Effects of system response delay on perceived humanness and usability

Yu-Lan Dibbits, Harmjan Lever, Alexia Ntantouri, Kaja Philippens, and Pepijn Stoop

Utrecht University

Methods in AI Research

Leendert van Maanen

Oct. 28, 2024

Author Note

Student numbers: 6679730, 6246974, 2958481, 6790526, and 9291164

Effects of system response delay on perceived humanness and usability

Dialogue systems are becoming more and more common. As their usage has become widespread, the need to better understand interactions between humans and these systems arises. A key difference between this technology and other information systems is that one of its goals is facilitating natural, human-like conversations. This humanness is an additional goal separate from the goal of usability, which implies the possibility of a trade-off between the two. One attempt to introduce more humanness to a dialogue system is introducing a delay in the responses it gives. Instead of receiving a near-instant reply, the user has to wait a certain amount of time, often less than a second, which could make the conversation feel more natural.

In existing literature, such a delay before the system's response does appear to increase humanness in some cases and decrease usability in others. A delay tends to have a positive effect on perceived 'humanness', especially when the user's motivations to use the system are related to social and relational factors (Brandtzaeg & Følstad, 2017). In such cases, a small delay can have a positive effect, making the system feel more human (Holtgraves & Han, 2007) and natural (Appel et al., 2012). Furthermore, small response delays can cause an increase in user satisfaction (Gnewuch et al., 2018). On the other hand, when fast response times are expected (Gnewuch et al., 2018) or productivity is a priority (Brandtzaeg & Følstad, 2017), the response delay has a negative effect (Gnewuch et al., 2022) and causes the system to be considered less likeable (Schanke et al., 2021).

We set out to test whether the presence or absence of a response delay significantly changes its perceived humanness and usability when testing for the same dialogue system. In other words, we want to show whether a delay can elicit a trade-off between humanness and usability when users have similar goals when interacting with one system. For this study, we used a dialogue agent that recommends restaurants based on preferences extracted during conversations with the user.

We created a rule-based model that can identify dialogue acts, e.g. 'bye' when the user's input contains 'goodbye' or 'negate' when the input contains 'no'. Next, we created a dialogue

system that welcomes the user with the following message: “Hello, welcome to the restaurant recommendations dialogue system! You can ask for restaurants by area, price range, or food type.” The user can provide the system with their preferences, and after a short conversation, receive restaurant recommendations that fulfil their needs. The system consists of eleven states, ranging from the ‘WELCOME_STATE’ to the ‘END_STATE’.

Dialogue acts determine whether information can be extracted or which state to proceed to next. First, the system tries to identify the most important preferences, which include the area, price range, and type of cuisine. It filters the list of restaurants accordingly and then asks if the user has any additional preferences, such as ‘romantic’ or ‘children allowed’. Once this is handled, the system recommends a restaurant. It can provide details such as the address and phone number. If there are no more recommendations, the system asks if the user needs anything else, allowing them to either repeat the process or end the conversation.

Based on existing research, we hypothesize that a small delay would increase the perceived humanness of our restaurant recommendation system but might decrease its perceived usability.

Methods

Participants

For this study, we interviewed 24 participants, all between the ages of 20 and 25. The gender distribution shows a higher proportion of female participants (60.9%) compared to male participants (34.8%), with a small percentage identifying as "other." All participants, except for two (a "Market Researcher" and a "Service Worker"), were students. Most students were enrolled in the MSc AI program at Utrecht University, followed by students in the MSc Bioinformatics program. These demographics were collected to ensure the homogeneity of our participants. The aim was to ensure that our participants would be between 18 and 30 years old as we did not want age to have a possible effect on the outcome. Furthermore, we asked about gender and occupation to rule out any confounding factors.

Materials

To conduct the experiment, a laptop was used to host both the dialogue system and the questionnaires that participants were required to complete. More specifically, Visual Studio Code was utilized to run the dialogues, which were executed in the system's terminal.

The system was initialized with a fixed delay of one (1) second for the first part of the experiment (see subsection Design). A fixed delay was chosen instead of a variable one, as we aimed specifically to test the effect of the presence or absence of a delay on perceived humanness and usability, rather than the effect of variability in delay on these perceptions. The length of the delay was set to 1 second. Previous comparable experiments have used a variety of delays (Holtgraves et al., 2007 : Gnewuch et al., 2018). Delays are already noticeable after a fraction of a second (Miller, 1968). However, we also wanted to mimic human-like responses. Given that the length of the dialogues when using the system varies, as well as the fact that we did not intentionally want the delay to cause a negative effect, we thought that 1 second would be optimal.

Furthermore, the dialogue system was configured with a Levenshtein edit distance of 2, allowing words misspelt by up to two characters to be classified by the system as correct. This configuration was selected to account for typographical and spelling errors, as well as differences

between UK and US English.

Based on existing research, we hypothesized that a small delay would increase the perceived humanness of our restaurant recommendation system but decrease its perceived usability.

Design

The experiment consisted of two parts. The first part included the experiment with a one-second delay, while the second part was conducted without that delay. To counterbalance learning effects, participants completed two different but similar tasks in total, one for each condition. Both the tasks and the conditions were counterbalanced within the group to prevent order effects. In total, there were four different configurations of the experiment. An overview of these configurations is detailed in Table 3 in the Appendix.

Each of the two tasks also consisted of two parts. Task 1.a required participants to find an expensive restaurant in the south of town and to record the address and phone number. Task 1.b required them to find a restaurant serving British food at a moderate price and to note the address and postal code. Task 2.a involved finding a cheap restaurant in the centre of town and recording the address and postal code, while Task 2.b required participants to find a restaurant serving Indian food at a low price and to record the address and postal code. The tasks were designed to be as similar as possible, specifying the price range and the area in one case (Task 1.a and Task 2.a), and the type of food and the price range in another case (Task 1.b and Task 2.b). Additionally, they required participants to obtain either the address and phone number or the address and postal code. Each task consisted of two parts, i.e., two conversations, to provide participants with enough time to become familiar with the system.

Before designing the experimental setup, we completed the Ethics and Privacy Quick Scan of Utrecht University. This Quick Scan revealed privacy risks associated with the personal data collected in the Informed Consent forms (see Section Procedure). We mitigated these risks by restricting access to and sharing of this data solely among members of our research group.

Procedure

Each participant received a short introduction to the experiment, stating that they will be engaging with a dialogue system that is specified in recommending restaurants, after which they have to fill in some questionnaires. They were asked to sign an informed consent and fill in their age, gender and occupation before the experiment started. It was decided beforehand in what order the participant would experience the conditions when interacting with the dialogue system. The system was then prepared by running the script and selecting the correct condition (with (1) or without delay (0)). The instructions for the person conducting the research were to intervene as little as possible with the participant. When intervening, the person experimenting had to refer the participant to the instructions and write down any complications that arose. After having finished the first task, the participant had to fill in the corresponding questionnaires. Then the system was restarted so the participant could complete the second task, but with a different condition. After completing the second task, the participant had to fill in the corresponding questionnaires again. Afterwards, it was possible to inform the participant about the nature of the experiment when shown interest.

Measurements

Measurements were taken during the experiment using two questionnaires. The Godspeed Questionnaire is regarded as a golden standard method in the field of Human-Robot Interaction and Human-Agent Interaction (Bartneck, 2023). It consists of five central relevant concepts to this field: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. The first four concepts relate to the user's opinion and impressions of the dialogue system during the interaction. The last concept refers to the user's perception of the level of danger and the level of comfort during the interaction with the system. The five concepts in the Godspeed Questionnaire are measured using a 5-item scale to indicate to what degree the systems acted human-like or not, with 1 being unhuman and 5 being human-like. (Bartneck et al., 2008).

The System Usability Scale questionnaire uses a ten-item scale to give a global view of subjective assessments of usability from the participant (Brooke, 1995). It has been widely used

in Human-Robot Interaction studies and has proven to be a reliable and robust tool.

The questionnaires were collected electronically through online forms. There were separate questionnaires for the separate conditions (with and without delay), so in total, four forms were created (two for each questionnaire). For this research, the English versions of both questionnaires were chosen to accommodate non-Dutch research participants. Both questionnaires were filled in by participants based on their perceptions and opinions about the interaction.

Results

Descriptive Statistics

In terms of number of inputs per dialogue (Figure 1), for Task 1 (with delay and without delay), participants averaged 9.04 number of inputs (SD = 3.72) for Dialogue 1 and 8.33 (SD = 3.70) for Dialogue 2. For Task 2, the average number of inputs increased to 10.92 (SD = 7.52) for Dialogue 1 and 9.63 (SD = 6.76) for Dialogue 2, suggesting that Task 2 required slightly more engagement, potentially due to increased task complexity or familiarity with the delayed system. The difference in the number of inputs is described further in the Inferential Statistics subsection.

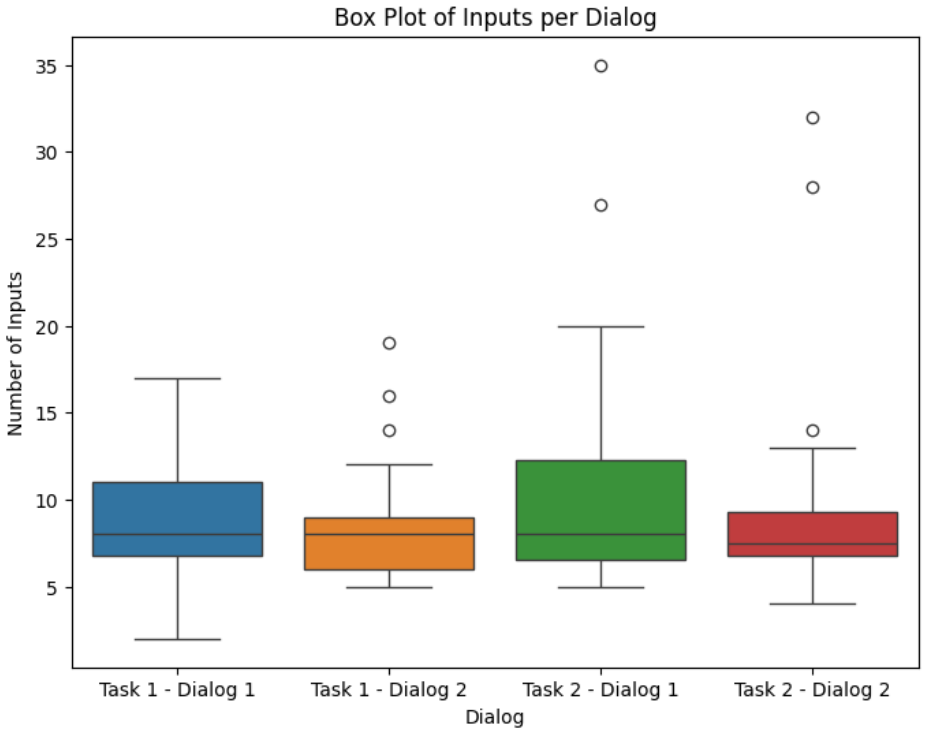

Figure 1

Box plot of the number of inputs per dialogue.

Summing the number of inputs per task, as shown in Figure 2, for Task 1 across both dialogues (with delay and without delay), participants averaged 17.38 number of inputs (SD = 6.24), while for Task 2, the average number of inputs increased to 20.54 (SD = 13.50).

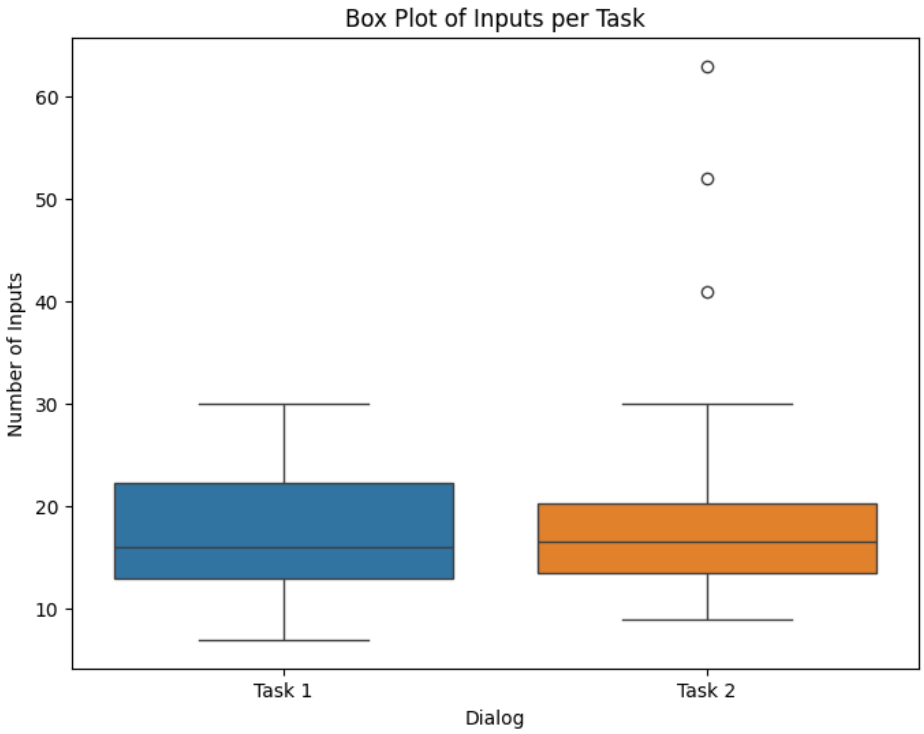

Figure 2

Box plot of number of inputs per task.

Figure 3 shows the number of inputs per participant for each task. We can see that participants 12, 13 and 14 struggled with Task 2 and had to provide way more inputs in comparison to the other participants increasing the average number of inputs for Task 2. This discrepancy could be attributed to their lack of experience using a dialogue system and the fact that Task 2 was the first task they had to do, as seen in Table 3 in the Appendix, that shows the task distribution of tasks among participants.

The results above suggest that Task 2 required slightly more engagement in comparison to Task 1, with an average of 8.04 inputs for Task 1 and 9.96 inputs for Task 2, but the difference was minor and is described further in the Inferential Statistics subsection.

In terms of System Usability Scale (SUS) scores, participants completed the SUS for both conditions (with delay and without delay). Figure 4 shows the violin plot of SUS scores with and without delay.

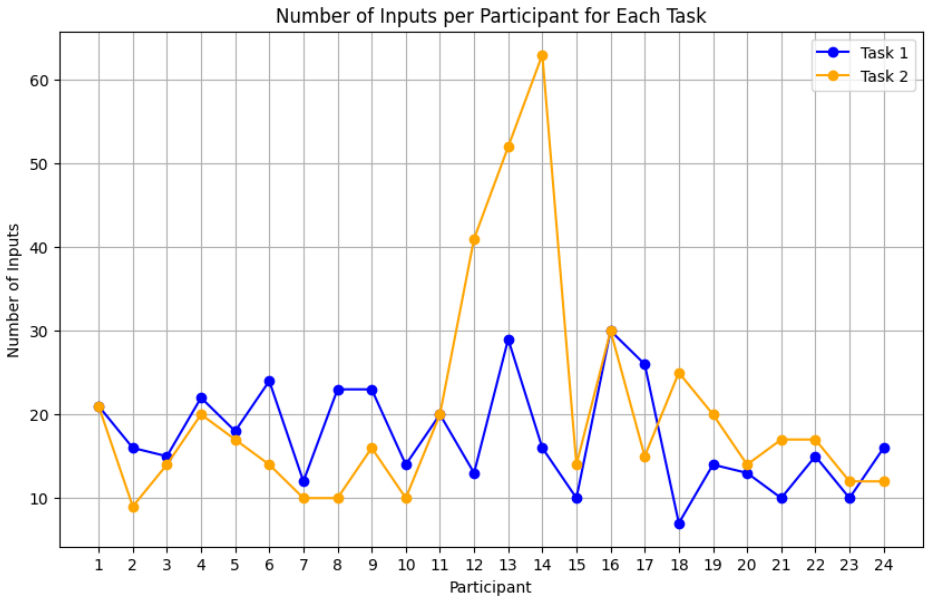

Figure 3

Number of inputs per participant for each task.

The average SUS score was 67.92 (SD = 18.75) without delay and 71.35 (SD = 17.94) with delay. These results show a slight increase in usability scores in the delayed condition, but the difference was minor and is described further in the Inferential Statistics subsection.

The average difference in SUS scores without minus with delay, was -3.44 (SD=13.27), as shown in Figure 5 which is a violin plot of the differences in SUS scores.

In terms of Godspeed Questionnaire Series (GQS) scores, participants completed the GQS for both conditions (with delay and without delay).

Table 1 shows the results of the GQS questionnaire per category. Between conditions, the Animacy and Likeability categories have higher scores for the condition without delay. Likeability has a score of 3.86 without delay and 3.79 with delay. Animacy has a score of 3.17 without delay and 3.13 with delay. On the other hand, the other categories have higher scores for the condition with response delay. Although the difference in scores between the conditions with and without delay is minor as described further in the Inferential Statistics subsection.

Across categories, Likeability has the highest scores for both conditions (3.86 without delay and 3.59 with delay) and Perceived Intelligence has the second highest score for both

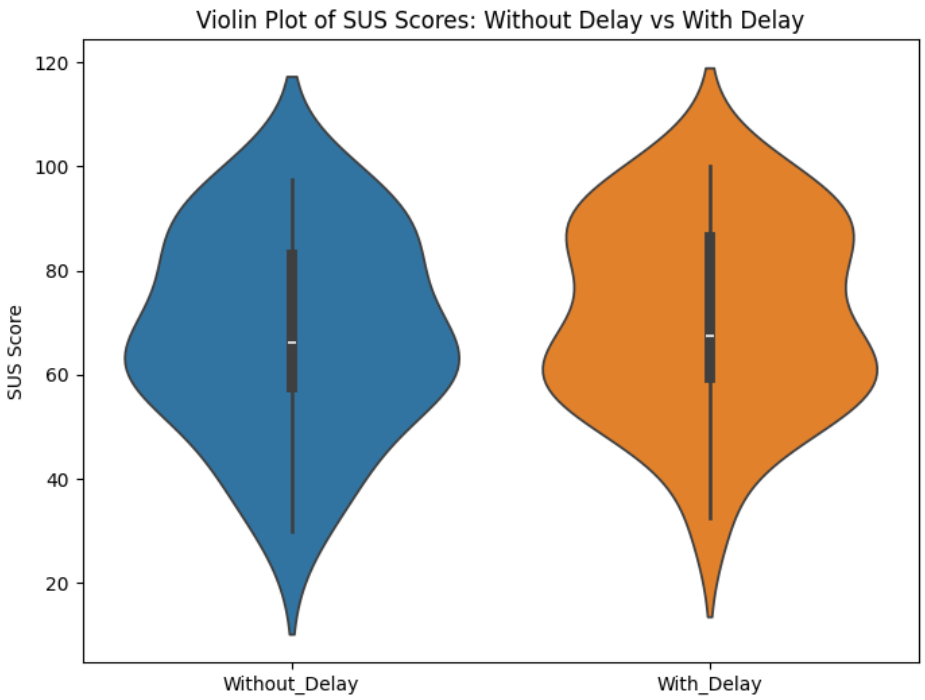

Figure 4

Violin Plot of SUS scores with and without delay.

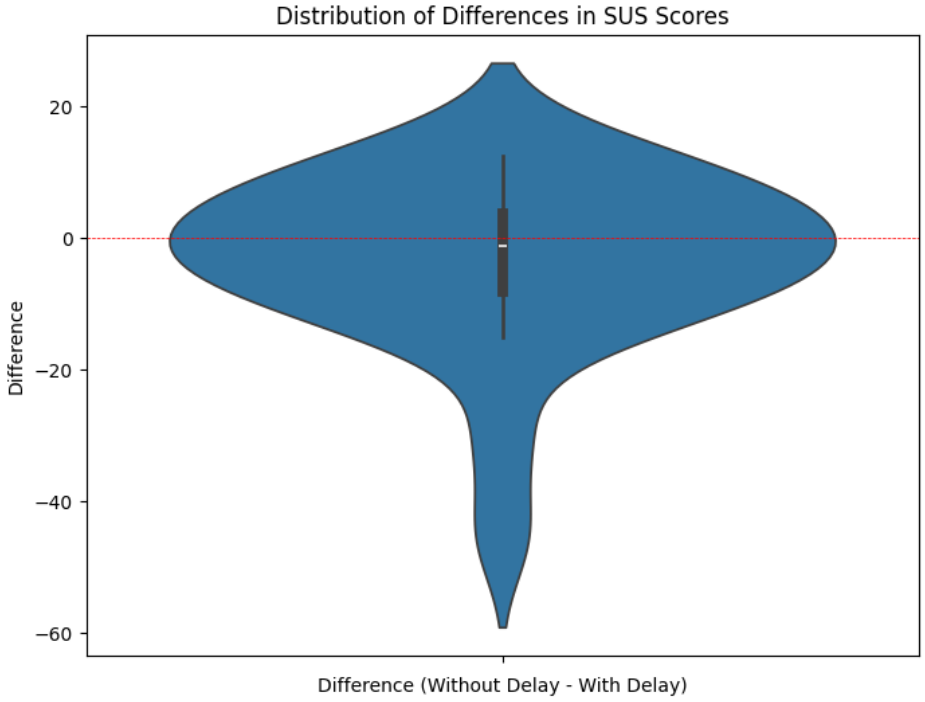

Figure 5

Violin plot of differences in SUS scores without minus with delay.

conditions (3.86 without delay and 3.59 with delay).

Category	Average		Std Dev	
	Without Delay	With Delay	Without Delay	With Delay
Anthropomorphism	2.78	2.88	0.95	0.99
Animacy	3.17	3.13	1.10	1.08
Likeability	3.86	3.79	0.87	0.89
Perceived Intelligence	3.39	3.59	0.96	1.02
Perceived Safety	2.99	3.07	1.33	1.21

Table 1

Godspeed Questionnaire Series Results (With and Without Delay)

Figure 6 shows the violin plot of GQS scores with and without delay. The average GQS score was 3.26 (SD = 0.59) without delay and 3.34 (SD = 0.65) with delay. These results show a slight increase in humanness scores in the delayed condition. However, Figure 7, which contains the violin plot of differences in GQS scores without minus with delay, shows that this difference was minor, more specifically the average difference was -0.08 (SD=0.44). The Inferential Statistics subsection elaborates further on this GQS score difference.

Inferential Statistics

Using a two-tailed t-test, there was no significant difference in the number of inputs between Task 1 and Task 2, $t(23)=1.14, p=0.27$, while the one-tailed t-test was $t(23)=-1.14, p=0.12$. The effect size, Cohen's d, was calculated as $d=-0.30$, indicating a small effect size and minimal practical difference in the number of inputs required between the two tasks. The 95% confidence interval for Task 1 user input counts ranged from [14.88, 19.87], and for Task 2, it ranged from [15.14, 25.94], with considerable overlap in these intervals, further confirming no statistically significant difference between the tasks in terms of input count.

Using a two-tailed t-test for SUS scores, there was no significant difference in usability

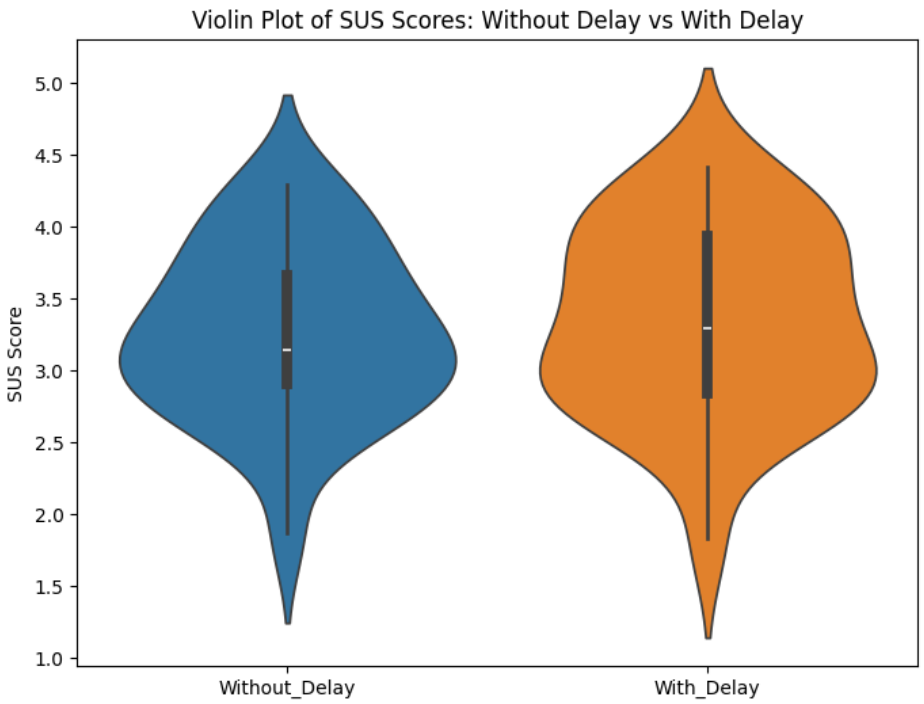

Figure 6

Violin Plot of GQS scores with and without delay.

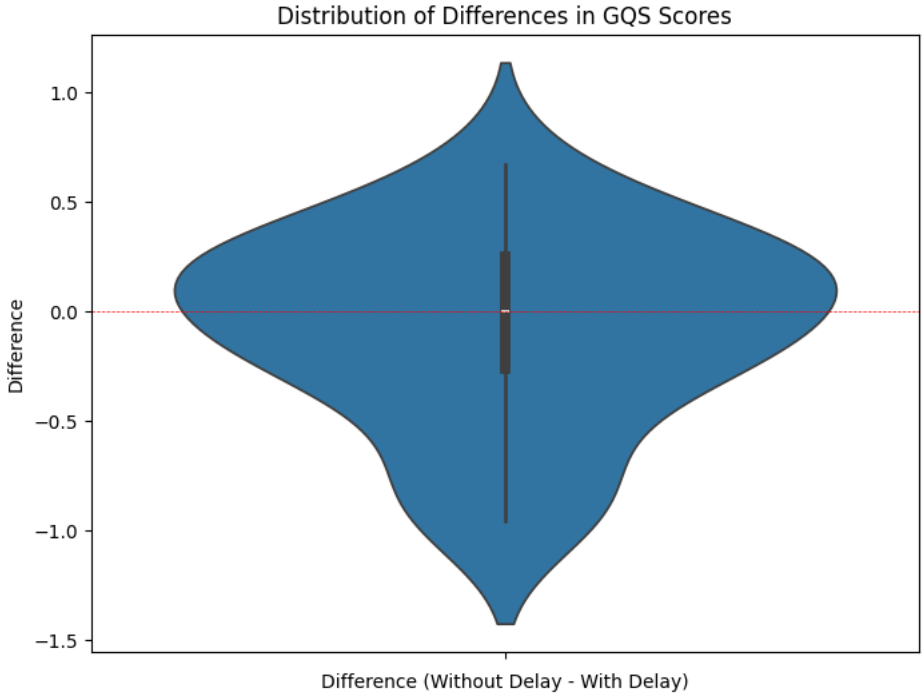

Figure 7

Violin plot of differences in GQS scores without minus with delay.

scores between the conditions with and without delay, t(23)=1.27,p=0.22, while the one-tailed t-test was t(23)=-1.27,p=0.11. The effect size, calculated using Cohen's d, was small (d=-0.19), indicating a negligible impact of delay on SUS scores. The 95% confidence interval for SUS scores ranged from [60.42, 76.46] without delay and [64.18, 78.53] with delay, overlapping considerably, which further confirmed no significant usability impact.

A two-tailed t-test for GQS scores indicated no significant difference in perceived humanness between conditions, t(23)=0.89,p=0.38. The one-tailed t-test was t(23)=-0.89,p=0.19. The effect size was also small, with Cohen's d=-0.13, suggesting a small effect size of delay on humanness perception. The 95% confidence interval ranged from [3.08, 3.60] with delay and [3.03, 3.50] without delay, overlapping considerably, which further confirmed no significant impact on perceived humanness. The results of this study indicate that introducing a response delay in the system did not significantly affect usability or perceived humanness. Specifically, the SUS scores showed only a small, non-significant increase in usability in the delayed condition, as indicated by large t-test p-values, a small effect size (Cohen's d) and overlapping confidence intervals. Similarly, the GQS scores displayed a slight, non-significant increase in perceived humanness with delay, which again was not significant as indicated by the large t-test p-values, a small effect size (Cohen's d) and overlapping confidence intervals.

General Discussion

Research question

The objective of our research was to determine whether the perceived humanness and usability of a dialogue system are influenced by its response time. We hypothesized that a response delay would increase perceived humanness while decreasing usability. To investigate this, we had participants test our restaurant recommendation system with and without a delay and then complete surveys to measure how natural and usable they perceived the system to be.

The findings presented in the Results section suggest that, for the task types and delay durations tested, response delay does not significantly affect users' perceptions of usability or humanness. Therefore, small delays can be introduced or omitted without detracting from the user experience, potentially allowing greater flexibility in system design.

On the other hand, the lack of a significant difference may be partially attributable to the limitations of our research. There are several limitations that we believe are important to convey.

Limitations of experiment

Firstly, the effectiveness of our research may have been affected by imperfections in our recommendation system. One issue was the Levenshtein distance, which we set to 2 to allow for minor user errors. However, this setting led the system to misinterpret several words. For example, the word 'hi' was frequently misread as 'Thai,' causing the system to search for Thai restaurants after a simple greeting. Additionally, an update intended to prevent the word 'the' from being misinterpreted due to the Levenshtein distance was not applied to all recommendation systems on the computers used for testing. This oversight led to misinterpretations in two out of the five dialogue systems used for testing, resulting in incorrect classifications that may have negatively impacted users' perceptions of the system.

The second flaw of the system was a conversation loop that participants could get stuck in. The 'ADDITIONAL_PREFERENCES' state required specific input to advance to the next state. Many participants got trapped in a loop where the system repeatedly asked, "Please give a valid additional preference," preventing them from completing the task. For the sake of consistency, we

were not allowed to provide participants with instructions on how to move the system to the next state. Both of these flaws may have interfered with participants' perceptions of the system, potentially reducing the effect of the response delay on their perceptions.

These issues could have been mitigated by conducting more thorough testing before beginning the experiments and by taking more time to evaluate the system between participants or replacing problematic trials with new participants.

Limitations of Questionnaires

Secondly, as mentioned in the Methods section, we used the Godspeed Questionnaire and the System Usability Scale to quantify participants' opinions of the recommendation system. These questionnaires rely on highly associative terms, which can be subjective and vague. While they work sufficiently well with a large participant group, the results can be quite unpredictable with a smaller sample size. We suspect that our group of 24 participants may have been too small to accurately quantify perceived humanness and usability. Nonetheless, we believe these questionnaires remain the best option for measuring such broad and subjective concepts.

Limitations of participant group

Thirdly, our participant group imposes some limitations on our research. We studied a small sample of about 24 participants, and there is no guarantee that the observed effects would be statistically significant in a larger population. Additionally, all of our participants were between 20 and 25 years old, which could introduce age bias. Finally, there was an imbalance in education level, as several participants were students in the AI master's program, and only two participants were not students. It is unclear whether participants had prior experience with dialogue systems. However, their age group and educational background might suggest they have an above-average amount of experience with dialogue systems, which could influence their expectations of the system.

The relationship between prior experience and expectations is illustrated by the work of Gnewuch et al., 2022. Their research concludes that the effects of a delay depend on user characteristics. They found that the delay had a positive impact on novice users. Further research

suggests that the user's motivation for interacting with the dialogue system—whether out of curiosity or for entertainment—can also influence the effect of delay (Brandtzaeg & Følstad, 2017). This is particularly true for users with less experience using such systems. Gnewuch et al., 2022 also point out that a delay may not work as effectively for a different group of users, namely those who are more experienced with dialogue systems or AI tools. For this group, the delay tends to harm their perceptions of the dialogue system. This research underscores the importance of discovering participants' prior experience with dialogue systems, as an imbalance in this area may skew the results.

One more interesting point to discuss is the chosen delay. As explained in the Methods section, we chose a response delay of one second. However, prior research that did find significant results uses a variable delay time, where the length of the delay depends on the number of words in the response (Gnewuch et al., 2018). It might be desirable to imitate the prior research in that regard and include a variable response delay instead of a fixed delay.

Conclusion and future work

As stated, no significant results were found in this experiment. Whether this is because there is truly no relationship between humanness, usability, and response time remains uncertain. However, we consider it unlikely that there is no difference, as this would contradict prior research. We believe that valuable contributions could be made to understanding the humanness and usability of AI systems by expanding on this experiment. For example, this could be achieved with a more optimized system and a variable response delay. Additionally, with sufficient time and resources to gather a larger and more diverse participant pool, the results could become more informative. We also believe it would be interesting to quantify participants' prior experience with and expectations of dialogue systems, as suggested by the research of (Gnewuch et al., 2022).

Overall, our findings highlight the complexity of user perceptions in dialogue systems and stress the importance of refining system responsiveness and understanding how timing nuances shape human-AI interactions.

References

Appel, J., von der Pütten, A., Krämer, N. C., & Gratch, J. (2012). Does humanity matter? analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction*, *2012*(1), 324694. https://doi.org/https://doi.org/10.1155/2012/324694

Bartneck, C. (2023, February). Godspeed questionnaire series: Translations and usage. https://doi.org/10.1007/978-3-030-89738-3_24-1

Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2008). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*, 71–81. https://doi.org/10.1007/s12369-008-0001-3

Brandtzaeg, P., & Følstad, A. (2017). Why people use chatbots. https://doi.org/10.1007/978-3-319-70284-1_30

Brooke, J. (1995). Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, *189*.

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. *Research Papers*, (113). https://aisel.aisnet.org/ecis2018_rp/113

Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2022). Opposing effects of response time in human–chatbot interaction. *Business & Information Systems Engineering*, *64*(6), 773–791. https://doi.org/10.1007/s12599-022-00755-x

Holtgraves, T., & Han, T.-L. (2007). A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods*, *39*(1), 156–163. https://doi.org/10.3758/BF03192855

Holtgraves, T., Ross, S., Weywadt, C., & Han, T. (2007). Perceiving artificial social agents. *Computers in Human Behavior*, *23*(5), 2163–2174. https://doi.org/https://doi.org/10.1016/j.chb.2006.02.017

Miller, R. B. (1968). Response time in man-computer conversational transactions. *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 267–277.

Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, *32*(3), 736–751. https://doi.org/10.1287/isre.2021.1015

Appendix

Table with contributions per student

Table 2 shows how many hours each group member contributed to the project.

	Alexia	**Harmjan**	**Kaja**	**Pepijn**	**Yu-Lan**
Meetings	10.5	8	10.5	7.5	10
Experiment	4	4	2	2.5	2
Peer Review		1	2	2.5	
Introduction Section		4	2		
Methods Section				2.5	2
Results Section	3			0.5	
Discussion Section		1	2	2	0.5
General Report	4	1.2	4	11.2	4
Total	**21.5 hrs**	**19.2 hrs**	**22.5 hrs**	**26.5 hrs**	**18.5 hrs**

Table 2

Overview of contributions

Experimental task set-up

	First experiment		Second experiment (after 10 mins)	
Participant 1	With delay	Task 1	Without delay	Task 2
Participant 2	With delay	Task 1	Without delay	Task 2
Participant 3	With delay	Task 1	Without delay	Task 2
Participant 4	With delay	Task 1	Without delay	Task 2
Participant 5	With delay	Task 1	Without delay	Task 2
Participant 6	With delay	Task 1	Without delay	Task 2

Participant 7	Without delay	Task 1	With delay	Task 2
Participant 8	Without delay	Task 1	With delay	Task 2
Participant 9	Without delay	Task 1	With delay	Task 2
Participant 10	Without delay	Task 1	With delay	Task 2
Participant 11	Without delay	Task 1	With delay	Task 2
Participant 12	Without delay	Task 1	With delay	Task 2
Participant 13	With delay	Task 2	Without delay	Task 1
Participant 14	With delay	Task 2	Without delay	Task 1
Participant 15	With delay	Task 2	Without delay	Task 1
Participant 16	With delay	Task 2	Without delay	Task 1
Participant 17	With delay	Task 2	Without delay	Task 1
Participant 18	With delay	Task 2	Without delay	Task 1
Participant 19	Without delay	Task 2	With delay	Task 1
Participant 20	Without delay	Task 2	With delay	Task 1
Participant 21	Without delay	Task 2	With delay	Task 1
Participant 22	Without delay	Task 2	With delay	Task 1
Participant 23	Without delay	Task 2	With delay	Task 1
Participant 24	Without delay	Task 2	With delay	Task 1

Table 3

Overview of task setup among the participants