# Methods in AI Research: Research Proposal
# Group 19

Aron Noordhoek
a.j.noordhoek@students.uu.nl

Teun Buwalda
t.c.buwalda@students.uu.nl

Maria Mouratidi
m.mouratidi@students.uu.nl

Alimohamed Jaffer
a.a.jaffer@students.uu.nl

Jichen Li
j.li20@students.uu.nl

November 9, 2023

## 1 Introduction

Human-computer interaction (HCI) often requires a degree of collaboration between the user and the system. When it comes to collaborative decision-making, the agency allocation between the user and the system can vary. This variation can be dependent on the system's abilities or can be predetermined by the system's developer. According to previous research, (Heer, 2019), an agent should never take over decision-making completely, because both the user and the system can offer different levels of expertise (Ren, Chen, & Qiu, 2023), such as in intuitive thinking tasks, where humans tend to perform better (Jarrahi, 2018).

While relying on the user's agency is desirable in high-stake situations where different levels of expertise may be needed, the opposite approach may be more productive for lower-importance daily tasks. For example, tasks like setting an alarm or getting a restaurant recommendation, should not occupy the user's entire cognitive resources in order to reserve resources for more important tasks. Indeed, research suggests that reducing the cognitive load of the human agent can improve task performance (de Melo, Kim, Norouzi, Bruder, & Welch, 2020) as well as trust in the virtual agent (Ahmad, Bernotat, Lohan, & Eyssel, 2019).

In support of the idea that assigning a lower degree of agency to the user may be a better design choice when creating an HCI system, the "paradox of choice" can be considered, as introduced by Schwartz (2005). The paradox of choice describes the phenomenon by which giving users the responsibility (or in this case, agency) to choose from a large array of options can cause more anxiety, and lead to poorer decision-making.

Moreover, as shown by Yan, Chang, Chou, and Tang (2015), the perception of greater choice variety can lead to better user satisfaction, only until it reaches a certain threshold. When the perceived variety becomes too large, user satisfaction drops. This evidence overall suggests that choice anxiety, as a result of excessive agency allocated to the user, could lead to poor task performance and lower satisfaction.

This study aims to investigate whether taking away some agency from the user in an everyday task like choosing a restaurant and allocating it to the system instead, can indeed bring better satisfaction and less choice overwhelmedness. Thus, the research question for this study is:

> *How does manipulating choice variety influence user satisfaction and overwhelmedness in decision-making scenarios with a digital assistant?*

In line with the previous research described above, we hypothesize that higher choice variety will lead to overwhelmedness and lower satisfaction, both for the interaction with the system, and the final decision of the restaurant. The context in which this study tests this hypothesis is a restaurant recommendation system, where the user has to get a recommendation given different preference scenarios.

# 2 Methodology

## 2.1 Participants

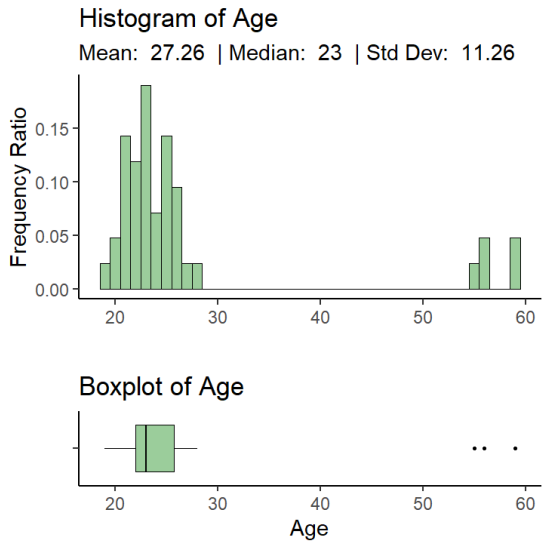
A total of 42 participants were recruited for the experiment. The demographics corresponding to this participant group are illustrated through a collection of plots in Figure 1. We report some generic statistics such as age and gender, as well as two task-specific descriptive statistics; level of familiarity with AI chatbots and level of English. The mean age of the participants was 27.26 with a standard deviation of 11.26, as seen in Figure 1a. 55% of the participants are male, 36% female, and the rest either did not specify or identify as 'other' (1b). Figure 1c shows that the average participant has a moderate familiarity with artificial dialogue systems. Lastly, the reported level of English proficiency is shown in Figure 1d. Most participants reported being fairly competent with the English language, so we can assume that language barriers did not affect the experiment. All participants consented to their data being processed anonymously for research purposes, and the data collection adhered to the Ethics and Privacy assessment guidelines provided by the Research Institute of Information and Computing Sciences, at Utrecht University.

## 2.2 Experimental Design

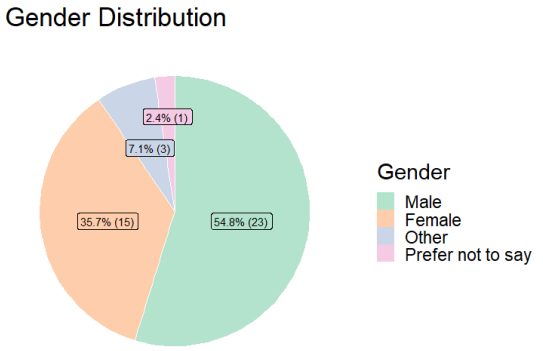
The experiment has a within-subject design and includes three sub-tasks, each randomly performed in one of the two conditions; control condition (A) and experimental condition (B). The condition sequences 'AAA' and 'BBB' were not performed as they are not valid for this study design. The participants need to interact with the system and eventually get a restaurant recommendation based on the given prompt for each sub-task. Condition A is a version of the recommendation system where after the user expresses their cuisine preference, the system prompts two random sub-categories from this cuisine. For example, if the user asks for an Asian restaurant in the cheap price range, the system will prompt the user to pick between Chinese and Korean. Condition B on the other hand, is a version of the recommendation system where after the user expresses their preference, the system will present 5 cuisine sub-categories. For example, when the user asks for an Asian restaurant, the system will ask the user to choose between Chinese, Japanese, Korean, Thai, or Indonesian food. The experimental condition (B) confronts the user with a wider array of options, while the control condition takes the liberty to eliminate some options to facilitate the user. With this design, we aim to test the effect of the number of options on the user satisfaction of the recommendation process, as well as the final recommendation and feeling of overwhelmedness.

## 2.3 Materials and Measures

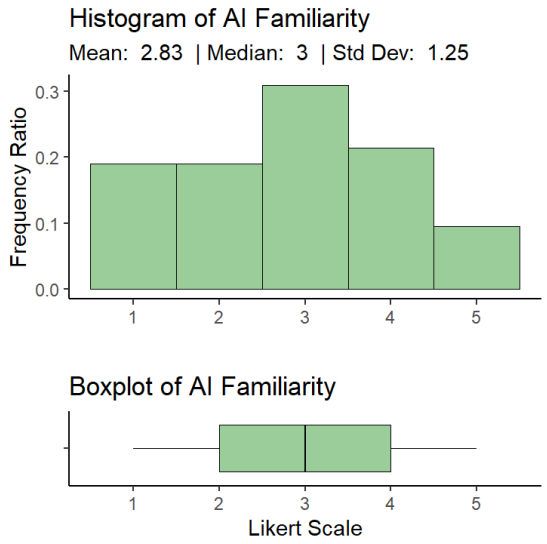
The experiment physically takes place on the experimenters' computers as the recommendation system is run in a local environment. User satisfaction with the recommendation process, final restaurant choice, and feeling of overwhelmedness are measured by means of a post-experimental survey, consisting of a demographic and experimental section. The demographic section of the questionnaire aims to model the demographic of the participants, as discussed in Section 2.1. The questions within the experimental section of the survey are organized on a Likert scale and aim to assess the participant's beliefs and perception of the recommendation system. The questionnaire was designed and distributed through Qualtrics (*Qualtrics*, 2023). The pre-processing of the data and the statistical analysis of the results was performed in R (RStudio Team, 2022). To measure the statistical difference of the conditions, we iteratively selected two random conditions for each participant, one in A and one in B, using bootstrap resampling of 500 permutations with replacement. For each permutation, we performed a paired t-test, with significance level of alpha = 0.05. Finally, we considered the mean and standard deviation of t-values and p-values across bootstraps, as well as the confidence intervals for Cohen's $d$ effect sizes. The effect sizes were compared based on the benchmarks suggested by Cohen (1988). This approach



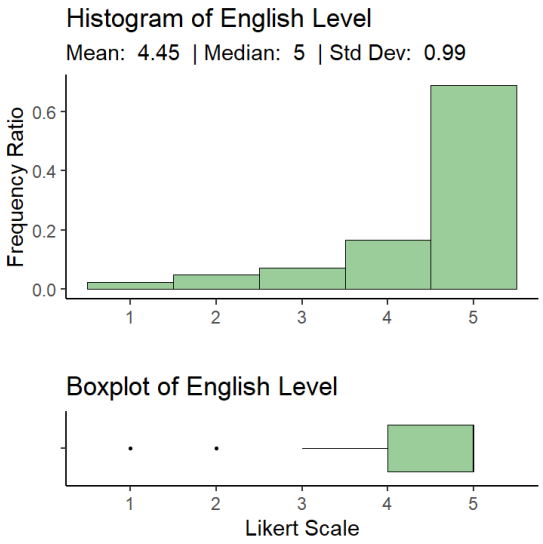

(a) Histogram and box-plot of self-reported participant age.




(b) Pie chart with raw and percentage values of self-reported gender.




(c) Histogram and box-plot of Likert scale responses for familiarity with AI chatbots. 1 represents unfamiliarity while 5 represents high familiarity with AI systems




(d) Histogram and box-plot of Likert scale responses for level of English ability. 1 represents limited English ability while 5 represents English proficiency

Figure 1: Descriptive plots of the sample's demographics

combines statistical with practical significance measures, and offers all necessary information regarding null-hypothesis testing; It includes the importance of the observed effect as well as the generalizability of the results through resampling, as suggested by Banjanovic and Osborne (2016). The analytical values are presented in Section 3.

### 2.4 Procedure

At the start of the experiment, the participants are given a story that places them within a context, e.g. they live in the south part of town, don't have much money, but would like to take their children out to eat at an Asian restaurant. They are then given a laptop with the recommendation system opened in the correct condition. Their task is to find a restaurant that conforms to their preferences and the context, and chat with the dialogue system for as long as they think is necessary. This process is repeated three times, in which both conditions (A, with 2 cuisine categories; and B, with 5 cateogries) occurs at least once. An example system output in condition A for the category *mediterranean* would be as follows:

```
System | We have greek, moroccan.  What would you like?
```

Whereas in condition B, this output could be:

```
System | We have italian, moroccan, greek, spanish, portuguese.
What would you like?
```

After the participants have finished their dialogue with the system, they are asked to fill out a short digital questionnaire about their experience, as described in section 2.3.

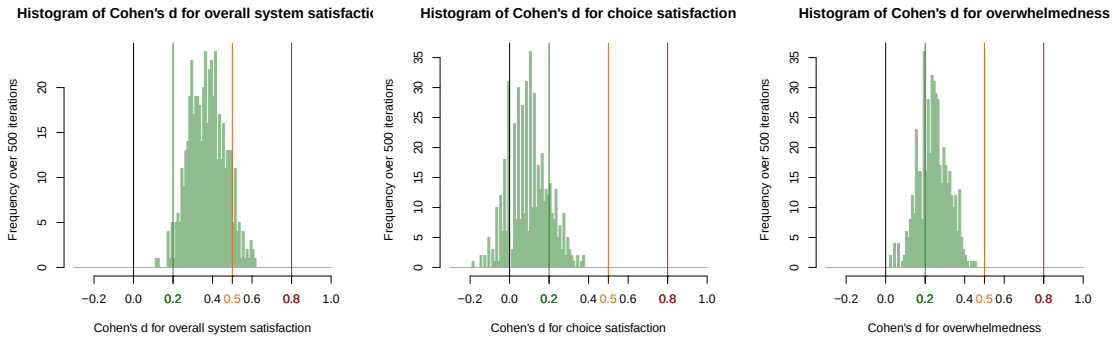
## 3 Results

The difference in system conditions did not have a significant effect on satisfaction with the recommendation process (average $t = -2.06, df = 43, p = 0.07$). However, 97% of the t-test permutations show an at least small effect size (Figure 2a), where respondents were more satisfied when given more options. Satisfaction with the final restaurant choice was also not significantly different between conditions (average $t = -0.63, df = 43, p = 0.52$), and the effect size across 83% of the permutations was negligible (Figure 2b). Finally, feeling of overwhelmedness did not differ significantly between conditions (average $t = -1.38, df = 43, p = 0.22$), and 68% of the t-test permutations showed a small effect size (Figure 2c), where more options made participants more overwhelmed. These results do not validate our hypothesis, which states that a larger variety of choices during the recommendation will undermine satisfaction and induce a feeling of overwhelmedness. In the next section, we discuss some possible reasons for the results and discuss improvements.

## 4 Discussion

### 4.1 Evaluation of Results

In this experiment, we empirically tested whether manipulating choice variety influences user satisfaction and overwhelmedness, by letting participants interact with two kinds of systems. One system (condition A) presented 2 sub-categories for each cuisine, whereas the other system (condition B) presented 5 sub-categories. Even though the given results did not validate our initial hypothesis, it is important to address the issues that may have led to this outcome, other than the possibility that our hypothesis is incorrect. This section discusses the implications for each of our three measurements; namely recommendation satisfaction, restaurant choice satisfaction, and overwhelmedness.




(a) Overall system satisfaction (b) Choice satisfaction (c) Overwhelmedness

Figure 2: Distribution of Cohen's d for all measures. The number of choices had a small to medium effect on overall system satisfaction, no significant effect on choice satisfaction, and a small effect on overwhelmedness.

#### 4.1.1 Overall system satisfaction

The number of choices did not significantly influence the overall satisfaction with the system. This result may have to do with the fact that the dialog system already gives rise to a complicated interaction as is. After all, the conditions only manipulated a total of 1-2 sentences in the conversation, and other aspects of the interaction may have overshadowed the overall impression of the system.

At the same time, analysis using Cohen's $d$ showed a small to medium positive effect of the increased number of choices on overall satisfaction (Figure 2a. This result could be explained by the idea that picky users were looking for a more specific cuisine and their needs were satisfied given a wider array of options.

#### 4.1.2 Restaurant satisfaction

Participants were not more satisfied with their ultimate choice when given a higher or lower number of choices. We speculate that this may be because the difference between 2 and 5 options may be not large enough to capture a noticeable effect in the final choice satisfaction. This relates to the choice variety threshold, mentioned in Section 1, where satisfaction only drops when too much variety is present. For example, in another experiment that showed the paradox of choice in jam marketing (Iyengar & Lepper, 2000), the few-choices condition included 6 jam options, while the many-choices condition involved 24 jams. Another possible reason for the negligible effect of choice variety on the ultimate choice satisfaction could be the simulating nature of the experiment. Because the participants only received a hypothetical recommendation, they may not have been able to assess their satisfaction realistically. In other words, there is no evidence that the participants expressed their satisfaction the way they would in a real setting. This speculation has consequences for the ecological validity of this study.

#### 4.1.3 Overwhelmedness

Participants showed no difference in overwhelmedness between conditions. This result may be attributed to two different factors; either the experiment design did not successfully manipulate the user's overwhelmedness, or the manipulation was overshadowed by other overwhelming events during the experiment.

The manipulation of overwhelmedness may have failed because of the small difference in the choice variety conditions, as discussed above. Thus, the difference may not have been enough to cause choice anxiety as suggested by the paradox of choice.

The reason why overwhelmedness may have occurred randomly due to other events is the random length of conversations in each condition. In either condition, participants may have felt unhappy about the recommendation and attempted to get more options, a course of action resulting in longer conversations that our recommendation system did not always handle well. Additionally, the number of restaurants that matched each prompt may have had an impact on the length of conversations. Since the list of restaurants was limited, some prompts could not be satisfied when the user made a certain choice (e.g. there were no Korean restaurants in the south area). Having initially fewer options made it more likely for the user to reach a dead end, meaning they were more likely to have to start over the conversation. This process may have been interpreted as overwhelming by the participant, although not the type the experiment aimed to capture.

### 4.2 Future Work

In the future, we believe that certain changes in the study design could improve the validity of this experiment. For example, using a larger difference between the number of suggestions between conditions, as conducted in previous research, could show an effect more aligned with our hypothesis. Moreover, using a more complete restaurant dataset could account for additional overwhelmedness naturally caused by a larger number of turns between conditions. Last, but not least, fine-tuning the system to more efficiently handle all interactions (such as when the user wants to modify their choices and change the recommendation development) may account for additional overwhelmedness that may act as a confound.

## References


Ahmad, M. I., Bernotat, J., Lohan, K., & Eyssel, F. (2019). Trust and cognitive load during human-robot interaction.

Banjanovic, E. S., & Osborne, J. W. (2016). Confidence intervals for effect sizes: Applying bootstrap resampling. Retrieved from `https://scholarworks.umass.edu/pare/vol21/iss1/5/` doi: 10.7275/DZ3R-8N08

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *New York, NY: Routledge Academic*.

de Melo, C. M., Kim, K., Norouzi, N., Bruder, G., & Welch, G. (2020, November). Reducing cognitive load and improving warfighter problem solving with intelligent virtual assistants. , *11*, 554706.

Heer, J. (2019, February). Agency plus automation: Designing artificial intelligence into interactive systems. *Proc. Natl. Acad. Sci. U. S. A.*, *116*(6), 1844–1850.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. doi: 10.1037/0022-3514.79.6.995

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business horizons*, *61*(4), 577–586.

*Qualtrics.* (2023). `https://www.qualtrics.com/`. (Accessed: November 9, 2023)

Ren, M., Chen, N., & Qiu, H. (2023, JUL). Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive intelligence. *INTERNATIONAL JOURNAL OF SOCIAL ROBOTICS*, *15*(7), 1101-1114. doi: 10.1007/s12369-023-01020-1

RStudio Team. (2022). Rstudio: Integrated development for r [Computer software manual]. Boston, MA. Retrieved from `http://www.rstudio.com`

Schwartz, B. (2005). The Paradox of Choice: Why More Is Less. *Vikalpa*, *42*, 265–267.

Yan, H., Chang, E.-C., Chou, T.-J., & Tang, X. (2015, March). The over-categorization effect: How the number of categorizations influences shoppers' perceptions of variety and satisfaction. *J. Bus. Res.*, *68*(3), 631–638.