# How Multitasking Influences the Usage of a Voice Assistant

**Kirti Biharie , Claudia Hauff**
TU Delft
k.s.biharie@student.tudelft.nl, c.hauff@tudelft.nl

## Abstract

This paper shows the influence of multitasking on the usage of a voice assistant. Voice assistants allow users to input queries over a speech-only channel, and as a result they do not require the same attention as a traditional search engine. Existing research describes the effects of the use of a voice assistant on driving and other demanding activities, however, there is no research that describes how that demanding activity influences the usage of the voice assistant. To research these effects, three sub questions have been constructed and answered. These tackled three characteristics to describe the usage: the query formulation, the knowledge gain and the user experience. To answer these questions a user study was conducted where the participants used a voice assistant in three situations. Two of these situations included a distraction in the form of a game. We found that the presence of dual tasking results in shorter queries. We also found that a higher intensity of the session can decrease the knowledge gain.

## 1 Introduction

Voice assistants are gaining more popularity. In 2022, 73% of the consumers are expected to use voice assistants in a car, compared to the 49% in 2019 [20]. The voice assistant allows the consumer to play music, check directions and control other car functions with only their voice. This is done by inputting a query over a spoken channel. This query can be a whole sentence instead of just keywords or commands. Because of this level of freedom, voice assistants can almost mimic a real conversation, which is what conversational search aims to do. Conversational search uses natural language processing which allows for more complex grammatical sentences and the use of context from previous queries. The returned results can be accompanied by auxiliary visuals. Voice assistants can make use of conversational search to process the queries, which decreases the gap between voice assistants and actual conversations. Popular voice assistants are Alexa, Siri and Google Assistant, of which the latter two are built-in in smartphones.

Traditionally, queries are input through a user interface with a traditional search engine. To input the query, physical actions are necessary. The search engine displays the results on the screen, which requires visual attention from the user. Previous research [17] has shown that cellphone usage can cause driving accidents. Four types of driver distractions are described in [21]: visual, cognitive, physical and auditory. A traditional search engine thus distracts a driver visually, physically and cognitively.

Since voice assistants work over a speech-only channel, they do not require the same visual or physical attention as a traditional search engine. A voice assistant primarily needs cognitive attention, however, hands-free cellphone usage does not come without accidents either [12]. Previous research has also shown what kind of distraction is caused by cellphone usage [3]. The effect of the speech recognition accuracy on driving performance is known as well [11]. However, there is currently no research on the influence of driving on the usage of a voice assistant. Driving is only one use case of a voice assistant. A notable use case is playing video games, because this has been used to simulate driving [18]. Other use cases include the bathroom, cooking, watching television, and working. These use cases describe a primary activity. The primary activity has the highest priority while the secondary activity, the voice assistant, has a lower priority.

Research to this date has not determined the influence of a primary activity and the presence thereof. This information can be used to improve the design of voice assistants. Even though in-car voice assistants are gaining popularity, nearly 50% of the consumers agree that the built-in voice assistants should be improved [20]. Though driving is only one use case of a voice assistant, it is the most researched one. However, for this research it is not possible to investigate driving, because of the difficulties and complications that come with it. Thus the use case of playing video games will be used.

To improve the usage of voice assistants as a secondary activity, we need to investigate the usage differences between a voice assistant as a primary activity and as a secondary activity. One usage characteristic is the query formulation, which can be expressed using query length, session length and number of queries. The knowledge gained from a session is another important factor and can explain how much attention was given to the voice assistant. The last factor, user satisfaction, accounts for the experience from the user's perspective. This paper will answer the question: "How does the usage of the system change when voice search is the secondary activity instead of the primary activity?" Three subquestions were constructed to answer this.

**RQ1:** How does the query formulation change when using the voice assistant, with conversational search, is the secondary activity instead of the primary activity?

**RQ2:** How does the knowledge gain change when using the voice assistant, with conversational search, is the secondary activity instead of the primary activity?

**RQ3:** How does the user experience change when using the voice assistant, with conversational search, is the secondary activity instead of the primary activity?

To answer these subquestions, a user study was conducted where the participants used a voice assistant while multitasking. While using the voice assistant, the participants had to simultaneously play a game to simulate multitasking.

The paper first gives some background information on multitasking, video games and conversational search. The third section is concerned with the methodology used for this study. Section 4 will discuss the experimental setup in detail. In section 5 the results will be presented. In section 6 the reproducibility of the experiment will be discussed briefly. This will be followed by the discussion of the results in section 7. The last section will summarize the research and formulate recommendations for future work.

## 2 Background

Multitasking is often defined as "doing multiple things at the same time" [5]. However, this can be interpreted in two ways. First, this can be seen as doing multiple activities at the exact same time, for example, listening to music while writing a report. Secondly, this can be seen as continuously switching between multiple activities, for example writing a report while playing fetch. In the second case, all of the attention is either given to writing or playing fetch and alternates between the two. The difference between the two interpretations can be described as simultaneous versus sequential execution of activities. In the scope of this paper, multitasking is described as the simultaneous execution of multiple activities.

### 2.1 Dual tasking

Single tasking, the opposite of multitasking, involves only one activity. Dual tasking involves exactly two activities. The two activities can have a different priority, such that one is the primary activity and the other the secondary activity. In the conducted user study, the primary activity was represented by a game, while the voice assistant represented the secondary activity. These activities can also be either active or passive, manual or automatic. Research has shown that response time generally increases with dual tasking [15]. To explain this effect, three theories have been proposed.

The bottleneck model assumes that a neural network of the brain can only process one task at once. When two tasks use the same neural network, the second task is delayed until the first task has finished processing. The performance of the first task does not differ from single tasking, while the second one has a higher response time and may have a worse performance. This model explains the interference between two continuous tasks that try to access the same neural network in the brain[15].

The capacity sharing model [19] assumes that some processes have a limited capacity. When the processing of two tasks overlap, the processing capacity is divided between the two. This leads to longer processing times for both tasks and longer response times than single tasking. If the difficulty of one tasks increases, the overall processing time increases, as well as the response time of the other task. This model allows multiple tasks to be processed at the same time, while the bottleneck model only allows the tasks to be processed sequentially.

The multiple resources model [22] claims that dual task interference will only occur if the two tasks use the same resources. If the two tasks use different resources, the response rate and performance should not differ from a single task situation. According to this model, tasks using different stimuli should not interfere with each other. Thus, if the game and voice assistant use different stimuli such as visual and auditory stimuli, they should not interfere with each other as much.

### 2.2 Games

Video games are a common leisure activity for many people. However, they can also be seen as cognitive consuming tasks. There has been various research [1; 4] on the positive and negative effects of gaming and games have also been used to simulate situations [18].

Games can be grouped into different categories that each use the brain in a different way. In [2], 20 different minigames were tested to link the games to cognitive abilities. The games were grouped in four different categories: reasoning, working memory, attention games and perceptual speed games. The perceptual speed games had the weakest links to the cognitive abilities, while the working memory and reasoning games had the strongest links to various cognitive capabilities. This means that perceptual games can serve as a low-level distraction, while working memory and reasoning games can serve as a high-level distraction.

### 2.3 Conversational search

Conversational search aims to mimic a human conversation as closely as possible. Users are allowed to create complex sentences and refer to previous interactions as they would while talking to other people. Natural language processing is used to understand a complex query and can translate it into a regular query which is inputted into a search engine. Natural language processing is capable of understanding reference words such as "that" or "then" and replacing them by the referenced words from the context.

Conversational search can be combined with voice assistants such that the interaction with a voice assistant is more similar to the interaction with a human assistant. Conversational search allows for queries such as "How old is he?", where *he* refers to the subject of the previous question. Google Assistant is a popular voice assistant that makes use of this system.

## 3 Methodology

To answer the subquestions, a user study was conducted. The setup was based on a study [7] where users take a knowledge test and perform a search task, followed by another knowledge test. The knowledge gain during the search task can be calculated, with these two knowledge tests. This naturally can answer the second research question about knowledge gain. The setup of their study can easily be adjusted to use a voice assistant and to answer the other two research questions regarding query formulation and user experience. Our study consists of three of those sessions, which is shown in figure 1 and set up as follows:

**Participants**
Participants were recruited from crowd-sourcing platform Prolific[1]. All of the participants were native English speakers. The participants needed a Prolific acceptance rate of at least 90% and at least 50 submissions. They also needed to use a desktop with access to a microphone and headphones or speakers. The participants agreed to use no external sources during the study. The participants filled in three general questions to familiarize themselves with voice assistants.

**Randomization**
Every participant was assigned to the three sessions and three topics in a random order.

**Sessions**
Every participant engaged in three sessions: (1) single task; (2) dual task with low distraction; (3) dual task with high distraction. With these three sessions, we can investigate two different scenarios. First, we can analyze the usage differences between session 1 and session 2/3 combined, which will show the influence of the presence of a primary task. Secondly, we can analyze the usage differences between sessions 2 and 3. This will show the influence of the intensity of a primary task. Every session revolved around one of the following topics taken from [7]: (1) American Revolutionary War; (2) Carpenter Bees; (3) USS Cole Bombing. The topics were chosen because the knowledge gain was comparable on average across the three topics. The topics and context did not contain any hard to pronounce words. This reduces the influence of the quality of the speech recognition. A session started with a knowledge test about a topic, followed by a task about the same topic and another knowledge test. The session was concluded with an evaluation.

**Voice assistant**
The microphone from the participant was used to record their speech. The participant did not need to activate the voice assistant for every query, instead the participant could start talking at any time during the task. The Google Speech API was used to convert the spoken query to text. The textual query was sent to a server running Macaw [23], which is a conversational information seeking platform. The server returned the answer to the query and used text-to-speech to read it aloud to the user.

**Knowledge tests**
The knowledge tests were structured according to [7]. The participant took a knowledge test before and after the task in all three sessions. A knowledge test consisted of 10 statements, to which the user could respond with "TRUE", "FALSE" or "I DON'T KNOW". The first knowledge test of the session was used to serve as a baseline for user's knowledge of the topic. The second knowledge test was exactly the same as the first knowledge test and the outcome represented the user's knowledge of the topic after the task. Together with the baseline, the knowledge gain during the session could be calculated.

**Tasks**
The single task session only contained the conversational search without a distraction. The user had a maximum of 7 minutes to ask questions about their topic to the voice assistant, however, the user was allowed to finish earlier after a minimum of 4 minutes. Some keywords from the topic were displayed on screen to help the participants come up with questions, as this was shown to be quite hard in a pilot. The pilot was held with colleagues and showed that formulating questions about a topic is hard without a lead. This was partly due to the fact that the participant could not input only the topic as a query to gain general information about the topic. Instead only targeted questions were allowed. The keywords were not included in the original study [7], but in that study a lot more information could be gained from a single query, which could then lead to the formulation of follow-up questions. The difficulties of the games were also decreased as a result of the pilot.

The low distraction dual task session contained a game in addition to the voice assistant. The game was based on Alphattack[2]. The goal of the game was to press the keys corresponding to the characters shown on the screen. No audio effects were used. The category of the game was perceptual speed [2], which needs the least cognitive ability of the games mentioned and can therefore serve as a low distraction. The high distraction dual task session contained a game based on Two Three[3] instead of Alphattack. In this game the user had to shoot down numbers, by bringing them exactly down to zero, by only using the numbers two and three. The mouse was used to aim and the keyboard keys to choose the projectiles. This category of this game was reasoning and required more cognitive capability.

**Evaluation**
At the end of each session, the participants filled in an evaluation. The participants answered on a 5-point Likert scale to measure their attitude. Three questions were asked as described in [9]:

1. *How satisfied are you with your experience in this task in general?*
2. *How much effort did you put in to complete this task?*
3. *How well did the system recognize what you said?*

The three questions evaluate different aspects of the user experience. The first question gives a general impression

[1] https://www.prolific.co/

[2] https://www.miniclip.com/games/alphattack/

[3] https://armorgames.com/play/863/twothree

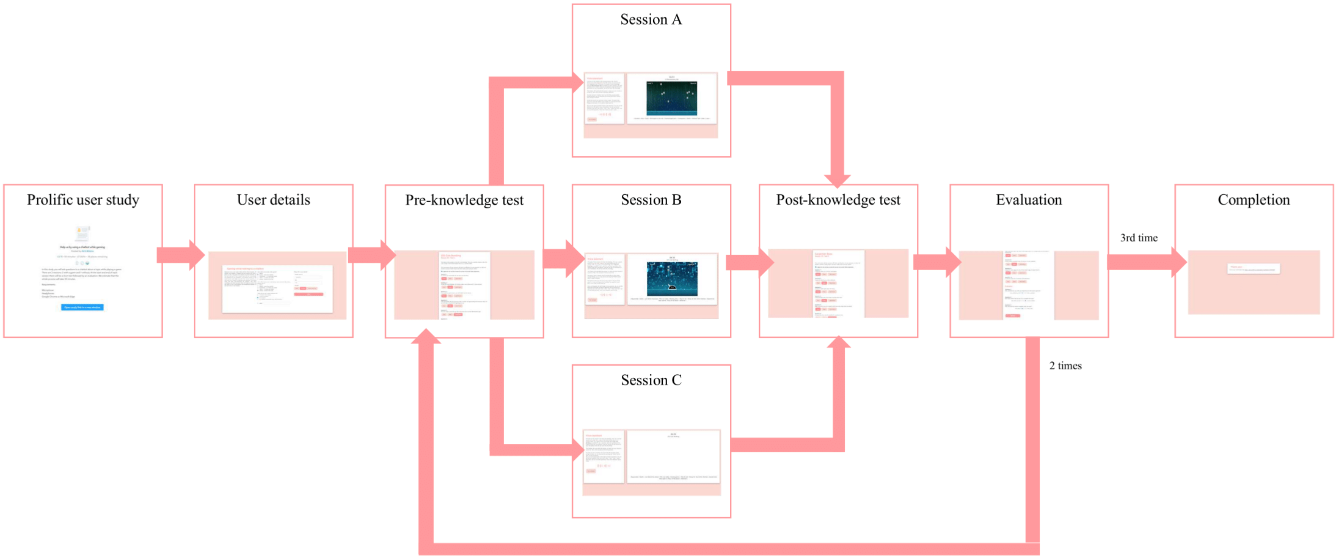

**Figure 1:** The process for one participant.

on how the participant experienced the session. The second question expresses how much effort was necessary for that session. With the results to this question, we can determine whether a distraction of higher intensity required more effort. The last question investigates whether the users's perception of the speech recognition quality differs based on the session. The three questions together can give an accurate image of the user experience.

**Metrics**
The first research question regarding query formulation used three different metrics. These metrics are the session length, the number of queries and the query length. The second research question uses the knowledge gain. This is calculated as the increase of correct answers of the post-knowledge test compared to the pre-knowledge test. An “I DON’T KNOW” was seen as an incorrect answer. The metrics for the first two research questions were also used in [7]. The third research question uses a Likert scale for each evaluation question. With the use of these metrics an accurate evaluation can be given to answer the main research question.

# 4 Experimental Setup

This section will first explain in detail how Macaw was used and what functionality was added to the framework to adapt it to our user study. This is followed by a setup of the server used to run the study. Finally, the recreation of the two games will be discussed.

## 4.1 Macaw

Macaw is an extensible conversational information seeking platform. Macaw provides several interfaces such as the command line and Telegram using a bot. It supports multiple types of tasks of which document retrieval and question answering are examples. It also supports a wizard of oz setup, a setup where researchers themselves provide answers instead of the search engine. Macaw is conversational, meaning that it stores consecutive queries and answers from users and uses those when processing a new query. The processing happens with DrQA, a system for reading comprehension. With the question answering task, the query is sent to a search engine, either Bing or Indri, which returns a web page containing the answer. DrQA searches through the page for an answer, which is then outputted by Macaw to the chosen interface.

For this user study, Macaw was setup inside a docker container running Ubuntu. Besides the original interfaces, an extra interface was added that starts the server and handles requests from the website. To increase the chance that DrQA could return an answer, text files were added to Macaw that explicitly contained the answers to the test questions. DrQA first searched through these text files, and only if no answer was found there, it would search the web using Bing. The text files contained multiple instances of the same answer, each using synonyms or with a different sentence structure. This increased the quality of the returned answers compared to the original code.

The main limitations originate from DrQA. DrQA currently does not support true or false questions, which has as a consequence that the participants cannot validate a statement. DrQA performs best with questions that start with “why”, “who”, “when”, “where” and “how”. Questions not starting with these words perform worse. The system also returns fairly short answers that usually consist of just one word. For example, the question “When did that happen?” can receive “2000” as an answer instead of the full date “12 October 2000”.

## 4.2 Website

To make the study easily accessible to participants, a website was published with the user study as its only purpose. The website had a simplistic design with no external distractions and used a soft color palette. This design enables the

**Table 1:** The session length (SL) in minutes, average query length (in words) and the average knowledge gain (KG) per participant.

| Session | Session Length (SL) (in mins) | #Queries | Query length (in words per query) | Query length per user (in words per query) | Knowledge Gain |
|---|---|---|---|---|---|
| Single tasking | $4.03 \pm 0.81$ | $24.88 \pm 7.88$ | $5.31 \pm 2.27$ | $5.48 \pm 1.33$ | $2.92 \pm 2.34$ |
| Dual tasking (easy) | $4.45 \pm 1.39$ | $22.69 \pm 10.82$ | $4.84 \pm 2.27$ | $4.95 \pm 1.01$ | $3.32 \pm 2.67$ |
| Dual tasking (hard) | $4.25 \pm 1.54$ | $20.69 \pm 11.88$ | $5.18 \pm 2.54$ | $5.21 \pm 1.60$ | $2.05 \pm 2.03$ |
| Dual tasking combined | $4.35 \pm 1.47$ | $21.69 \pm 11.40$ | $5.00 \pm 2.41$ | $5.08 \pm 1.35$ | $2.68 \pm 2.46$ |

**Table 2:** The average user satisfaction, effort and speech recognition quality based on the three evaluation questions: 1) *How satisfied are you with your experience in this task in general?* 2) *How much effort did you put in to complete this task?* 3) *How well did the system recognize what you said?*

| Session | User Satisfaction (1-5) | Effort (1-5) | Speech Recognition (1-5) |
|---|---|---|---|
| Single tasking | $2.94 \pm 1.20$ | $4.50 \pm 0.71$ | $2.34 \pm 1.11$ |
| Dual tasking (easy) | $2.88 \pm 0.93$ | $4.50 \pm 0.71$ | $2.38 \pm 0.93$ |
| Dual tasking (hard) | $2.56 \pm 1.22$ | $4.19 \pm 0.81$ | $2.13 \pm 0.86$ |
| Dual tasking combined | $2.72 \pm 1.10$ | $4.34 \pm 0.77$ | $2.25 \pm 0.90$ |

**Table 3:** The average knowledge gain per topic

| Topic | Knowledge Gain |
|---|---|
| USS Cole Bombing | $4.39 \pm 2.47$ |
| Carpenter Bees | $2.69 \pm 1.87$ |
| Revolutionary War | $1.20 \pm 1.71$ |

participants to focus more on the task. The user study in its entirety is done on this website and does not require any external sources. The surveys were embedded in the website with HTML forms. The common alternative for surveys, Google Forms, would require users to fill in their Prolific ID for every survey. Since there were 6 surveys per participant, this would be prone to error if the participants would have to fill in their exact Prolific ID repeatedly. Another downside to Google Forms is that it has no straightforward way to confirm that the participant indeed submitted the survey. The participant could finish the user study and receive a contribution without filling in a survey. Therefore, HTML forms were a better alternative.

To create the server side of the website, Flask was used. Flask is a lightweight web application framework and allows for the creation of a server with little code. The requests sent from the website were all processed by Flask. The server also handled most of the logging. The submitted forms were stored in a csv file on the server. Besides that, the user details and session lengths were also logged. However, the queries and answers were stored with MongoDB, as this was already implemented in Macaw. To get the textual queries from the spoken queries, the Web Speech API was used. This API is an experimental technology that offers both a way to convert speech to text as well as a way to convert text to speech. The participants used their microphone to input the query. This required the browser to access their microphone. This again required an HTTPS connection, which in turn required an SSL certificate. Even though the website does not handle any sensitive data or personal information, an SSL certificate gives the users more trust in the user study.

The Web Speech API is an experimental technology and currently only supported in Microsoft Edge and Google Chrome. Other popular browsers such as Firefox and Safari are thus not supported. This limits the participants eligible for the user study. The Web Speech API has some limitations itself. The quality of the speech recognition is good, but not flawless. It is unable to differentiate similar sounding words. It recognizes “eyes” as “ice”, for example. Harder, uncommon words such as “Xylocopa”, the genus of carpenter bees, are often not recognized.

### 4.3 Games

The user study used two games to simulate a primary activity: Alphattack and TwoThree. Both games are hosted on minigame websites. The games required the users to enable Adobe Flash. Adobe Flash is known to have many vulnerabilities and will receive no further support starting 2021. It would not be responsible to use these games as is and expose the personal computers of participants to these vulnerabilities. The games also had different difficulty settings and required users to restart the games when they lost. This would give a different experience of the game to different participants. Because of these two reasons, the games were recreated for this study.

The games were recreated using Cocos Creator, a game development engine, and exported to HTML5, an alternative to Adobe Flash. All the assets used were either free to use or created by the authors. The games were simplified to their main mechanics. For Alphattack, this meant that letters fell down on the screen. The users only had to press the corresponding keys on their keyboard to make the letters disappear. Other mechanics, such as bombs which removed all the

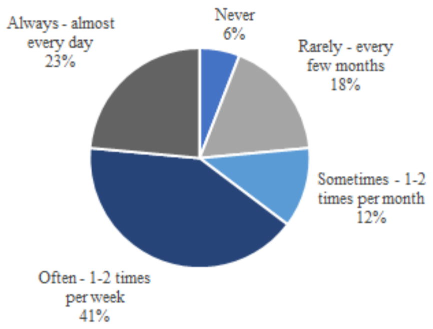

**Figure 2:** The answers to the first general question "How often do you play video games?"

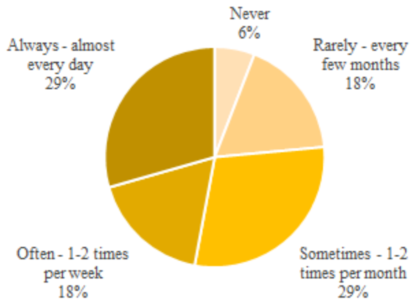

**Figure 3:** The answers to the second general question "How often do you use a chatbot or voice assistant (Siri, Alexa, Cortana, etc.)"

letters, or numbers for bonus points were not included. For TwoThree, the recreated game was fairly similar to the original game. The users had to press the 2 and 3 keys to bring the falling numbers down to zero. For both games, negative points were not included and thus losing was not possible.

In the original games, the difficulty increased with levels which simultaneously increased the attention necessary. For this user study, however, it was necessary for the games to require a consistent amount of attention throughout different levels. For this purpose, the levels were still included in the recreated games to keep a sense of progression, but the difficulty did not increase. In Alphattack, the difficulty is defined as the amount of letters on screen combined with the speed. In TwoThree, the difficulty also is decided by the value of the falling number: the higher the value, the higher the difficulty. The difficulty of the games were adjusted to increase the contrast between Alphattack and TwoThree, while keeping it reasonably possible to reach the given goal, namely level 10, in both games within the time limit.

# 5 Results

This section will first discuss the details of the participants. After that the three research questions will be handled.

## 5.1 Participants

19 Participants completed the entire study, however, 3 participants were filtered out because they either did not enter any queries or they mostly entered random queries not related to

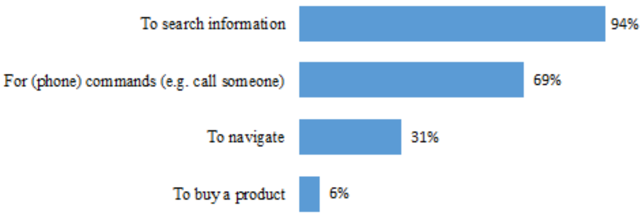

**Figure 4:** The answers to the third general question "What do you use a voice assistant for?"

the task topic at all. An issue also arose when the participant used speakers during the study. While the answer to the queries was being played over the speaker, it was also being picked up by the microphone as a new query from the user. This resulted in a lot of queries from users consisting of only "no response has been found". Answers being picked up by the microphone happened with 5 from the 16 remaining participants. Because it was clear that these queries were not actually spoken by the user, and to have somewhat enough data points, these queries were removed and not taken into account with the results. The participants also asked true or false questions, which were not supported by DrQA and did not return answers.

The 16 participants consisted of 5 males and 11 females between ages 19 to 49. Figure 2 shows the answers to the first general question regarding video games. 81% of the participants play video games at least every month. A report [14] showed that around 67% of Americans played video games at least every month, which is less than in this study. The answer to the second general question about voice assistants is shown in figure 3. 76% of the participants use a voice assistant at least every month, compared to 35% of Americans [16]. The uses of the voice assistant are shown in figure 4. The user details of the participants did not show any correlations with the other metrics using Pearson's correlation coefficient.

## 5.2 Query formulation

The average session length is shown in table 1. The session length was calculated as the time of the last query minus the time of the first query as done in [7]. This has as a result that the session length can be less than the specified minimum session length in the Method section of 4 minutes. The dual task sessions have a higher average session length than the single task session. The significance of this difference and the other differences in this section were each determined using a T-test. This difference turned out to be not significant (p = 0.42). The same holds for the difference between the two dual task sessions themselves. While the low distraction session has a higher average session length than the high distraction session, the difference is not significant (p = 0.70).

The average number of queries are also shown in table 1. The average number of queries from the dual task sessions are lower than the single task session, however, this difference is insignificant (p = 0.32). The difference between the low distraction and high distraction is also insignificant (p = 0.62). The correlation between the number of queries and the session length was also calculated to investigate whether a longer session length results in more queries as one would expect. Pearson's R shows a somewhat positive correlation between the two of R = 0.60 with p = 0.02.

The last characteristic of the query formulation is the query length (QL) shown in table 1. The average QL of the dual task sessions is less than the single task session and this difference is significant (p = 0.04). This would imply that more distraction results in shorter queries. However, queries from the high distraction session are longer than queries from the low distraction session, however, this difference is not significant (p = 0.07). The average QL per user was also calculated to test correlations. The QL per user is higher than the normal QL in all 3 sessions. The QL per user of the dual task sessions is less than the single task session, though this difference is not significant (p = 0.34). The difference between the low distraction and high distraction session is neither significant (p = 0.59). Using Pearson's R we tested for correlation with the session length and number of queries, but they were not strongly correlated.

### 5.3 Knowledge gain

The knowledge gain is measured with the knowledge gain described in [7] and is shown in table 1. The knowledge gain of the dual task sessions is lower than the singe task session, which would imply that the presence of a second task lowers the knowledge gain. The knowledge gain of the high distraction session is lower than the low distraction session, which would imply that the intensity of the second task negatively influences the knowledge gain. Since the topic can also influence the knowledge gain, we used a two way ANOVA to test these two hypotheses. The two way ANOVA showed no evidence for the first hypothesis (p = 0.80). However, there was strong evidence for the second hypothesis (p = 0.016), namely that the intensity of the second task negatively influences the knowledge gain.

As for the topics, the knowledge gain is shown in table 3. The knowledge gain can be seen to vary a lot between topics. The two way ANOVA showed very strong evidence (p = 0.00003) that the topic influenced the knowledge gain. This is different from the results in [7], where the difference in knowledge gain between the three topics is small.

### 5.4 Evaluation

In the evaluation part of the session, the participants were asked 3 questions, which they rated on a scale of 1 - 5, with 1 being the worst. The results of the evaluation can be seen in table 2. The first question revolved around the user satisfaction and was on average negative to neutral. The user satisfaction of the dual task sessions was lower than the single task session and the high distraction session user satisfaction was lower than the low distraction session. This would imply that the user satisfaction decreases because of more distraction, however, both differences were insignificant (p = 0.42 and p = 0.53 respectively). We also used Pearson's R to investigate correlations with any of the previous metrics, but no correlations were found.

The second evaluation question discussed the amount of effort necessary for the task. On average this took quite some effort (4-5). The dual task session has a lower effort on average than the single task session, but this difference is insignificant (p = 0.50). The difference between the low distraction and high distraction session is insignificant as well (p = 0.25). This question also showed no correlation with the previous metrics.

The last evaluation question discussed the quality of the speech recognition. On average the participants rated the quality negative (2-2.4). The difference between the dual task sessions and the single task session, and the difference between the low distraction and the high distraction session were insignificant (p = 0.68 and p = 0.43 respectively). This question showed no correlation with the previous metrics. Also, the three evaluation questions showed no correlation between themselves.

## 6 Responsible Research

High quality research has to be responsible. In this section, two types of responsible research will be discussed, namely ethically responsible and epistemically responsible. After that some tensions between the two types will be mentioned.

### 6.1 Ethically Responsible

For research to be ethically responsible, moral rules are applied to the "collection, analysis, reporting, and publication of information about research subjects" as stated on Encyclopedia [6]. The ethics checklist for human research by the TU Delft was filled in to verify that the subjects would not be harmed during the study. All checklist items, such as "Is pain or more than mild discomfort likely to result from the study?" were answered with no. Thus this study adheres to the TU Delft norms for human research.

Gillen [8] introduced four norms for ethical research: respect for autonomy, non-maleficence, justice and beneficence. Respect for autonomy is adhered to since the participants can choose whether to participate or not after reading the description of the study and they can stop the experiment at any moment in time if they want to. The second norm, non-maleficence, means avoiding harm for the participants. This norm was verified with the TU Delft checklist. This checklist also verifies the third norm, justice, since all laws should be complied to. The fourth norm, beneficence, means balancing the cost of the research to the benefits. This study has relatively little cost for the participants while the outcomes can be helpful, and therefore adheres to this norm.

The most applicable ethics norm to this study is privacy. For this study, no identifiable information about the participants is stored. The only personal information stored is relevant to this study, such as their gender and age. The information and queries are stored together with their Prolific ID, which is necessary to link information from different sessions together. The voice recordings during the sessions are never stored and will not be made available to the public.

### 6.2 Epistemically Responsible

Epistemically responsible research is reliable, reproducible and legitimate. This study was set up while complying to the guidelines of the Netherlands Code of Conduct for Research

Integrity [10]. Merton [13] introduced four norms for epistemically responsible research: universalism, communality, disinterestedness and organized skepticism.

The first norm, universalism, implies that the claims should be hold to pre-established criteria. This study followed the scientific method, where applicable, and thus uses pre-established criteria. The second norm, communality, revolves around open communication about the results and findings to contribute to scientific progress. All the relevant results have been included in this paper, so this norm is complied to. We had no interest or benefits in the outcome of the study and thus were disinterested in the findings of the study which complies to the third norm. Organized skepticism means that the work should be reproducible in order for other researchers to verify the claims made. The methodology used for this study has been thoroughly explained which allows for the reproduction of the conducted study. The software used (Macaw, DrQA, Google Speech API) are all available only.

### 6.3 Tensions

There is a tension between an ethical norm and an epistemical norm. According to the epistemical norms, all the results should be shared to make the claim verifiable by others. However, sharing the voice recordings would invade the privacy of the participants. Because of that we decided to not store the voice recordings.

## 7 Discussion

### 7.1 Main findings

The aspects of query formulation include session length, number of queries and query length. We found that the session length and the number of queries were not significantly influenced by dual tasking. The query length was significantly influenced by dual tasking in the sense that the average query length was shorter in a dual task session. However, the intensity of the dual task session did not influence the query length significantly. The session length positively correlates with the number of queries in a session, which is consistent with the findings of [7]. The query length did not correlate with either of the other two.

We found that the high distraction session had a significantly lower knowledge gain than the low distraction session. However, neither of the two sessions had a significantly different knowledge gain than the single task session. This might imply that a decrease in knowledge gain only happens after a certain intensity of the dual task session and that dual task sessions with a lower intensity level do not influence the knowledge gain. The three different topics had a significantly different knowledge gain. This likely influenced the results.

We found that the three evaluation questions did not correlate with each other, which implies that the questions each touched upon different areas. The three questions also did not correlate with any of the other metrics. The sessions had no significant influences on the answers to the three questions. On average the user satisfaction and speech recognition quality were rated poorly, while the effort necessary to complete the tasks was quite high.

### 7.2 Limitations

This study had several limitations. One of these is the topics chosen. Initially the topics were chosen from [7] especially because of the comparable knowledge gain. However, the results show that the topics still differed significantly. This can be either because of the nature of the topic or because of the specific questions that were paired with the topic.

Since the server for the study ran on a personal computer, the website did not have a continuous uptime. This limited the timeframe in which users from Prolific could participate in the study. Because of this, the sample size was quite small which did not allow for a hard substantiation for the findings.

A major limitation was the quality of the answers returned by Macaw. Only queries of a certain structure were correctly processed, which prevented users from asking queries exactly how they would like. The importance for fact-check queries could also be seen from the many queries with a true/false structure. Gathering information was one of the main objectives for the participants, which was prevented every time Macaw could not find an answer to the query. This directly influenced their answers to the post-knowledge tests and therefore their knowledge gain.

Another limitation was the quality of the speech recognition, which was rated quite poorly by the participants. This also prevented participants from personalizing their queries. Even though the relevant evaluation question did not correlate with any of the other metrics, speech recognition is still a central part in this user study and could have influenced the results.

## 8 Conclusions and Future Work

Our study was designed to determine the effect of dual tasking on the usage of a voice assistant. We conducted a user study to investigate the influences on the query formulation, knowledge gain and user experience. We tested this with a single task session, a dual task session with low distraction and a dual task session with high distraction. We found a significant effect of the dual task sessions on the query length. This implies that the presence of dual tasking results in a shorter average query length. We also found a significant effect of the intensity of the dual task session on the knowledge gain, which implies that a higher intensity can decrease the knowledge gain. However, we did not find a significant effect of the intensity on the query length. We did not find evidence that the presence of dual tasking influences the knowledge gain. Likewise, we did not find a significant effect on any of the other metrics.

A future study could further investigate the usage of a voice assistant by repeating the study in an improved environment. More participants, better answers and speech recognition could result in more significant results. More information on topics, that have almost no influence on the knowledge gain and other metrics, would help to establish a greater degree of accuracy on this matter. This research can form a basis for further development or refinement on voice assistants and conversational search.

## References

[1] ANGUERA, J. A., BOCCANFUSO, J., RINTOUL, J. L., AL-HASHIMI, O., FARAJI, F., JANOWICH, J., KONG, E., LARRABURO, Y., ROLLE, C., AND JOHNSTON, E. Video game training enhances cognitive control in older adults. *Nature 501*, 7465 (2013), 97–101.

[2] BANIQUED, P. L., LEE, H., VOSS, M. W., BASAK, C., COSMAN, J. D., DESOUZA, S., SEVERSON, J., SALTHOUSE, T. A., AND KRAMER, A. F. Selling points: What cognitive abilities are tapped by casual video games? *Acta psychologica 142*, 1 (2013), 74–86.

[3] BARON, A., AND GREEN, P. Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review.

[4] BIGDELI, S., AND KAUFMAN, D. Digital games in health professions education: Advantages, disadvantages, and game engagement factors. *Medical journal of the Islamic Republic of Iran 31* (2017), 117.

[5] DZUBAK, C. M. Multitasking: The good, the bad, and the unknown. *The Journal of the Association for the Tutoring Profession 1*, 2 (2008), 1–12.

[6] ENCYCLOPEDIA. https://www.encyclopedia.com/social-sciences/dictionaries-thesauruses-pictures-and-press-releases/research-ethics.

[7] GADIRAJU, U., YU, R., DIETZE, S., AND HOLTZ, P. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (2018), pp. 2–11.

[8] GILLON, R. Medical ethics: four principles plus attention to scope. *Bmj 309*, 6948 (1994), 184.

[9] KISELEVA, J., WILLIAMS, K., JIANG, J., HASSAN AWADALLAH, A., CROOK, A. C., ZITOUNI, I., AND ANASTASAKOS, T. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (New York, NY, USA, 2016), CHIIR '16, Association for Computing Machinery, p. 121–130.

[10] ”KNAW”. Nederlandse gedragscode wetenschappelijke integriteit.

[11] KUN, A., PAEK, T., AND MEDENICA, Z. The effect of speech interface accuracy on driving performance. In *Eighth Annual Conference of the International Speech Communication Association* (2007).

[12] MCEVOY, S. P., STEVENSON, M. R., AND WOODWARD, M. The contribution of passengers versus mobile phone use to motor vehicle crashes resulting in hospital attendance by the driver. *Accident Analysis & Prevention 39*, 6 (2007), 1170–1176.

[13] MERTON, R. K. *The sociology of science: Theoretical and empirical investigations.* University of Chicago press, 1973.

[14] NPD EEDAR. Gamer segmentation. Tech. rep., 2019.

[15] PASHLER, H. Dual-task interference in simple tasks: data and theory. *Psychological bulletin 116*, 2 (1994), 220.

[16] PETROCK, V. Us voice assistant users 2019. Tech. rep., eMarketer, July 2019.

[17] REDELMEIER, D. A., AND TIBSHIRANI, R. J. Association between cellular-telephone calls and motor vehicle collisions. *New England Journal of Medicine 336*, 7 (1997), 453–458.

[18] TAKAYAMA, L., AND NASS, C. Driver safety and information from afar: An experimental driving simulator study of wireless vs. in-car information services. *International Journal of Human-Computer Studies 66*, 3 (2008), 173–184.

[19] TOMBU, M., AND JOLICŒUR, P. A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance 29*, 1 (2003), 3.

[20] WINKLER, M., BUVAT, J., AGGARWAL, G., MEHL, R., PUTTUR, R. K., AND SHAH, H. Voice on the go. Tech. rep., Capgemini Research Institute, 2019.

[21] WORLD HEALTH ORGANIZATION. Mobile phone use: a growing problem of driver distraction. Tech. rep., January 2011.

[22] YOGEV-SELIGMANN, G., HAUSDORFF, J. M., AND GILADI, N. The role of executive function and attention in gait. *Movement disorders: official journal of the Movement Disorder Society 23*, 3 (2008), 329–342.

[23] ZAMANI, H., AND CRASWELL, N. Macaw: An extensible conversational information seeking platform, 2019.