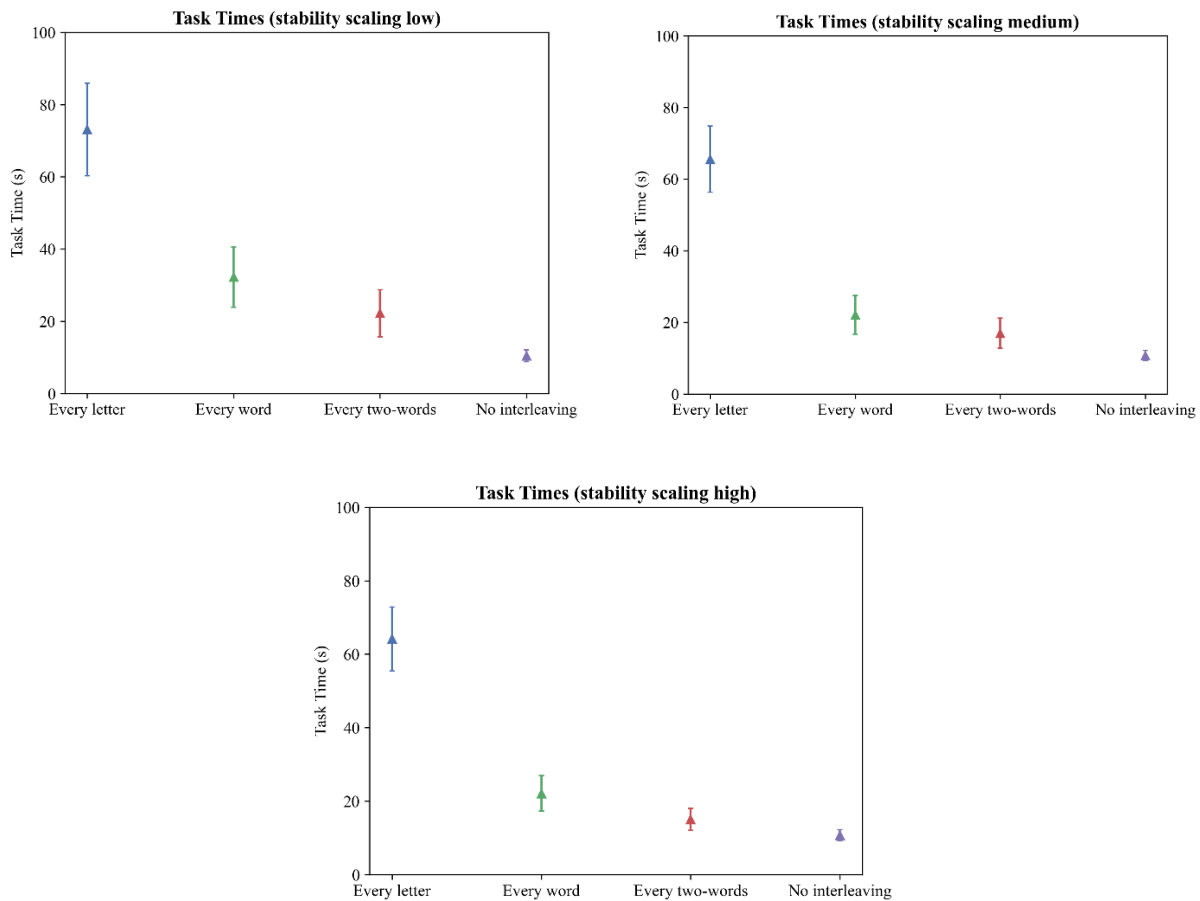


Figure 3.2 The mean task times per strategy and stability factor

The differences in task times when looking at the medium and high stability values are less pronounced compared to the task times in the low scaling graphs. This confirms our findings during the preliminary tests in the methods section where we observed the same relation between the scaling factor and average task times. However, in the next section, we will see that there is a clearer difference in driving performance when we analyze the differences in lateral deviation between those models.

3.1.2 Lateral deviation analysis

For the lateral deviation, we will go through the same procedure as for the task times. We will first look at the models with their different strategies over the different sentence groups. The lateral deviation is measured during the moment of distraction, i.e., during the start of the task until the end. In *Figure 3.3*, this represents the part of the graph where *Typing Task* has the value of 1. Another method valid method would be to take the lateral deviation values from where *Distracted* is 1, although we are not doing that here as we eventually want to compare this data against the empirical data. If we could accurately derive from the empirical data exactly when the participants are interleaving, this method could be more accurate.

Besides the lateral deviation when the secondary task is active, we also want to acquire a baseline driving performance. This is simply done by taking the mean of all the data points when *Typing Task* has the value of 0, which results in an overall average of driving performance. In all lateral deviation measurements, we take the absolute value, as we are only interested in the absolute deviation from the center of the lane.

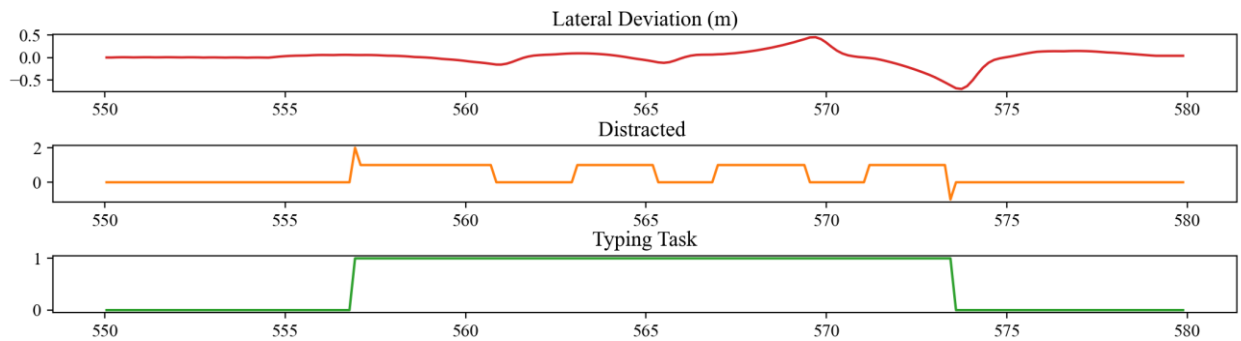


Figure 3.3 Lateral deviation is measured when *Typing Task* has the value of 1

Sampling all the lateral deviations data points when performing a complete typing task and taking the mean of these we get the results as shown in *Figure 3.4*. The black line indicates the baseline lateral deviation, which is the baseline performance when the model is not in the state of being distracted (i.e., the driving model base performance).

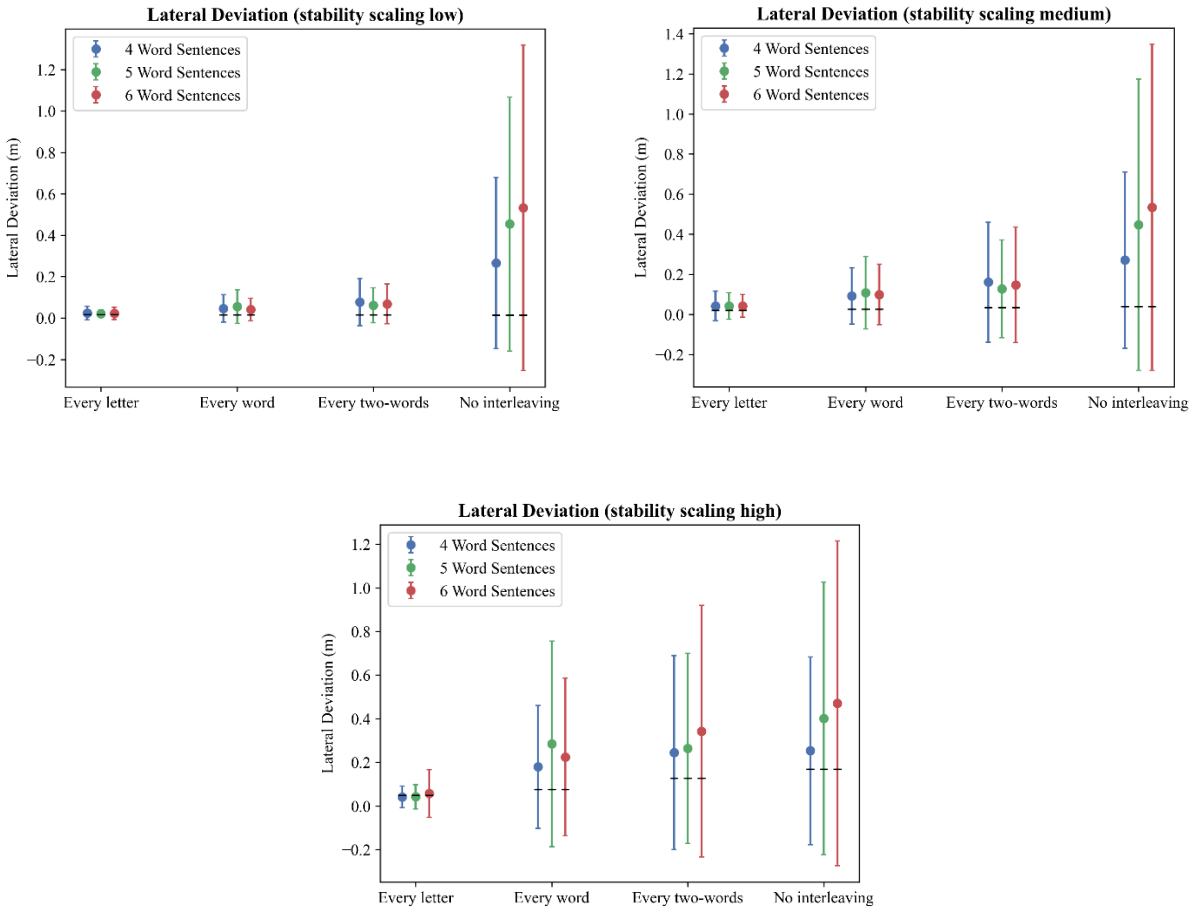


Figure 3.4 The mean lateral deviations per strategy, sentence group and stability factor

When examining these results, we can see that the baseline driving performance is quite accurate overall. This is the driving performance that the standard parameters from the driving model yield. These values being rather low is probably due to the highway scenario itself, with mostly straight sections and rather gentle curves, plus factoring in a limited speed of the car itself. When the secondary task is performed, we see varying results. The differences between the individual interleaving strategies and the length of the sentence are less pronounced compared to the results that we have seen before concerning the task times. This could partially be explained by the rather large standard deviations, which could be affecting the mean values. Consequentially, this could mean that the sample size might be too small. Another possible reason is that the moments of interleaving (i.e., the time it takes when the car is restabilized during the interleaving) are contributing to a proportionally large part of the measured lateral deviation samples. However, as

we are using the identical method for measuring the data in the empirical data this should not pose a problem for the final comparison.

As the individual differences in the length of the sentences do not provide meaningful additional insights, we will again look at the aggregated data when we combine all sentence groups, which is illustrated in *Figure 3.5*.

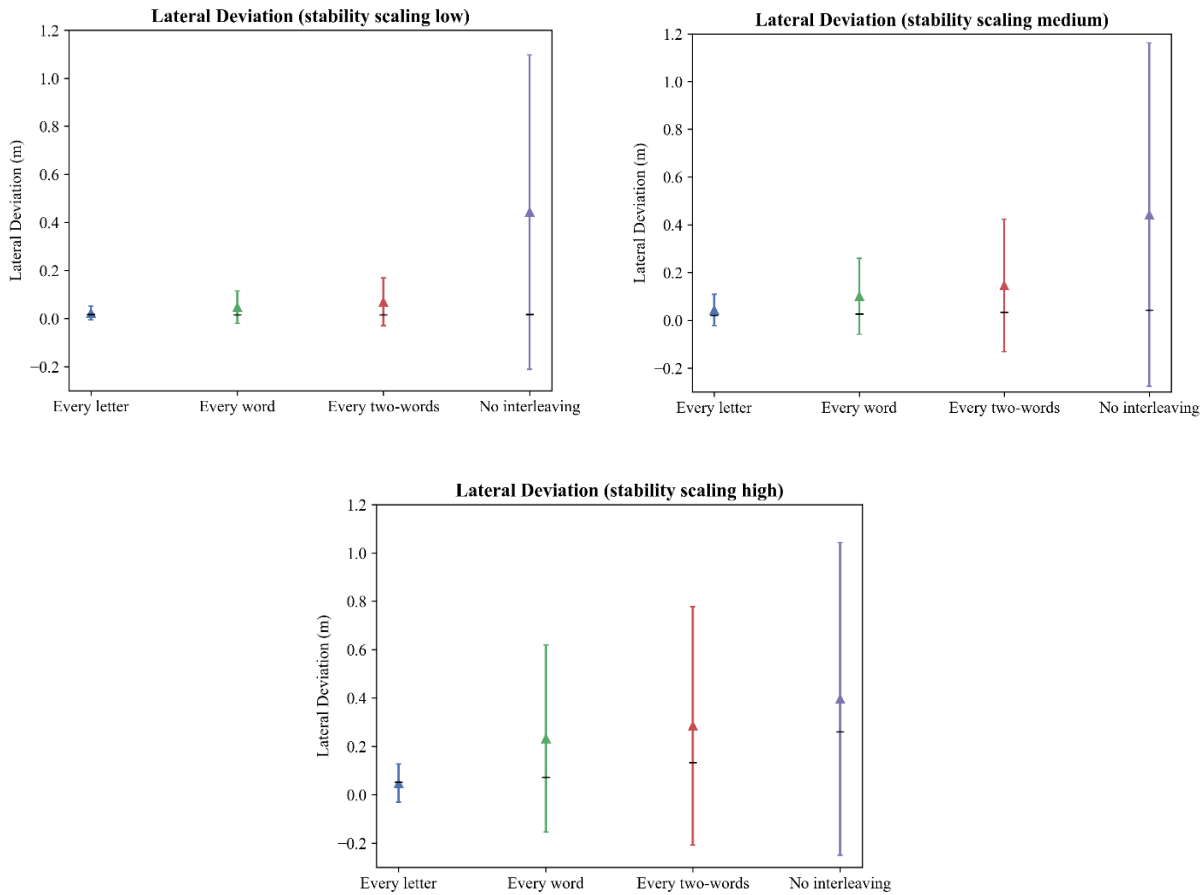


Figure 3.5 The mean lateral deviation per strategy and stability factor

When examining the aggregated data, the results seem to be more trivial. We can clearly identify that a higher stability scaling factor yields higher overall lane deviation for the every- and two-word interleaving strategies. This is to be expected, as the ‘swerving’ effect will be exaggerated during the moments of interleaving, since the secondary task will be continued when the car is in a less than ideal position as the stability assessment threshold becomes higher. We distinguish the same effect on a very small level for the “Every letter” interleaving strategy, but the effect is not present when “No interleaving” is utilized. Also, the baseline lateral deviation will increase due to the less strict interleaving policy when the stability scaling is high; further proving that the overall baseline driving performance is degrading when using higher values.

As mentioned earlier, when comparing the task times, we saw that the high and medium stability scaling models yielded comparable results. However, in this instance, the low and medium models respectively are providing the most similar results.

3.2 Empirical Study

In this section, we will analyze the empirical data that has been collected during the experiment. In total, $N=540$ tasks have been performed by the participants, adding up to over more than 350km driven in the simulator. However, by filtering out unfinished or failed tasks $N=487$ remains, which means that approximately there are 3 failed tasks per participant. In practice, this often meant that the ‘submit’ button was prematurely pressed when picking up the phone.

3.2.1 Task time analysis

First, we will look at the distribution of all the task times for all the participants to identify if there are outliers. The sample mean $x=18.09$ with a standard deviation of $s=6.05$. As can be seen in *Figure 3.6* and *3.7*, the data is mostly normally distributed, although being a little right-skewed. However, there is one outlier with a mean of 35.75. Considering that this outlier is still less than 3 standard deviations away from the mean (+2.86), we will not omit this data from the dataset.

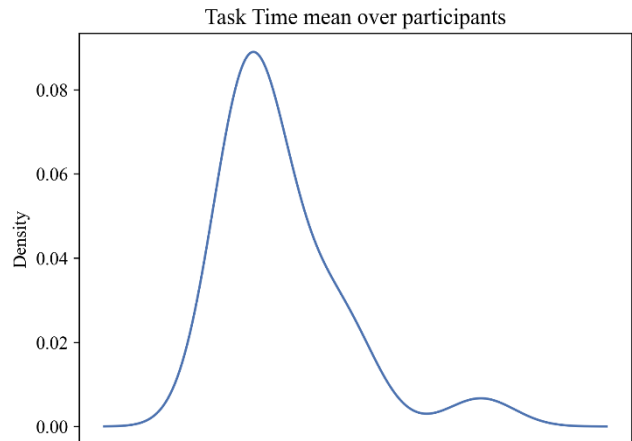


Figure 3.6 The distribution of the mean task times per participant

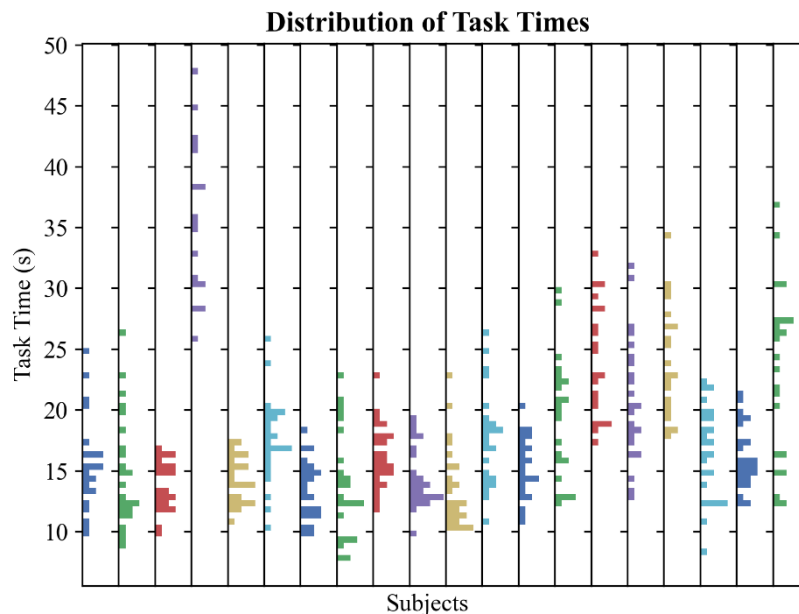


Figure 3.7 Task time frequencies over individual participants

We are also interested in finding out what the difference is in the task times over the different sentence groups. As expected, the tasks containing sentences with fewer words were performed faster which can be seen in *Figure 3.8*. The distribution of these task times per sentence group is shown in *Figure 3.9*.

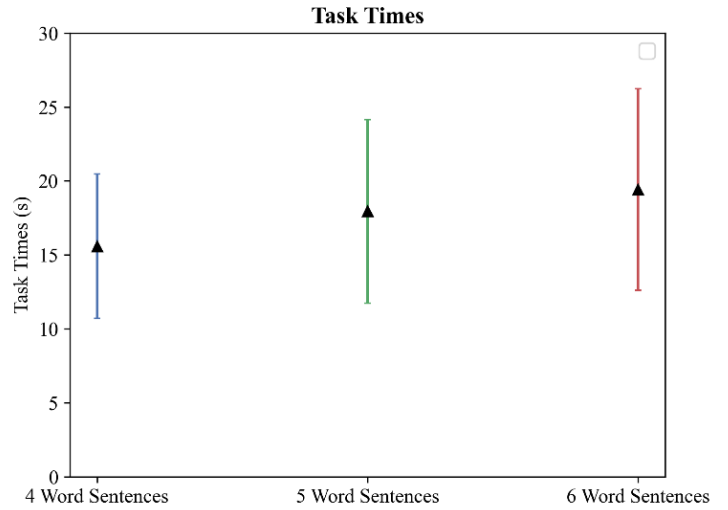


Figure 3.8 Mean task time per sentence group over all participants

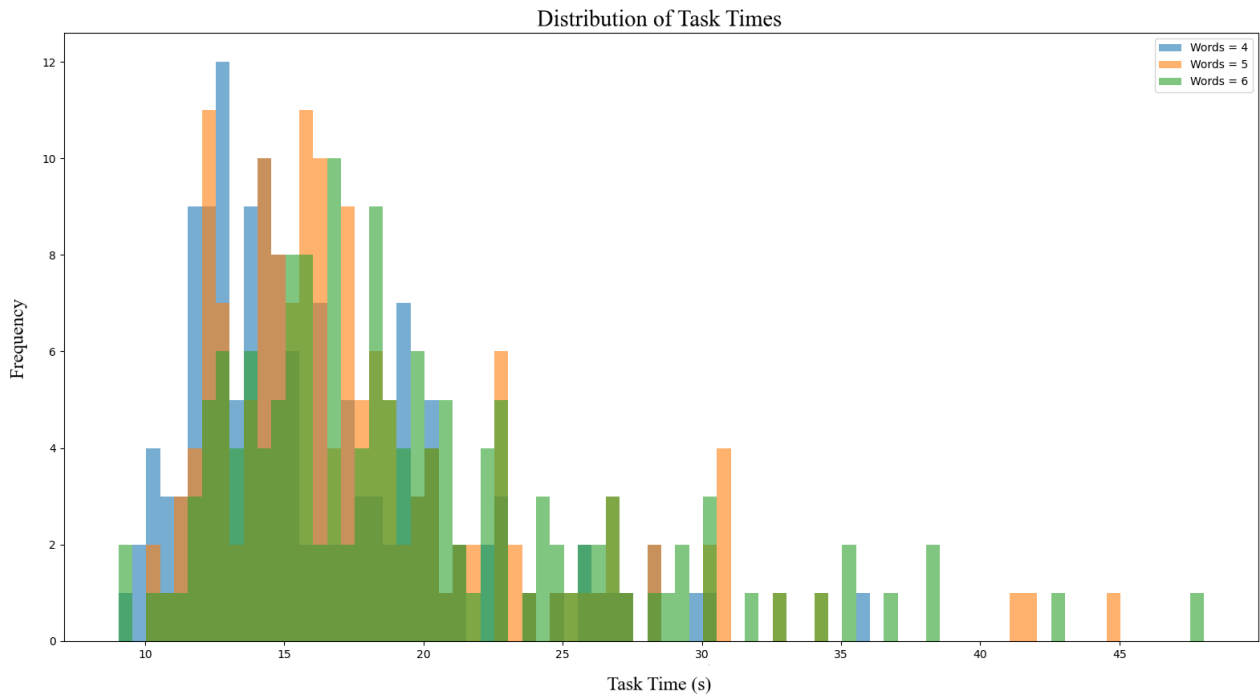


Figure 3.9 Histogram of the task times per word category

Multilevel modeling analysis

To determine the covariance between the task times and the individual participants, we can use a linear mixed model. If we apply this statistical analysis on all the task times grouped by the individual subjects, we get an intraclass correlation coefficient (ICC) of 0.631 , thus 63.1% of the total variance is accounted for by individual differences. Since this value is higher than 0.5, we can state that the difference in observed task times between the individuals explains most of the variance (i.e., there are a lot of individual differences in task time) (Park & Lake, 2005). When we add the number of words as a predictor variable to this model, we find that the ICC becomes 0.671 , which with a 4% increase is only marginally higher.

We can also look at the 3-level model; where the number of words, instead of being a predictor, is interpreted as an additional grouping variable within the individual participants. When we add the number of words as a grouping variable to our previously established null model, we can see in *figure 3.10* that the number of words attributes around 10% of the overall variance. If we test this with an ANOVA test between the three models, we get the following results:

Model	Parameters	Df	Pr(>Chisq)
1 Level model	2		
2 Level model	3	1	$< 2.2e^{-16}$
3 Level model	4	1	$9.662e^{-15}$

By using an alpha level of $\alpha = .05$, we can conclude that both the individual subjects as well as the number of words within the subjects are contributing significantly to the overall variance in task times.

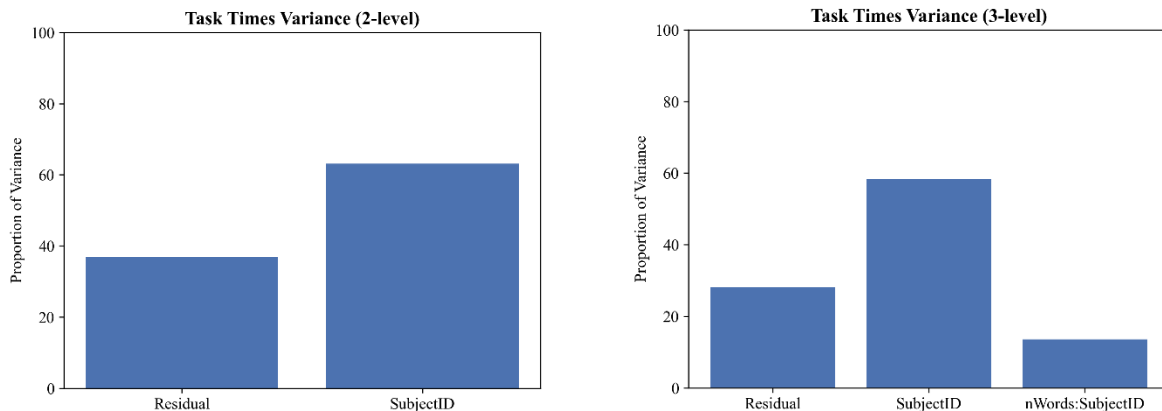


Figure 3.10 Variance decomposition of task times

3.2.2 Lateral deviation analysis

For the lateral deviation during tasks the data mostly follow a normal distribution, although there is a small outlier here as well. The sample mean is $x=0.42$ with a standard deviation of $s=0.36$. As the outlier has a value of 0.78 , this value is only one (+1) standard deviation away from the mean which means we will not drop this data.

In this part we will perform the same analysis as conducted with the model data; we need to establish the baseline performance of the participants so we can compare this with the models, as individual differences (i.e., individual overall driving skills) are likely to be more of influence.

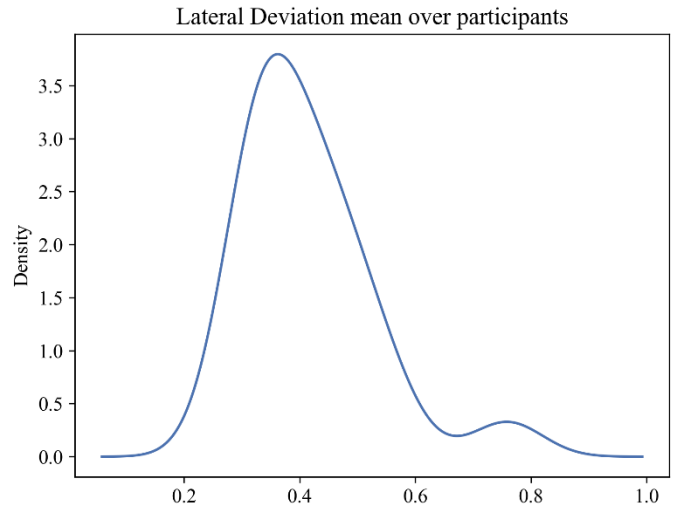


Figure 3.11 The distribution of the mean lateral deviations

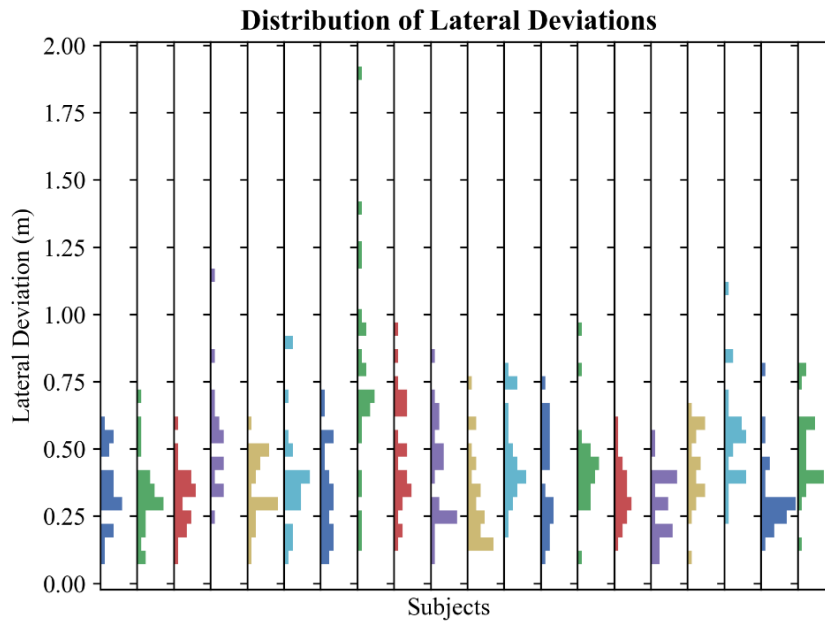


Figure 3.12 Lateral deviations over individual participants

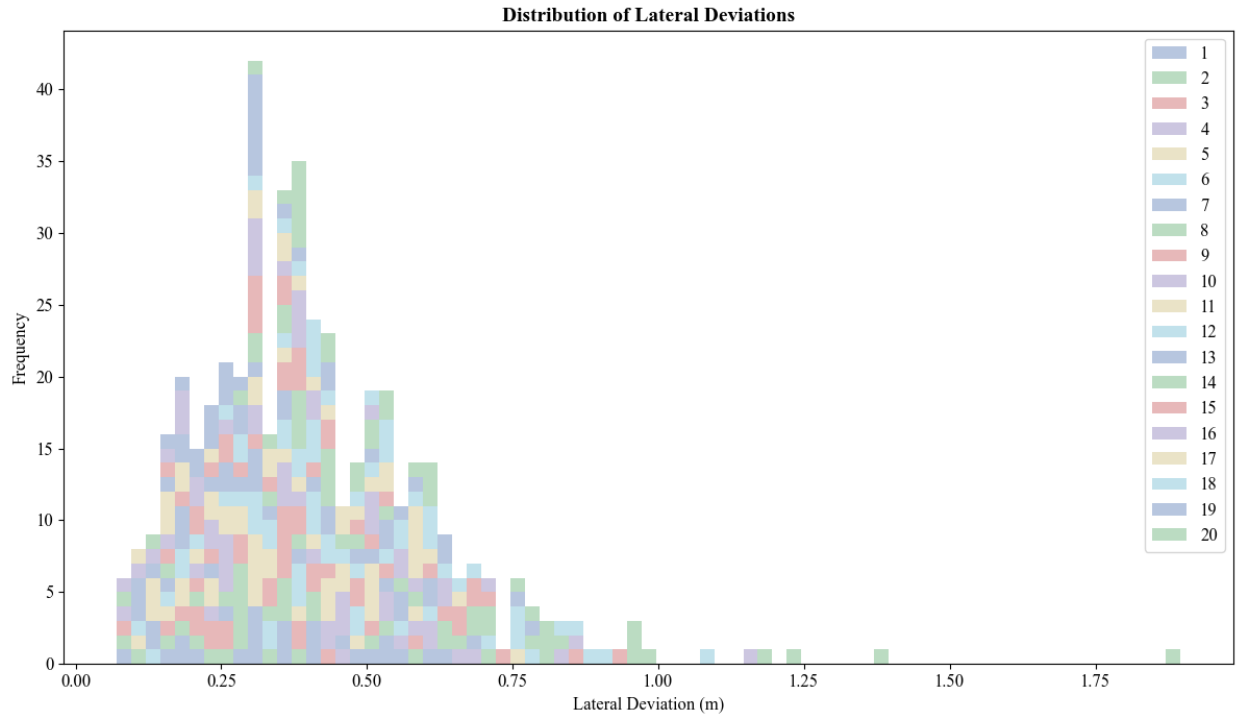


Figure 3.13 Stacked histogram of all individual lateral deviations

By computing the baseline driving performance, i.e., the driving performance when not performing a task, we find the values as indicated by the horizontal black lines in *figure 3.14*. The difference between the baseline and the mean lateral deviation during a task is the measurement of absolute lateral deviation, which in turn we then can then compare with the models.

Similar to the task times, we find that for longer sentences the mean values do increase, although the effect is less obvious. The standard deviations show that there is quite some variance, which can either be explained by the individual driver characteristics in the ability to keep the car centered for a particular participant or that this varies between separate tasks.

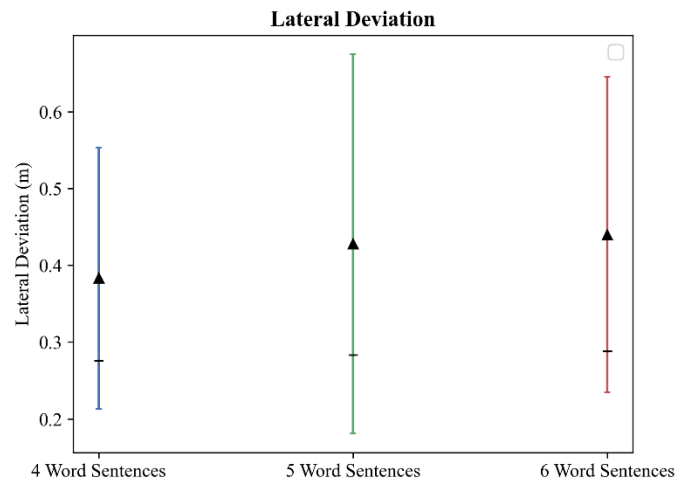


Figure 3.14 The distribution of the mean lateral deviations

Multilevel modeling analysis

Similarly, as before with the task times, we can also apply a linear mixed model here for the lateral deviations. If we apply this statistical analysis on all the mean lateral deviations grouped by the individual subjects, we get an intraclass correlation coefficient (ICC) of 0.231 , thus 23.1% of the variance is accounted for by the individual subjects. Since this value is lower than 0.5, we can interpret this as there being not much difference in lateral deviation between the participants (Park & Lake, 2005). This means that the group is more homogeneous in overall driving performance while being distracted if we were to compare this to the individual task times. Again, if we add the number of words to this model as a predictor variable, we find an ICC value of 0.229 , which shows that the influence of the number of words is even lower on lateral deviations as it was to the task times.

If we examine the 3-level model, where the 4,5- or 6-words groups are nested within the individual subjects, we can see in *figure 3.15* below that the number of words only attributes to a small proportion of the overall variance. If we test this with an ANOVA test between the three models, we get the following results:

Model	Parameters	Df	Pr(>Chisq)
1 Level model	2		
2 Level model	3	1	$< 2e^{-16}$
3 Level model	4	1	0.1422

This shows that by using an alpha level of $\alpha = .05$, the individual subjects attribute significantly to the variance. However, by adding in the number of words within the individual subjects, we find that this variable is insignificant (for an alpha level of $\alpha = .05$). Thus, it is evident that the number of words attributes significantly smaller to the lateral deviation variance than it did for the task time variance.

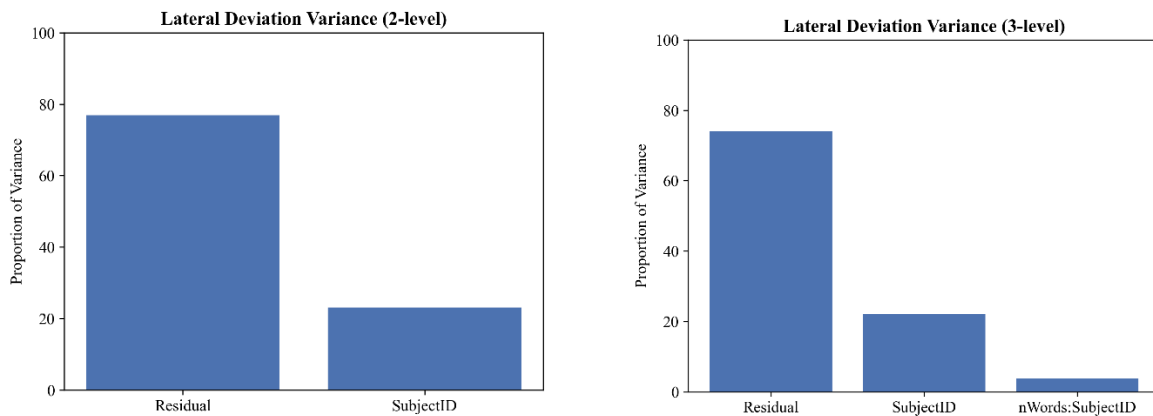


Figure 3.15 Variance decomposition of lateral deviation

3.3 Data Comparison

With the complete results from the simulations and the study, we can proceed by comparing the simulated data with the empirical data. In the discussion section, a more detailed analysis can be found.

3.3.1 Task time

Let us first examine the task times from the three different models with their different strategies and compare them with respectively the same sentence groups as from the empirical data in *Figure 3.16*.

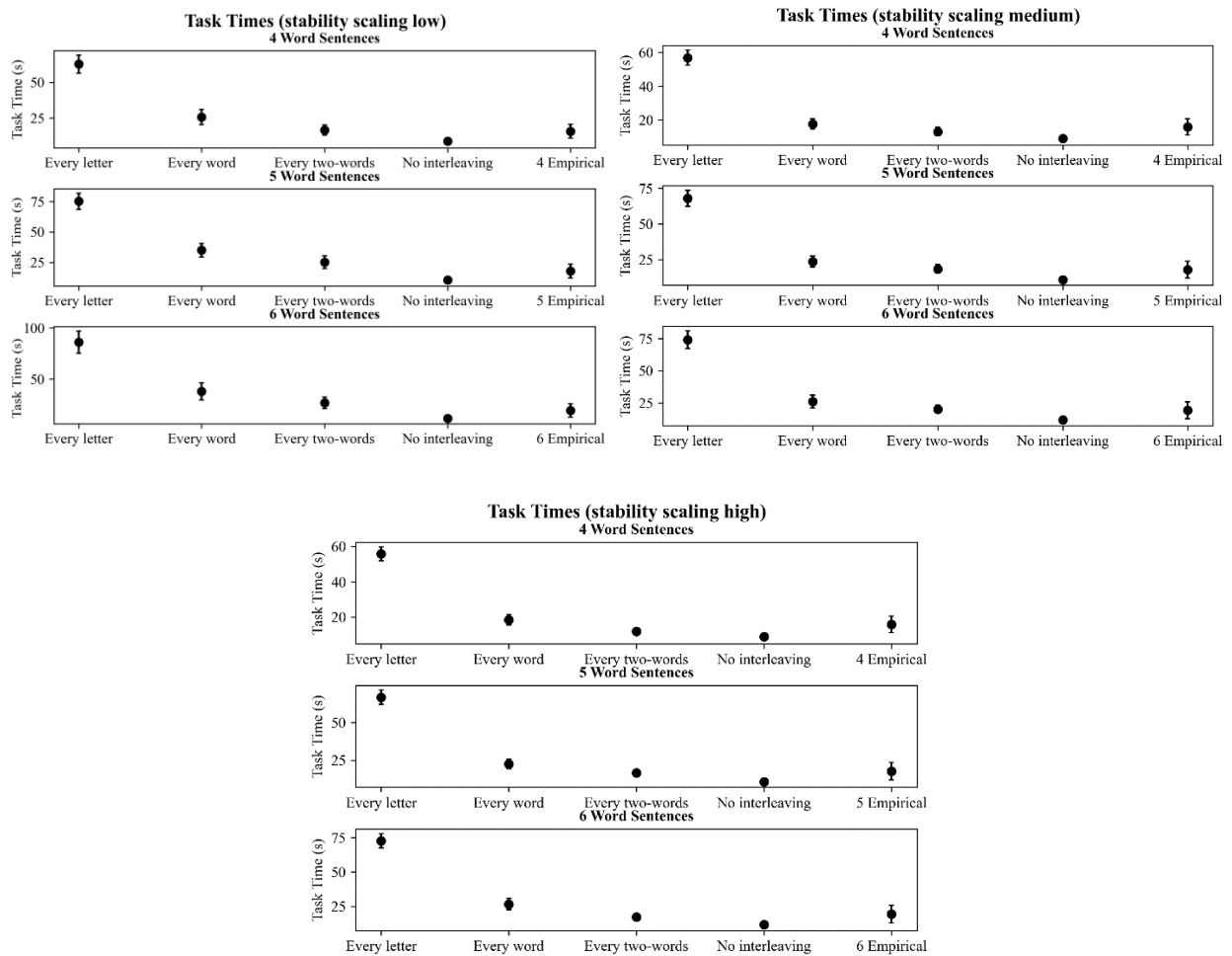


Figure 3.16 Task time compared with all strategies, sentences and models

As we have seen before in both the model and empirical data, the influence of the number of words per sentence did yield fairly different results. Only for the low scaling parameter, the empirical data falls out of range of the “Every word” interleaving strategy. Furthermore, the “Every two-words” is within range of the empirical data, independent of the stability scaling multiplier. Evidently, the “Every letter” strategy is least identical to the empirical data, followed by the “No interleaving” strategy.

If we were to aggregate all the sentences together, we find the results as shown in *Figure 3.17*. By taking out the *number of words* subcategories, both the single as two-word strategies are within range of the empirical data. Again, interleaving after every letter yields completely different results and by using no interleaving at all the task times are apparently too low.

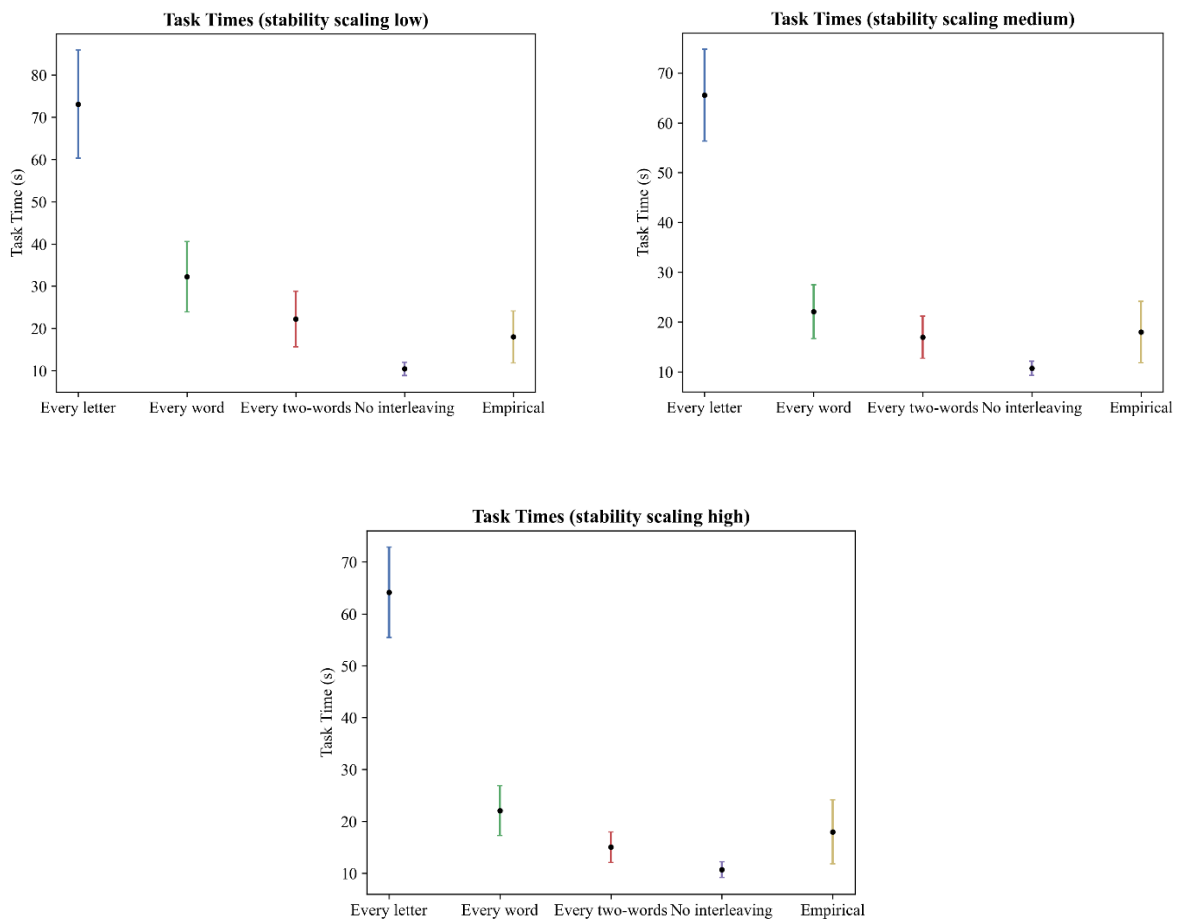


Figure 3.17 Task time aggregated over all sentence groups

3.3.2 Lateral deviation

As mentioned before, in the comparison between the models and the empirical data the baselines will be shifted to be aligned with the baselines of the models. This means that in the next graphs, the actual lateral deviation is higher (as in *Figure 3.5* and *Figure 3.14* respectively). In *Figure 3.18*, the prefix indicates the sentence group (e.g., number of words) and the suffix stands for the interleaving strategy which is shown in the table below.

0	1	2	3	E
Every letter	Every word	Every two-words	No interleaving	Empirical Data

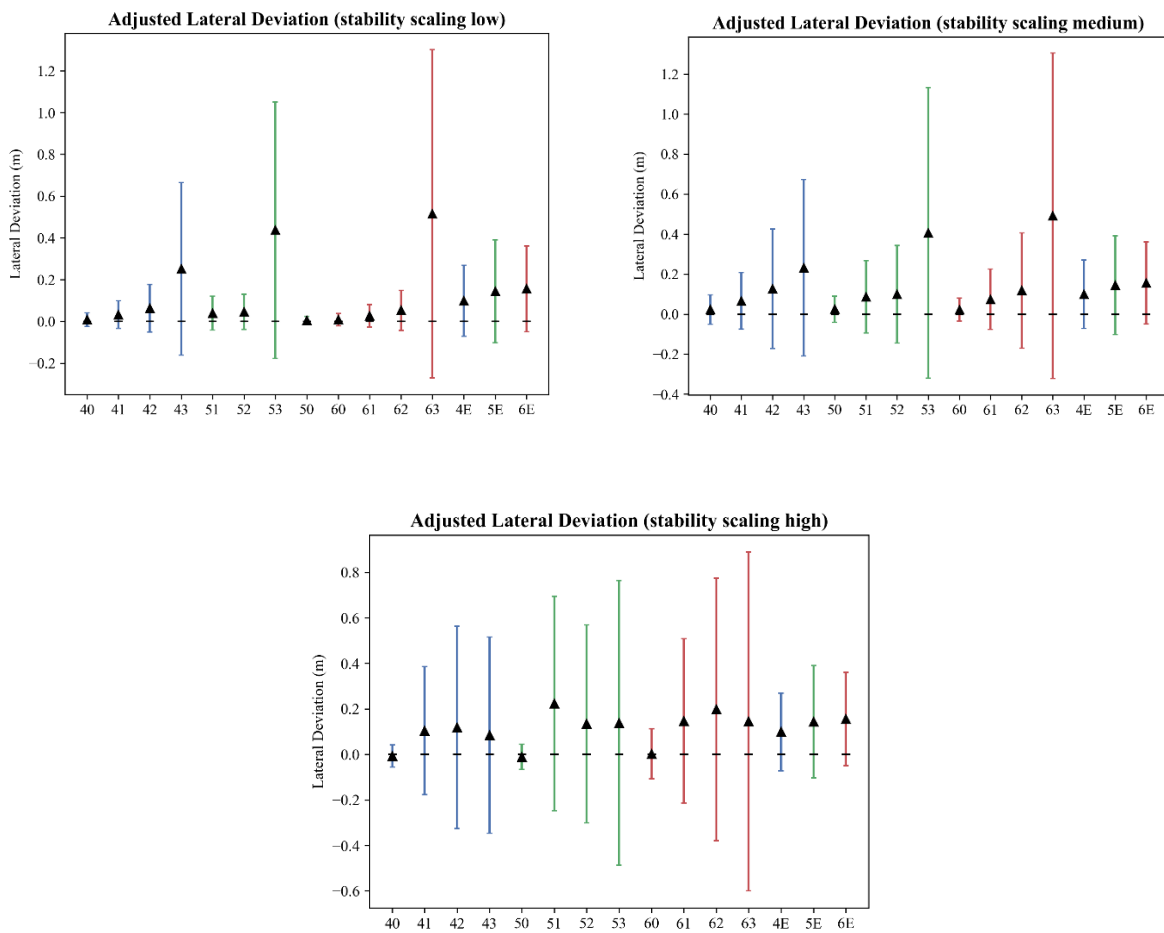


Figure 3.18 Absolute mean lateral deviation over all models and empirical data (blue: 4-word, green: 5-word, red: 6-word sentences)

In these graphs, it is interesting to see that the standard deviation of the empirical data is within range of all the outcomes of the simulated data, which makes the comparison less trivial. If we aggregate all the sentences, we get a better overview of how the model and the empirical data compare as is shown in *Figure 3.19*.

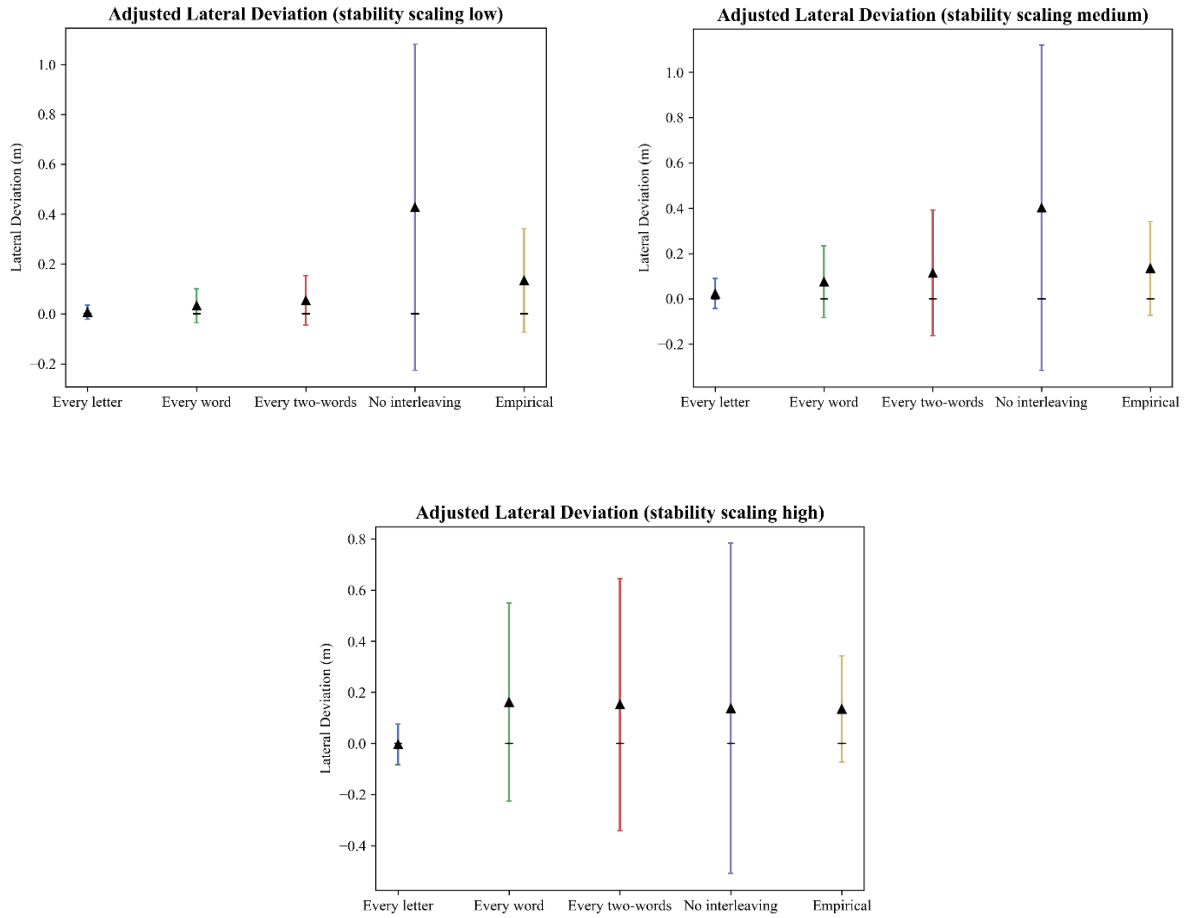


Figure 3.19 Lateral deviation (adjusted) aggregated over all sentence groups

4: General Discussion

The goal of this study was to model and simulate distracted driving behavior while texting. To evaluate the accuracy of these results, we have also conducted an empirical study. Since we have now gathered all the required results; in this section, we will discuss and interpret these within the scope of our predefined research questions and relevant previous studies on the subject.

By having 12 different models that generated an extensive amount of simulated driving data together with 20 participants having performed those same tasks, it is rather challenging to directly and quantitatively compare both datasets. Since cognitive models aim to explain the underlying processes of human decision-making and the resulting behavior from it, this means that one of the shortcomings of this approach is that humans are usually modeled as invariants rather than individuals, as data is mostly aggregated or averaged over subjects. This inherently means that the modeling itself assumes that there are no individual differences between subjects (Lee & Webb, 2005). This in itself does not pose an immediate threat for creating a general model of human multitasking, but it is something to keep aware of that not all individual differences are captured within the resulting model, which makes the quantitative comparison to the empirical data more challenging.

As different sentence lengths provided similar results, we will only discern the aggregated data; i.e., all sentence groups combined. From a more cognitive psychology interest, it would also be interesting to see whether this data can provide any insight into what interleaving strategies have been utilized when performing these tasks.

4.1 Task time

With the implementation of the transcription typing model, we find that the results when no interleaving is applied during driving are mostly identical to the predicted time it takes when the model is run stand-alone with those same sentences. Even though we did not measure the transcription typing times without driving, we assume these to be correct as the model has been extensively tested (Cao et al., 2018). The task times within the empirical data are longer than those singular task times, which is most likely the effect of the speed-accuracy trade-off, as the participants are balancing between performing the secondary task and keeping the car reasonably under control (Brumby et al., 2007). The question here though is whether one of our predefined strategies falls in the most general exercised interleaving strategy.

We can clearly recognize that the least comparable models (i.e., interleaving strategies) are those with every-letter and no-interleaving strategies. Even when taking the low and high stability scaling models into account, these strategies produce values that are the furthest from the empirical data. Therefore, we could make a case that those strategies seem unlikely to account for an accurate representation of the actual interleaving strategies used by the participants. However, we have to note here that by making this claim, we are utilizing the implicit assumption that the models are in fact able to simulate human behavior to some extent. Although this assumption could be refuted,

the mean task times of the most likely interleaving strategies; being the one or two words interleaving, are reasonably close to the empirical data mean task times. This is especially true for the medium stability assessment model.

It would be difficult, however, to make such a claim for which strategy is more likely or does resemble the truth of the remaining two. Unless when we are looking at the low stability scaling model, both models' mean task times are within one standard deviation of the mean task time within the empirical data. Although according to the simulated data from the models, the closest fit independent of the stability scaling value, remains the two-word interleaving model. It is interesting to see, that in a qualitative approach, the predicted task times of the two most likely models are actually within the range of the empirical data.

4.2 Driving performance

Unlike task time, driving performance was shown to be more ambiguous to interpretation. We can evidently see this in the rather large and overlapping standard deviations within both the empirical and simulated data. In the first instance, when looking at the raw data, there was a significantly large mismatch between the results of both datasets. This is why, in the results section, we have determined the baseline driving performance for both the models and the participants. By doing so, it is easier to compare the two, as the model has a much better baseline performance than the participants (i.e., mean lateral deviation when not performing any secondary tasks).

This mismatch between the model and human baseline driving performance is to be expected when we consult earlier studies. For example in an earlier study by Salvucci & Macuga (2002), the same driving model was used and here the graphs between the model and the human data are scaled as well to provide a better comparison. Also note that the variance within the empirical data is higher in this study, which confirms our findings as well (Salvucci & Macuga, 2002). The standard deviation values from the human data in this study are also comparable to an earlier empirical study based on handheld text entry (He et al., 2014).

With the absolute lateral deviation calculated, we can compare the two datasets as we are now analyzing the decrease in driving performance when performing the typing task. Here we can also make the case that the every-letter and no-interleaving strategies are most unlikely, as they fit the data the least. However, for the high stability scaling model, this is not entirely true; but we have to take into account that the baseline lateral deviation has risen there as well; which can lead to inaccurate results. If we would solely focus on the low and medium stability models, we can see that the every-word and two-word interleaving models again have the best fit in terms of driving performance. As in the case of the task times, the means of the two-word interleaving models have the closest resemblance to the empirical data.

Besides comparing the quantitative values, interestingly enough, we have seen the same type of behavioral patterns occurring within the simulated data and the empirical data. This is prevalent when the attention switches back from the secondary task to the driving itself. If the offset of the car in means of lane deviation exceeds a certain threshold, we see the same type of overcorrecting behavior occurring within both the models as the empirical data. This is to be expected from the model, as less frequent steering control actions result in larger steering actions (Mörzl et al., 2017).

4.3 Practical Implications

In theory, the data that resulted from simulating the models could be used for the training of a machine learning algorithm. Even though the applied training of such systems is beyond the scope of this study, the behavioral patterns that emerge from the distracted driving cognitive model can in fact be used for the classification of whether the driver is currently performing the secondary task. The question remains here though what level of accuracy this could theoretically reach. Another unknown is how well this could generalize, in performing the same classification for real humans. We could argue though that through the use of transfer learning fewer human data would be required since we see the same patterns emerge in the simulated data, albeit without the personal individual differences.

As the driving model parameters are kept at default values and are not varied over the different models, adjusting these could lead to more ‘individual’ results. This could be beneficial for the training of, for example, an advanced driving assistance system; as we have found a rather large mismatch between the human and simulated non-distracted driving performance (i.e., the models’ baseline driving performance is too accurate).

Since the results of the strategies “every word interleaving” and “every two-word interleaving” yield close results within the empirical data, a possibility would be to use both for the training of a classifier. As there might not be a single interleaving strategy utilized by humans, but the applied interleaving strategy depends on an unknown function of both individual preferences and prioritizing between safe driving or performing the secondary task as quickly as possible (i.e., the speed-accuracy trade-off (Janssen et al., 2012)). If the combination of both strategies would converge more to the actual perceived data, this could implicate that for tasks such as these it would be better to not focus on a single strategy for generating simulated data.

4.4 Limitations and Future Work

A limitation of the implementation was that the models all had to be run in real-time as the model was controlling the driving simulator *TORCS* in real-time. If the models could run faster or in parallel, it would be easier to vary over a large range of different parameter values. This would be most beneficial in analyzing a wider range of different interleaving strategies to find the best fit, as done with finding optimal strategies for phone dialing models (Janssen et al., 2012). This in turn could lead to better utilization of the threaded cognition theory (Salvucci & Taatgen, 2008), in contrast to the rather hard assumptions on interleaving that we have constrained the models to in this research.

Validation of the models would also be easier when the actual moments of interleaving would be further analyzed as performed by the participants in the study. Even though we have recorded all the touch input and the steering wheel input, determining the precise moments of interleaving are too ambiguous to lead to a conclusion. For future research use of eye or gaze tracking could be used to determine these moments more accurately, rather than deriving them from only the user input.

Comparable studies on distractions while driving has measured the driving performance based on other measurements than solely lane deviation, such as longitudinal deviation, braking response time and headway distance to other cars (Cao & Liu, 2013; Oviedo-Trespalacios et al., 2016; Salvucci, 2019; Salvucci & Macuga, 2002). It could be insightful for future studies to use a more detailed and real-world high-way scenario to examine the influence on these other performance measures.

4.5 Remarks

For being rather exploratory research, the results are looking promising. Following our reasoning, if we can confidently exclude the every-letter and no-interleaving strategies, the simulated data gives a good estimate of the task times and the decrease in driving performance, with the latter being more disputable and prone to variance. As for now, the model that resembles the empirical data within both modalities the closest, is the two-word interleaving model. This would also be backed by the literature, as the word-span typically lies between 2 and 8 words according to Salthouse (1986). However, we have to remark that the differences between one- and two-word interleaving strategies are often quite small, thus it is difficult to come to a definitive conclusion based on this data alone.

It is worth noting again, that by conducting this study we are implicitly assuming that *all* humans are using the same type of interleaving strategy. Although this could converge as shown in previous literature concerning similar tasks, to a somewhat optimal (or average) task performance trade-off. In that case, it would be promising to create computational cognitive models within cognitive frameworks as ACT-R to simulate human behavior in a variety of different multitasking scenarios, as simple assumptions with limited empirical validation could lead to possibly quite accurate simulations of real-world situations. This data in turn could be used as initial training data for more

advanced AI systems and fine-tuned to real humans by utilizing techniques such as transfer learning on the individual level (Yang et al., 2020). Especially since certain behavioral patterns tend to arise from the simulated data that reflects the empirical data quite well, having this data as prior could be beneficial for such systems.

References

- Alm, H., & Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis and Prevention*, 27(5), 707–715.
[https://doi.org/10.1016/0001-4575\(95\)00026-V](https://doi.org/10.1016/0001-4575(95)00026-V)
- Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, 111, 1036–1060.
<https://doi.org/10.1037/0033-295X.111.4.1036>
- Bechtel, W., & Graham, G. (Eds.). (1999). *A Companion to Cognitive Science* (Reprint edition). Wiley-Blackwell.
- Bhattacharyya, R., Wulfe, B., Phillips, D., Kuefler, A., Morton, J., Senanayake, R., & Kochenderfer, M. (2020). Modeling Human Driving Behavior through Generative Adversarial Imitation Learning. *ArXiv:2006.06412 [Cs]*. <http://arxiv.org/abs/2006.06412>
- Brumby, D. P., Howes, A., & Salvucci, D. D. (2007). A cognitive constraint model of dual-task trade-offs in a highly dynamic driving task. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 233–242.
<https://doi.org/10.1145/1240624.1240664>
- Cao, S. (2013). *Queueing Network Modeling of Human Performance in Complex Cognitive Multi-task Scenarios*.
- Cao, S., Ho, A., & He, J. (2018). Modeling and Predicting Mobile Phone Touchscreen Transcription Typing Using an Integrated Cognitive Architecture. *International Journal of Human–Computer Interaction*, 34(6), 544–556.
<https://doi.org/10.1080/10447318.2017.1373463>

- Cao, S., & Liu, Y. (2011). Integrating Queueing Network and ACT-R Cognitive Architectures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55, 836–840. <https://doi.org/10.1177/1071181311551174>
- Cao, S., & Liu, Y. (2013). Concurrent processing of vehicle lane keeping and speech comprehension tasks. *Accident Analysis & Prevention*, 59, 46–54. <https://doi.org/10.1016/j.aap.2013.04.038>
- CDC. (2020, December 8). *Distracted Driving*. https://www.cdc.gov/transportationsafety/distracted_driving/index.html
- Chakraborty, B., & Nakano, K. (2016). Automatic detection of driver's awareness with cognitive task from driving behavior. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. <https://doi.org/10.1109/SMC.2016.7844797>
- Choudhary, P., & Velaga, N. R. (2017). Analysis of vehicle-based lateral performance measures during distracted driving due to phone use. *Transportation Research Part F: Traffic Psychology and Behaviour*, 44, 120–133. <https://doi.org/10.1016/j.trf.2016.11.002>
- Choudhary, P., & Velaga, N. R. (2019). Effects of phone use on driving performance: A comparative analysis of young and professional drivers. *Safety Science*, 111, 179–187. <https://doi.org/10.1016/j.ssci.2018.07.009>
- Deng, C., Cao, S., Wu, C., & Lyu, N. (2019). Modeling Driver Take-Over Reaction Time and Emergency Response Time using an Integrated Cognitive Architecture. *Transportation Research Record*, 2673(12), 380–390. <https://doi.org/10.1177/0361198119842114>
- Deng, C., Wu, C., Cao, S., & Lyu, N. (2018). Modeling the effect of limited sight distance through fog on car-following performance using QN-ACTR cognitive architecture.

- Transportation Research Part F: Traffic Psychology and Behaviour*, 65, 643–654.
<https://doi.org/10.1016/j.trf.2017.12.017>
- Deng, C., Wu, C., & Lyu, N. (2017). Traffic sign recognition task cognitive integration model based on the ACT-R cognitive structure. *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, 337–342.
<https://doi.org/10.1109/ICTIS.2017.8047786>
- Fitch, G., Soccolich, S., Guo, F., McClafferty, J. A., Fang, Y., Olson, R. L., Perez, M. A., Hanowski, R., Hankey, J., & Dingus, T. (2013). The Impact of Hand-Held and Hands-Free Cell Phone Use on Driving Performance and Safety-Critical Event Risk. *Undefined*.
<https://www.semanticscholar.org/paper/The-Impact-of-Hand-Held-and-Hands-Free-Cell-Phone-Fitch-Soccolich/22b1600e2b1e9a147df0d1cdfb4057100e3edbae>
- Foreman, A. M., Friedel, J. E., Hayashi, Y., & Wirth, O. (2021). Texting while driving: A discrete choice experiment. *Accident Analysis & Prevention*, 149, 105823.
<https://doi.org/10.1016/j.aap.2020.105823>
- Groeger, J. A. (2000). *Understanding driving: Applying cognitive psychology to a complex everyday task* (pp. xvi, 254). Psychology Press.
- Haddington, P., & Rauniomaa, M. (2011). Technologies, Multitasking, and Driving: Attending to and Preparing for a Mobile Phone Conversation in a Car. *Human Communication Research*, 37(2), 223–254. <https://doi.org/10.1111/j.1468-2958.2010.01400.x>
- Haring, K., Ragni, M., Konieczny, L., & Watanabe, K. (2012). The use of ACT-R to develop an attention model for simple driving tasks. *CogSci*. <https://doi.org/10.17265/2159-5542/2013.04.002>
- Haring, K. S., Ragni, M., & Konieczny, L. (2012). *A Cognitive Model of Drivers Attention*. 6.

- He, J., Chaparro, A., Nguyen, B., Burge, R. J., Crandall, J., Chaparro, B., Ni, R., & Cao, S. (2014). Texting while driving: Is speech-based text entry less risky than handheld text entry? *Accident Analysis & Prevention*, *72*, 287–295.
<https://doi.org/10.1016/j.aap.2014.07.014>
- Huang, L., Guo, H., Zhang, R., Wang, H., & Wu, J. (2018). Capturing Drivers' Lane Changing Behaviors on Operational Level by Data Driven Methods. *IEEE Access*, *PP*, 1–1.
<https://doi.org/10.1109/ACCESS.2018.2873942>
- Huang, Z., Wu, J., & Lv, C. (2021). Driving Behavior Modeling Using Naturalistic Human Driving Data With Inverse Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 1–13. <https://doi.org/10.1109/TITS.2021.3088935>
- Janssen, C. P. (2012). *Understanding Strategic Adaptation in Dual-Task Situations as Cognitively Bounded Rational Behavior*.
- Janssen, C. P., Brumby, D. P., & Garnett, R. (2012). Natural Break Points: The Influence of Priorities and Cognitive and Motor Cues on Dual-Task Interleaving. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 5–29.
<https://doi.org/10.1177/1555343411432339>
- Khan, I., Rizvi, S. S., Khusro, S., Ali, S., & Chung, T.-S. (2021). Analyzing Drivers' Distractions due to Smartphone Usage: Evidence from AutoLog Dataset. *Mobile Information Systems*, *2021*, e5802658. <https://doi.org/10.1155/2021/5802658>
- Khodayari, A., Ghaffari, A., Ameli, S., & Fлахatgar, J. (2010). A historical review on lateral and longitudinal control of autonomous vehicle motions. *2010 International Conference on Mechanical and Electrical Technology*, 421–429.
<https://doi.org/10.1109/ICMET.2010.5598396>

- Kujala, T., & Salvucci, D. D. (2015). Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control. *International Journal of Human-Computer Studies*, 79, 66–78. <https://doi.org/10.1016/j.ijhcs.2015.02.009>
- Lee, M., & Webb, M. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, 12, 605–621. <https://doi.org/10.3758/BF03196751>
- Liu, Y. (2009). QN-ACES: Integrating queueing network and ACT-R, CAPS, EPIC, and soar architectures for multitask cognitive modeling. *Int. J. Hum. Comput. Interaction*, 25, 554–581. <https://doi.org/10.1080/10447310902973182>
- Martínez-Díaz, M., & Soriguera, F. (2018). Autonomous vehicles: Theoretical and practical challenges. *Transportation Research Procedia*, 33, 275–282. <https://doi.org/10.1016/j.trpro.2018.10.103>
- Minoiu Enache, N., Netto, M., Mammar, S., & Lusetti, B. (2009). Driver steering assistance for lane departure avoidance. *Control Engineering Practice*, 17(6), 642–651. <https://doi.org/10.1016/j.conengprac.2008.10.012>
- Mörtl, P., Festl, A., Wimmer, P., Kaiser, C., & Stocker, A. (2017). *Modelling driver styles based on driving data*. <https://www.semanticscholar.org/paper/Modelling-driver-styles-based-on-driving-data-M%C3%B6rtl-Festl/886b43e7aa679586a16ca15ac6868539829a14bf>
- Newell, A. (1990). *Unified theories of cognition* (pp. xvii, 549). Harvard University Press.
- Oulasvirta, A., Kristensson, P. O., Bi, X., & Howes, A. (Eds.). (2018). *Computational Interaction*. Oxford University Press.
- Oviedo-Trespalacios, O., Haque, Md. M., King, M., & Washington, S. (2016). Understanding the impacts of mobile phone distraction on driving performance: A systematic review.

- Transportation Research Part C: Emerging Technologies*, 72, 360–380.
<https://doi.org/10.1016/j.trc.2016.10.006>
- Park, S., & Lake, E. T. (2005). Multilevel Modeling of a Clustered Continuous Outcome. *Nursing Research*, 54(6), 406–413.
- Pekkanen, J., Lappi, O., Rinkkala, P., Tuhkanen, S., Frantsi, R., & Summala, H. (2018). A computational model for driver's cognitive state, visual perception and intermittent attention in a distracted car following task. *Royal Society Open Science*, 5(9), 180194.
<https://doi.org/10.1098/rsos.180194>
- Pohl, J., Birk, W., & Westervall, L. (2007). A driver-distraction-based lane-keeping assistance system. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 221(4), 541–552.
<https://doi.org/10.1243/09596518JSCE218>
- QN-ACTR-Release*. (2019). [Java]. HOMlab. <https://github.com/HOMlab/QN-ACTR-Release>
(Original work published 2017)
- R Core Team. (2020). *R*. <http://www.r-project.org/index.html>
- Reback, J., jbrockmendel, McKinney, W., Bossche, J. V. den, Augspurger, T., Cloud, P., Hawkins, S., gfyong, Roeschke, M., Sinhrks, Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., ... Seabold, S. (2021). *pandas-dev/pandas: Pandas 1.3.3*. Zenodo. <https://doi.org/10.5281/zenodo.5501881>
- Rehman, U., Cao, S., & MacGregor, C. (2019). Using an Integrated Cognitive Architecture to Model the Effect of Environmental Complexity on Drivers' Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 812–816. <https://doi.org/10.1177/1071181319631313>

- Rossum, G., & Drake, F. (2009). Python 3 Reference Manual. *Undefined*.
<https://www.semanticscholar.org/paper/Python-3-Reference-Manual-Rossum-Drake/fc1d1ba95d4b1ce49d50ac82f553b6236305b0b6>
- RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL
<http://www.rstudio.com/>.
- Salthouse, T. A. (1986). Perceptual, cognitive, and motoric aspects of transcription typing.
Psychological Bulletin, 99(3), 303–319. <https://doi.org/10.1037/0033-2909.99.3.303>
- Salvucci, D. (2002). *Modeling Driver Distraction from Cognitive Tasks*.
<https://doi.org/10.4324/9781315782379-171>
- Salvucci, D. D. (2006). Modeling Driver Behavior in a Cognitive Architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(2), 362–380.
<https://doi.org/10.1518/001872006777724417>
- Salvucci, D. D. (2019). Modeling Driver Distraction from Cognitive Tasks. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (1st ed., pp. 792–797). Routledge.
<https://doi.org/10.4324/9781315782379-171>
- Salvucci, D. D., Boer, E. R., & Liu, A. (2001). Toward an Integrated Model of Driver Behavior in Cognitive Architecture. *Transportation Research Record*, 1779(1), 9–16.
<https://doi.org/10.3141/1779-02>
- Salvucci, D. D., & Macuga, K. L. (2002). Predicting the effects of cellular-phone dialing on driver performance. *Cognitive Systems Research*, 3(1), 95–102.
[https://doi.org/10.1016/S1389-0417\(01\)00048-1](https://doi.org/10.1016/S1389-0417(01)00048-1)

- Salvucci, D. D., Mandalia, H. M., Kuge, N., & Yamamura, T. (2007). Lane-Change Detection Using a Computational Driver Model. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(3), 532–542.
<https://doi.org/10.1518/001872007X200157>
- Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*(1), 101–130.
<https://doi.org/10.1037/0033-295X.115.1.101>
- Salvucci, D. D., & Taatgen, N. A. (2011). Toward a Unified View of Cognitive Control. *Topics in Cognitive Science*, *3*(2), 227–230. <https://doi.org/10.1111/j.1756-8765.2011.01134.x>
- Salvucci, D. D., Zuber, M., Beregovaia, E., & Markley, D. (2005). Distract-R: Rapid prototyping and evaluation of in-vehicle interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '05*, 581.
<https://doi.org/10.1145/1054972.1055052>
- Salvucci, D., & Gray, R. (2004). A Two-Point Visual Control Model of Steering. *Perception*, *33*, 1233–1248. <https://doi.org/10.1068/p5343>
- Stothart, C., Mitchum, A., & Yehnert, C. (2015). The Attentional Cost of Receiving a Cell Phone Notification. *Journal of Experimental Psychology. Human Perception and Performance*.
<https://doi.org/10.1037/xhp0000100>
- Sun, R. (2008). Introduction to Computational Cognitive Modeling. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 3–20). Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772.003>
- Wymann, B., Dimitrakakis, C., Sumnery, A., & Guionneauz, C. (2015). *TORCS: The open racing car simulator*.

Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). *Transfer Learning*. Cambridge University Press.

Appendix

Informed Consent

The study “**Validation of a Cognitive Driving Multitasking Model**” is conducted as part of an Artificial Intelligence Master’s thesis carried out at Utrecht University.

By signing this document, I confirm that:

- I have read and understood the general information of this research, and have been informed of all the information listed in the information letter.
- I have been given the opportunity to ask questions and I had sufficient time to decide whether I participate.
- I know that my participation is completely voluntary. I know that I can refuse to participate and that I can stop my participation at any time during the study. And I can withdraw permission to use my data up to 2 months after my participation.
- I understand and agree with how my data will be handled.
- I agree with my participation in this study.

Name:

Date:

Signature:

Information Letter

Welcome to the study “Validation of a Cognitive Driving Multitasking Model”. This study is conducted as part of an Artificial Intelligence Master’s thesis, carried out at Utrecht University. In this document, you will read about the purpose of the study, study procedure, management of your data, and your rights as participants. Please read these statements carefully before you proceed.

Purpose: Driving while not having full attention on the road is an important reason for the occurrence of accidents. In this study, we try to gain a better understanding of the behavior that emerges when one is distracted while driving. This to research how well we can predict specific situational behavior with the use of cognitive models that have been created prior to this experiment. In this experiment, we will monitor your actions while doing two different tasks. The main objective is to drive a car down the highway and maintain your lane as well as possible. Your secondary objective is to type certain given sentences on a smartphone when prompted to. Even though this is a simulated environment, we hope you can immerse yourself in the experiment and feel as natural as possible, since this will help us gain the best overall insight.

Procedure: After giving the informed consent, you will start the experiment. You will drive a car on a highway without any other cars or other distractions. After you have familiarized yourself with the simulated car driving, the main experiment will begin. From this point on, whenever an audio notification plays on the smartphone, you will be instructed to type the sentence that is shown on the screen. The experiment itself is expected to take around 15-20 minutes.

Voluntariness & Anonymity: Your participation in this research is voluntary. You can withdraw at any time without consequences of any kind. All materials associated with the study are confidential and will be used only for the purpose of this research. No identifying information will be collected. The resulting data will be published anonymously and cannot be assigned to you.

Privacy & Data Management: All materials associated with the data collection are confidential. The data that will be recorded during this experiment includes all interactions with the smartphone and the driving simulator. Your data will only be stored only locally and stored in such a way that this data cannot be associated with you. Additionally, the anonymized data will be made publicly

accessible in the final results of this study. According to the General Data Protection Regulation (GDPR), you have the following rights:

- Right of access by the data subject (Art. 15)
- Right to withdraw your consent (Art. 7)
- Right to rectification (Art. 16) - Right to erasure (Art. 17)
- Right to restriction of processing (Art 18)

Questions and complaints: If you have more questions about the study, please contact f.a.j.fikkert@students.uu.nl. If you have any concerns, please contact the complaint officer (klachtenfunctionaris-fetsocwet@uu.nl) and the central privacy (privacy@uu.nl).