Pre-print of full paper presented at CONVERSATIONS 2022 – the 6th International Workshop on Chatbot Research, Applications and Design, hosted by the University of Amsterdam, Netherlands, November 22-23. The final version of the paper will be published in the forthcoming Springer LNCS post-workshop proceedings.

Voice your Opinion! Young Voters' Usage and Perceptions of a Text-based, Voice-based and Text-Voice combined Conversational Agent Voting Advice Application (CAVAA)

Christine Liebrecht✉[0000-0002-6621-2212], Naomi Kamoen[0000-0002-8719-2417], and Celine Aerts

[1] Tilburg School of Humanities and Digital Sciences, Department of Communication and Cognition, Tilburg University, Tilburg, The Netherlands

C.C.Liebrecht@tilburguniversity.edu,
N.Kamoen@tilburguniversity.edu

Abstract. Conversational Agent Voting Advice Applications (CAVAAs) are chatbot-based information retrieval systems for citizens who aim to inform themselves about the political issues at stake in times of political elections. Previous studies investigating these relatively young tools primarily focused on the effects of CAVAAs that include a text-based chatbot. In order to further optimize their design, current research compared the effects of CAVAAs with a text, voice, and combined chatbot. In an experimental lab study among young voters (*N*=60) these three modalities have been compared on usage measures (the amount of information retrieved from the chatbot and miscommunication), evaluation measures (ease of use, usefulness, and enjoyment), and political measures (perceived and factual political knowledge). Results show that the three CAVAA modalities score equally high on political measures and the perception of enjoyment. At the same time, the textual and combined CAVAA outperform the voice CAVAA on several aspects: the voice CAVAA received lower ease of use and usefulness scores, respondents requested less additional information, and they experienced more miscommunication when interacting with the voice chatbot. Analyses of the usage data also indicate that in the combined condition users hardly use the voice-option and instead almost exclusively rely on text-functionalities like clicking on suggestion buttons. This seems to suggest that using voice is too much of an effort for CAVAA users; we therefore recommend the usage of text-bots in this specific usage context.

Keywords: Voting Advice Applications; Conversational Agents; Chatbot Modality; Usefulness; Ease of Use; Enjoyment; Political Knowledge

1 Introduction

While CAVA is Spanish sparkling wine, CAVAA is the abbreviation for Conversational Agent Voting Advice Application. Similar to regular Voting Advice Applications (VAAs) that are very popular in elections times [10], such as the Dutch *Stemwijzer*, German *Wahl-O-Mat*, and Swedish *Valkompassen*, CAVAA users answer political attitude statements about relevant political issues (e.g., 'Taxes on housing should be increased') and receive a voting advice based on their answers [6, 19]. In contrast to a regular VAA, however, a CAVAA has an integrated chatbot functionality that users can address if they experience comprehension problems when answering the political attitude statements. The chatbot in a CAVAA is trained to provide information relevant for solving frequently occurring comprehension problems (also see [13]); it can for example provide information about the definition of political terms ('What is taxes on housing?'), or about the current state of affairs with respect to the political issue ('How high is the taxes on housing at the moment?'). In contrast to other chatbots in the political domain that provide subjective information about the standpoints of a single candidate running in the elections [17], a CAVAA hence aims to provide objective information voters can use to form a well-considered answer to the political attitude statements, which should ultimately lead to a valid voting advice.

Research demonstrates that CAVAAs are valuable information retrieval systems for citizens, as these tools receive better user experience scores than regular VAAs without a chatbot functionality [14]. Moreover, citizens score higher on political knowledge measures after having worked with a CAVAA than after working with a regular VAA [14]. In light of these positive findings, it is now time to further optimize the design of the chatbot in a CAVAA. Since the young audience that forms the prime user group of (CA)VAAs is becoming more and more experienced with voice technology in their daily lives [7], the current study will explore the effects of different chatbot modalities for young voters (aged 18-25), comparing a text-based CAVAA to a voice-based CAVAA and a CAVAA that combines these two modalities.

The three CAVAA versions are compared with respect to three types of outcome measures. First, we compare the tools with respect to a set of usage measures (the number of questions asked to the system, and the amount of miscommunication occurring) to get an idea of how modality might affect the extent users feel invited to ask questions to the chatbot (compare [14]). Second, we include subjective tool evaluation measures (ease of use, usefulness, enjoyment) that are commonly used in the field of chatbot research to evaluate a chatbot's design characteristics (compare [32]). Finally, we compare the three versions with respect to political knowledge measures (perceived and factual political knowledge) that are common outcome measures in studies on VAAs (compare [15]), which ultimate goal is to boost the user's political knowledge in order to increase the chance of casting a vote [11].

In comparing these three chatbot modalities on a range of dependent variables in a specific goal-oriented usage context and for a homogeneous group of young voters, current study will not only contribute to research on (CA)VAAs, but also to chatbot research in general since there is a strong call for studies on specific chatbot design characteristics in specific domains and across specific user groups [8]. This call is

especially relevant for modality effects, as these has been shown to depend on the interaction between the user, the technology and the task the user is trying to accomplish [27]. Another relevant aspect of current research for the domain of chatbot research is that not only a full text-based version and a voice-based chatbot are compared, but also a third version combining text and voice; there has been a call for the integration of such a version in previous chatbot studies [18, 27].

1.1 Modality

Text-based and voice-based chatbots each have their own distinctive qualities. For example, typing a question in a text-based chatbot allows users to check their input for correctness before sending a message to the system [2, 18]. Moreover, a chatbot's written response can be read at the user's own pace [18], which might increase the user's perception of control over the interaction [30]. Finally, in situations where users are not able – or not willing – to use speech, a text-based tool is more accessible [24].

By contrast, voice-based chatbots are easier to use in situations where users cannot type messages themselves or cannot read the chatbot's written answers, for example when using a chatbot while cooking or driving [24]. Furthermore, voice-based technology is generally more intuitive and speaking is also faster than typing [2, 26]. This, however, does not imply that using a voice-bot is more efficient than using a text-based bot, as in a voice-context the user is forced to listen to the full output of the chatbot whereas in a text-based chatbot the user can skim or even skip information efficiently [26]. A final advantage of voice-based chatbots is that voice interaction promotes human-like perceptions, which may result in more enjoyment during the interaction [25, 33].

While both text and voice have their distinctive qualities, or perhaps, *because* both modalities have their own distinctive qualities, neither can be seen as superior. Several studies show that user perceptions of a chatbot's modality depend both on characteristics of the task and characteristics of the user working with the application. This for example shows from [5], who found voice interaction to lead to more positive attitudes than text interaction, but only with utilitarian tasks (e.g., 'When is Father's Day in 2017?') and not for hedonic ones (e.g., 'Tell me a bedtime story'). Moreover, [27] compared two task contexts, a goal-directed information search task (i.e., searching for a restaurant with predefined instructions) and an experiential search task (i.e., searching for a restaurant without predefined instructions) and found that compared to users of the text-based chatbot users of the voice-based chatbot did experience more cognitive work load and more enjoyment when performing the goal-directed task. Though, the task did not impact voice-based users' perceptions of efficiency. Finally, [26] established that there are correlations between certain user characteristics and the preference for a text or voice modality, consequently extending prior studies that focus on demographic characteristics such as age [28].

The Task-Technology Fit theory (TTF; [12]) can be used to explain these differences in the usage and perception of text-based and voice-based chatbots dependent on characteristics of the task and the user. The theory postulates that the task, the individual user, and the functionalities of the technology should match to result in positive

performance outcomes. If individuals perceive a high fit between the task and the technology, they experience the technology to be more effective and efficient. By contrast, in case of discrepancies between the user, the task and the technology, the system will receive less good evaluations [12].

In light of TTF, it is relevant to explore how user, task and technology interact in the specific context of CAVAAs. This context can be regarded as utilitarian and goal-directed, since the user wants to gather information about politics and ultimately aims to receive a concrete voting advice from the application. Compared to previous chatbot modality studies that were mainly conducted in (fictitious) customer service contexts (e.g., [20, 26, 27]), the CAVAA context can hence be seen as more cognitively demanding since users try to solve their (real) comprehension problems about political issues by asking questions to the chatbot. As for user characteristics, it is known that CAVAA users make only a minimal effort to actually gather the required information before answering the political statements (see the findings of [13, 14]). Hence, our study can be seen of a study on modality effects in a cognitively demanding context where users make a minimal effort.

It is hard to predict how a text-based, a voice-based and a combined chatbot will be used by young voters in the specific usage context of CAVAAs. On the one hand, reasoning from the TTF, one might expect that a combined CAVAA leads to most intensive usage, and hence, to users asking most questions to the system. This might be expected because users can decide themselves which modality they use and they might even switch between modalities during the interaction. On the other hand, since (CA)VAA users have been shown to make only a minimal effort when working with the tool [13, 14], it might also be the case that switching between modalities is too much an effort and that users keep working in one of the two modalities in the combined version. For the CAVAAs that contain one modality (voice versus text), it can be reasoned on the one hand that the voice-based chatbot will lead to more information requests, since this modality is more intuitive to use and speaking is faster than typing [2, 26]. On the other hand, interpreting voice-output correctly is harder than interpreting text-output that users can process at their own pace [18, 26].

In light of these different possible scenarios it is hard to formulate a concrete hypothesis about the effect of modality on usage measures. As how users evaluate the tool and also to how much political knowledge they retain are expected to be the result of the actual usage of the tool, we will also refrain from formulating concrete hypotheses about these dependent variables. Instead, we will explore the differences between the three CAVAA modalities for both the usage measures, the evaluation measures, and the political measures.

2 Method

2.1 Design

In a between-group experimental study, we compared a CAVAA with a text-based chatbot, a voice-based chatbot, and a combined chatbot on several outcome measures.

In the experiment, the CAVAAs were distributed in a laboratory setting to a homogenous group of eligible Dutch young voters. Each participant worked with only one of the three CAVAA versions, and subsequently filled out a survey in which the evaluation measures (perceived ease of use, usefulness, enjoyment), and political measures (perceived and factual political knowledge) were measured. In addition, the actual usage of the chatbot modalities was measured by analyzing the chatlogs of participants' CAVAA conversations on the types of information requested, the usage of predefined buttons or free input to obtain information, and the appearance of miscommunication with the chatbot. On December 16, 2021, the research project received ethical approval from Tilburg University's Ethics Review Board (TSHD_RP174).[1]

2.2 Participants

We recruited a convenience sample of 60 young Dutch voters between 18 and 24 years old via the participant pool of our university (M_{age} = 20.3 years; *SD* = 1.88). Of them, 13 participants (21.7%) identified themselves as male, 46 as female (76.7%) and 1 participant (1.7%) identified outside the gender binary. All participants had Dutch as a native language and were registered with a Dutch municipality, and hence eligible to vote. Of the participants, 11 (18.3%) had never voted before, 37 participants (61.7%) had voted in one previous election and 12 participants (20%) had voted in multiple previous elections.

We compared participants in the text (*N*=19), voice (*N* = 20) and combined (*N* = 21) condition with respect to the above mentioned demographic characteristics and found no differences in prior voting experience (χ^2 (4) = 2.02, *p* = .73), gender (χ^2 (4) = 3.52, *p* = .48) and age (*F* (2, 57) = 3.70, *p* = .36). This implies that there is no reason to assume that there were *a priori* differences between the participants in the text condition, the voice condition, and the combined condition.

2.3 Materials

Development process. The three CAVAA versions were developed in collaboration with chatbot developer Genius Voice (geniusvoice.nl). This company designed the look and feel of the chatbots, and trained them to improve intent recognition. To check the functionalities, we pretested our three CAVAA versions among nine participants (three per version) before the start of the experiment. Based on these pretests, several improvements to the CAVAAs were made. Below we will describe the experimental materials as they were used in the final experiment. All experimental materials can be found in Dataverse (https://doi.org/10.34894/MNMLAT).

Modality. In the text-based condition, users could interact with the chatbot either by clicking on suggestion buttons or by typing in their messages in an open chat window themselves. They hence always used typed text or clicking to consult the chatbots, and they always received an answer in written text in return.

[1] The research project was approved by the university's Ethics Review Board (TSHD_RP174). Data collection was initiated prior to final approval, following liaising with the ethics board.

In the voice-based condition, the user and the chatbot communicated via speech. Just like in the text condition, users in the voice condition were shown suggestion buttons in written indicating the types of information they could request, but these buttons were not clickable; instead, users had to read the suggestion buttons out loud to activate them. Moreover, they could also formulate questions themselves by means of free speech, comparable to the open chat function in the text condition. When the user asked a question in the voice condition, the chatbot would display the answer in text on the screen and also read out the answer aloud.

In the combined condition, users could communicate with the chatbot via both modalities and were able to switch between text and voice during the conversation. This means that they could activate the suggestion buttons by either clicking on them, or by using the voice functionality to activate the content. Moreover, the chatbot's answers were visible on the screen, and when users would activate the sound button, the chatbot answers were also read out loud. Only in the combined condition it was hence possible to switch between the two modalities during the course of filling out the CAVAA. Figure 1 provides an example of the combined condition and describes the look and feel of this condition related to the other two conditions.

Statements. The CAVAAs' content and conversational flow were based on the experimental materials of [15], who developed a CAVAA for the Dutch National Elections in 2021. In total, 16 political attitude statements from that study were also included in the current research, as they were still topic of debate at the time we developed the materials for the current research. We added two new statements to come to 18 political attitude statements in total, which is the minimal number of VAA statements identified in a corpus study analyzing VAAs in national elections [31]. A (translated) example of a statement is 'There should be a binding referendum with which citizens can stop laws being implemented'.

Users could indicate their opinion towards each statement by answering 'agree', 'neutral' or 'disagree'. These answer options were visualized with a green ('agree'), grey ('neutral'), and red ('disagree') button below each statement (see Figure 1). After answering all attitude statements, the CAVAA provided the user with a personalized voting advice in which the user's standpoints were matched with the standpoints of the eight most prominent political parties of the Netherlands (similar to [15]).

Information Types. The chatbots were developed with conversational framework Rasa (rasa.com) and trained to recognize the intents of the users on the basis of an extensive list with potential questions users could ask per attitude statement, including synonyms (e.g., 'disadvantages', 'downsides', 'cons') and abbreviations (e.g., 'Partij van de Arbeid', 'PvdA'). These training data led to a NLU-model; the intent-entity combinations subsequently determined the chatbots' output to the user. Based on the user's input, the chatbots provided users with four types of information for each attitude statement in the tool; these types of information were based on the types of questions users have when answering political attitude statements [13].

First, the chatbots were trained to provide semantic information, which means that the chatbots could explain the meaning of a difficult word in the question (e.g., 'What does a binding referendum mean?'). Second, the chatbots were trained to provide pragmatic information about the current state of affairs with respect to the political issue in

the statement (e.g., ‘What is the current status with respect to referendums in the Netherlands?’). In addition to semantic and pragmatic information, the chatbot was also able to provide information about the advantages and disadvantages of the policy in the statement (e.g., ‘What is an advantage of implementing binding referendums?’), and about the standpoints of the political parties towards the statement (e.g., ‘What is the standpoint of the PvdA on binding referendums?’). The four information types were shown below the statement by means of four suggestion buttons, but users could also access information by phrasing questions themselves. The information the chatbot provided in response to users’ questions was preformulated by the researchers and always based on reliable resources, such as government websites, online dictionaries, news articles, and existing voting aids (similar to [14, 15]).

Conversation Flow. The conversation between the CAVAA and the user started with the chatbot greeting the user. Thereafter, the first statement was shown. The user could choose to either respond directly to the statement, or to ask for additional information first.

To enhance the dialogical character of the chatbot, we added conversational sentences in three different ways. First, an information request was always introduced with a conversational sentence (e.g., ‘Thanks for your question’, ‘I looked it up for you’). Second, after showing the additional information, the chatbot repeated the statement preceded by a conversational sentence (e.g., ‘So, the statement was …’, ‘Is there anything else you’d like to know before answering the statement?’). Third, the transition between statements was marked with a conversational sentence (e.g., ‘I have registered your answer, let’s move on to the next statement’). In all three chatbot versions, the chatbot randomly selected conversational sentences from a list, so that the three experimental conditions contained the same variation in conversational elements.

The chatbots were also equipped with a set of error responses in case they did not understand the user’s input or could not find a fitting answer. These responses consisted of an error notification (e.g., ‘Sorry, I don’t understand your question’) and a repair strategy (e.g., ‘Could you reformulate it in different words?’). In case miscommunication still occurred, the chatbot’s error response hinted on the four types of information that could be requested by the user (e.g., ‘Unfortunately, I cannot answer that question. I can give you some information on …’), which is proven to be a successful recovery strategy in chatbot conversations [3, 4].

Fig. 1: Screenshot of the combined CAVAA showing the first statement, the three answer options, and four suggestion buttons. Below the suggestion buttons, there are two icons (marked with a dotted line in the figure) that could be used to turn the sound (for output) and microphone (for input) on and off. Moreover there is an open text field to enter a question in written (marked with a dashed line in the figure). In the voice-condition, the open text field was not present and the sound and microphone icons were always on, as in this condition voice was the only modality that could be used to control the CAVAA; in the text-based condition the open text field was visible and the icons to turn the sound and microphone on/off were not displayed, as this CAVAA could only be controlled by using text.

2.4 Usage Measures

For the usage measures we analyzed a sample of 60 (participants) * 18 (statements) = 1,080 respondent and item combinations. This sample was coded on the types of information requested by the participants, and whether miscommunication occurred between the user and the chatbot. For the combined condition, we also scored which modality the participants used to request information. A second coder coded a random subsample of 17 chatbot conversations (28%), divided across the three CAVAA versions. The intercoder reliability was always acceptable (semantic information $\kappa = 0.97$, pragmatic information $\kappa = 0.96$, party standpoints $\kappa = 1.00$, (dis)advantages $\kappa = 0.95$; miscommunication $\kappa = 0.76$).

2.5 Evaluation measures and Political Measures

In an online survey, the evaluation measures were examined first, followed by the political measures. Except for the factual knowledge questions, all survey questions could be answered on a seven-point scale ranging from ‘fully disagree’ to ‘fully agree’.

Enjoyment. The questionnaire started with three statements to measure participants’ enjoyment while using the CAVAA. The three items were adapted from a survey in an earlier study by [21] and modified to fit the context of the current study (e.g., ‘I found using the chatbot a pleasant experience’). The three items showed to group well together ($\alpha = .88$).

Ease of Use. The ease of use of the chatbot was measured with five items, adapted from the study of [1] (e.g., ‘I found this chatbot user friendly’). The five items clustered well together ($\alpha = .81$).

Usefulness. Usefulness was measured with four items based on [1] and modified to fit the context of the current study (e.g., ‘Using this chatbot enabled me to answer the statements better than a regular voting aid’). The four items provided a reliable measure ($\alpha = .73$).

Perceived Political Knowledge. Participants’ perception of political knowledge after using the CAVAA was measured by adapting four statements from a study by [29] (e.g., ‘By using this chatbot, I gained more knowledge about the political landscape’). The four items showed to group well together ($\alpha = .69$).

Factual Political Knowledge. Eight true/false statements were presented on topics related to the political attitude statements in the CAVAA. Participants answered these

statements with either ‘true’, ‘false’ or ‘I don’t know’. The latter answering option was included to avoid guessing behavior of participants that could affect the reliability. For the data analysis, the answers to the eight knowledge questions have been recoded. The correct answers have been coded with 1 and both the incorrect answers and the ‘I don’t know’ answers have been coded with 0. This led to a factual political knowledge score between 0 and 8 for every participant.

2.6 Procedure

The study was conducted in December 2021, approximately three months prior to municipal elections in the Netherlands. All participants were recruited via the Human Subject Pool of our university and took part in the experiment in the lab (taking the Corona measures at the time into account). Before starting the experiment in one of the sound proof cabins, participants were given a brief instruction on what Voting Advice Applications are and how to specifically use the CAVAA in the current study. Subsequently, they started the study and were asked to provide informed consent for the usage of their data. It was stressed that participation was completely voluntary and participant could stop at any point in time. After having provided informed consent, participants answered several questions about demographic variables. Next, participants could click on a link that directed them to one of the three CAVAA versions that opened in a new window. After having answered all 18 political statements in the CAVAA, a voting advice was provided. Thereafter, the participant was redirected to the online survey that included the evaluation measures and the political knowledge measures. The questionnaire ended with a debriefing in which the participants were informed about the purpose of the study. In total, the experiment took approximately 20 minutes and all participants received a partial course credit in return.

3 Results

3.1 Usage Measures

The means and standard deviations for the usage measures are shown in Table 1. Across all four types of information a respondent could request, there was a difference between the chatbot conditions: in both the text condition and the combined condition respondents more frequently requested at least one of the four forms of information than in the voice condition (text vs. voice: $\chi^2 = 4.52$, $p = .03$; combined vs. voice: $\chi^2 = 6.76$, $p = .009$). There was no difference between the text and combined condition ($\chi^2 = 0.03$, $p = .86$).

If we split out this analysis per type of information, it can be seen that in the voice condition less information about the advantages and disadvantages of a certain policy was requested compared to the combined condition ($\chi^2 = 7.28$, $p = .007$), and also that in the voice condition less information was looked up about the party stances compared to the text condition ($\chi^2 = 5.76$, $p = .02$). All other contrasts failed to reach significance. Also, no significant differences between the three CAVAA conditions were found with

respect to the retrieval of semantic and pragmatic information (in all cases: $\chi^2 < 3.55$, $p > .06$), although the tendencies for differences (p-values between 0.06 and 0.1) indicate that in a larger sample there may be more information requests found for both the text and the combined condition compared to the voice condition.

Table 1. Proportion of times a type of information was requested (Logit and SE between brackets), and the accompanying variances (in Logit) (*M*) for each experimental condition.

	Semantic	Pragmatic	(Dis)advantages	Party Stances	Total
Text	20.8% (-1.34; 0.33)	19.6% (-1.41; 0.25)	29.7% (-0.86; 0.24)	19%* (-1.45; 0.32)	59.2%* (0.38; 0.30)
Voice	15.3% (-1.71; 0.33)	12.2% (-1.97; 0.26)	21.7% (-1.29; 0.23)	7.2% (-2.55; 0.33)	43.1% (-2.76; 0.27)
Combined	21.4% (-1.30; 0.34)	18.8% (-1.46; 0.19)	40.1%* (-0.40; 0.25)	11.4% (-2.05; 0.32)	60.6%* (0.43; 0.26)
$S^2_{resp.}$ Text	0.12	0.46 (0.27)	0.86 (0.34)	1.67 (0.64)	0.83 (0.35)
$S^2_{resp.}$ Voice	0.04	0.50 (0.32)	0.78 (0.32)	1.45 (0.68)	0.54 (0.25)
$S^2_{resp.}$ Comb.	0.29	0.07 (0.13)	1.05 (0.37)	1.67 (0.66)	0.51 (0.23)
S^2_{items}	1.51	0.31 (0.14)	0 (0)	0 (0)	0.57 (0.22)

* indicates a significant difference (p <.05) with the voice condition

We also run a Loglinear Multi-level model similar to [14] to compare the number of times a respondent experienced miscommunication. There was no miscommunication observed in the combined condition (0%), only very little miscommunication in the text condition (1.4%), whereas in the voice condition in about 12.8% of the respondent and item combinations some form of miscommunication occurred. The differences between the text and voice condition were indeed found to be significant ($\chi^2 = 3.20$, $p < .001$; see Table 2). The combined condition could not be included in the analysis due to a lack of variance.

As there was no miscommunication in the combined condition and quite a lot of miscommunication in the voice condition, a relevant question is as to how frequently users used the voice option in the combined condition. The combined condition contained 21 participants and they all responded to 18 statements about politics, so there were 378 respondent and item combinations. In only 12 of these cases (3.2%) respondents used their voice to request information. This means that the voice functionality was hardly used and respondents used the text option even if they had the possibility to control the chatbot with their voice.

Table 2. Proportion of times that miscommunication occurred (Logit and SE between brackets), and the accompanying variances (in Logit) (*M*) for each experimental condition.

	Miscommunication
Text	1.4%* (-4.21; 0.69)
Voice	12.8% (-1.92; 0.18)
Combined	0 (0)
$S^2_{resp.}$ Text	5.49
$S^2_{resp.}$ Voice	0.20
$S^2_{resp.}$ Comb.	0
S^2_{items}	0

* indicates a significant difference (p <.05) with the voice condition

3.2 Evaluation and Political Measures

In the survey, participants evaluated the CAVAA's enjoyment, ease of use, and usefulness. Furthermore, both perceived and factual political knowledge were measured. The means and standard deviations of all dependent variables can be found in Table 3.

Table 3. Means (*M*) and standard deviations (*SD*) between brackets per dependent variable and per experimental condition.

	Enjoyment	Ease of Use	Usefulness	Perceived Knowledge	Factual Knowledge
Text (N =19)	5.60 (1.09)	6.13 (0.68)*	6.12 (0.80)*	4.91 (0.85)	5.11 (1.79)
Voice (N=20)	5.40 (1.12)	5.21 (1.30)	5.34 (1.11)	5.14 (0.93)	5.25 (2.00)
Combined (N = 21)	5.63 (0.49)	6.18 (0.53)*	6.01 (0.70)	5.11 (0.86)	5.57 (1.40)

* indicates a significant difference (p <.05) with the voice condition

For each dependent variable, a Factorial ANOVA was conducted to examine whether this dependent variable was dependent on the modality of the CAVAA. For ease of use, there was a modality effect (F (2, 57) = 7.38, p = .01). A post hoc test (Bonferroni) indicated that both the text and the combined condition were easier to use than the voice condition (p = .007 and p = .003 respectively). Also, a modality effect was found for usefulness (F (2, 57) = 4.55, p = .02). A post hoc test (Bonferroni) indicated that in the text condition the CAVAA was evaluated to be more useful than in the voice condition (p = .02), and that there was also a tendency (p = .05) for CAVAA to be evaluated as more useful in the combined condition than in the voice condition. For enjoyment and the two political knowledge measures, no modality effects were observed (enjoyment: F (2, 57) = 0.37, p = .70; perceived political knowledge: F (2, 57) = 0.39, p = .68; factual political knowledge: F (2, 57) = 0.38, p = .69).

4 Discussion

We explored the effects of chatbot modality (text, voice, or combined) in the specific usage context of Conversational Agent Voting Advice Applications (CAVAAs). In contrast to earlier chatbot studies on modality effects in customer service contexts (e.g., [20, 26, 27]), the current usage context in which users tried to understand political attitude statements can be seen as more cognitively demanding and goal-oriented. In our study, we focused on a homogeneous group of young voters (aged 18-25), who are known to expose satisficing behaviour when working with a CAVAA, which means that they are only willing to make a minimal effort to find information [13, 14].

Results show that users' perceptions of ease of use and usefulness of the tools differ: both the text and the combined condition scored higher on these measures than the voice condition. From the results of our content analysis comparing the tools of several usage measures, two possible explanations can be formulated for these findings. First, users experienced more miscommunication when they used speech input. This miscommunication sometimes occurred when the user tried to request additional information, e.g., when the chatbot did not understand a question like 'Wait are the advantages' when the user probably meant '*What* are the advantages'. Most miscommunication, however, occurred when the user tried to answer the political attitude statements saying 'Agree', 'Neutral', or 'Disagree'. The voice bot for example sometimes thought the user said 'Eend' (the Dutch word for 'Duck') or 'Aids' ('Aids') when the user probably wanted to say 'Eens' (the Dutch word for 'Agree'). Similarly, the chatbot sometimes understood 'Centraal' ('central') when the user probably meant 'Neutraal' ('Neutral'), or 'Online' ('online') when the user meant 'Oneens' ('Disagree'). These different forms of miscommunication probably caused the user to feel less in control in the voice condition [30], which may have led to lower scores for ease of use and usefulness.

A second explanation for the lower scores on ease of use and usefulness for the voice condition is that in this condition users felt less invited to ask questions to the chatbot, which may have lowered the perception of ease of use and especially usefulness. A relevant question therefore is as to *why* users file less information requests in the voice condition. One explanation may be related to the previous point suggesting that users were afraid to experience miscommunication. However, as miscommunication was more frequently occurring when the user answered the political attitude statement rather than when asking a question to the system, an alternative explanation, is that users, who are known to make only a minimal effort to request information [13, 14], found using their voice a too big an effort. In our view, this explanation very plausible especially in light of our finding that in the combined condition users first and foremost used text (clicking) to request information and not voice. The occurrence of miscommunication does not count as an alternative explanation for the reliance on text in the combined condition, as we observed no miscommunication whatsoever in the combined condition. In our view, it is therefore likely that a textual communication mode simply fits the user better in the specific usage context of CAVAAs. To be more certain of what has caused the lesser amount of information requests in the voice condition, and therefore probably the lower scores on usability measures, however, it would be worthwhile to conduct a replication study with an improved version of the voice-based chatbot.

This chatbot should then be trained better to recognize respondents' answers to the political attitude statements. In addition, it would be valuable to combine such an experimental study with a cognitive interview afterwards asking users to indicate explicitly how much an effort they thought asking a question was.

Another result of the current study is that no differences between the three CAVAA versions were found for perceived enjoyment. This finding is in contrast with earlier studies showing that users of voice-based chatbot frequently enjoy the interaction [25, 33]. It seems that in the current study users overall enjoyed working with all three CAVAA versions a lot, showing from the relatively high mean scores for enjoyment (around 5), as well as the open comments users made at the end of the survey, such as: 'I really liked using the chatbot!', 'The chatbot helped me to understand the topics in the voting advice application; I really enjoyed using a chatbot in a voting advice application' and 'It felt naturalistic to talk to the chatbot. It is much nicer and more personal to do than just answering questions'. A possible explanation for as to why not just the voice-based CAVAA but all three versions received high scores for enjoyment, might be that CAVAAs are relatively new tools in general. Therefore, a novelty effect [9, 16] may have occurred across all three versions since experiences of enjoyment and novelty are closely related [23].

A final results of the current study is that we found no differences between the three modalities for perceived and factual knowledge. As we expected the effects of the political measures to be the result of the actual usage and evaluation of the tool, and as we did find modality effects for these latter measures, the absence of differences for the political measures may be unexpected. In all three conditions relatively high scores on perceived (means around 5 on a 7-point scale) and factual (means around 5 on an 8-point scale) knowledge were obtained. This may suggest that answering political attitude statements in a CAVAA, irrespective of modality, leads to relatively high scores on these political measures. To further understand this finding it would therefore be interesting for a future study to include not only a post-CAVAA measure of perceived and factual knowledge, but also a pre-CAVAA measure. This way the delta of these two measures can be calculated and used as a more fine-grained measure of perceived and factual knowledge.

5 Conclusion

The goal of this study was to explore how people use and perceive chatbots that differ in modality in the cognitively demanding context of Conversational Agent Voting Advice Applications. The participants' scores on perceived and factual political knowledge, as well as their perceived enjoyment scores did not differ between the chatbot conditions. However, differences were found for usefulness and ease of use: the voice-based CAVAA was both considered less easy to use and less useful than the other two modalities. The content analysis of the chatlogs revealed that users request more information in both the text condition and the combined condition as compared to the voice condition. Moreover, more miscommunication occurred between the tool and the user in the voice condition than in the other two conditions. Finally results showed that

in the combined condition users hardly used the opportunity to control the chatbot using voice and they relied on the type and click functionality in most of the cases. All in all, these results suggest that the combined condition in practice resembled the text condition and that these two conditions outperformed the voice condition in various respects. In order to achieve an optimal fit between users, task, and technology – as formulated by the TTF theory – chatbot developers in the context of political CAVAAs could best develop text-based chatbots, since such chatbots do not only avoid miscommunication, but also stimulate users to request more information in an easy way. This way CAVAAs can best help citizens to find political information. This should ultimately lead to more voters actually casting a vote in real-life elections and to a stronger democracy.

Acknowledgements. The authors would like to Tilburg University's Fund (project number ESF2021-2) for the financial support to develop the CAVAAs. A summary of the results of this study has also been published in the Dutch popular-scientific magazine *Tekstblad* [22].

References

1. Ahn, T., Ryu, S., Ahn, T., Ryu, S., Han, I.: The impact of Web quality and playfulness on user acceptance of online retailing. Information & Management **44(3)**, 263-275 (2007). https://doi.org/10.1016/j.im.2006.12.008.
2. Angga, P. A., Fachri, W. E., Elevanita, A., & Agushinta, R. D.: Design of chatbot with 3D avatar, voice interface, and facial expression. In 2015 International Conference on Science in Information Technology (ICSITech), pp. 326-330. IEEE (2015, October).
3. Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D.: Resilient chatbots: Repair strategy preferences for conversational breakdowns. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-12) (2019, May).
4. Bohus, D., & Rudnicky, A.: Sorry and I Didn't Catch That! An Investigation of Non- understanding Errors and Recovery Strategies. In Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue (pp. 128-143) (2005, September).
5. Cho, E., Molina, M. D., & Wang, J.: The effects of modality, device, and task differences on perceived human likeness of voice-activated virtual assistants. *Cyberpsychology, Behavior, and Social Networking*, **22(8)**, 515-520 (2019).
6. De Graaf, J.: The irresistible rise of Stemwijzer. In: Cedroni, L, Garzia, D. (eds.) Voting Advice Applications in Europe: The State of the Art, pp. 35-46. Napoli, Scriptaweb (2010).
7. Direct Research: De Nationale Voice Monitor 2021 (2021). Retrieved on February 8, 2022, from https://www.directresearch.nl/blogs/de-nationale-voice-monitor-2021/.
8. Følstad, A., Araujo, T., Law, E. L. C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., Ischen, C., Wald, R., Catania, F., Meyer von Wolff, R., Hobert, S., & Luger, E.: Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, **103(12)**, 2915-2942 (2021).

9. Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z.: Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, **75**, 461-468 (2017).
10. Garzia, D., Marschall, S.: Voting Advice Applications under review: The state of research. International Journal of Electronic Governance **5(3-4)**, 203-222 (2012).
11. Gemenis, K., & Rosema, M. (2014). Voting advice applications and electoral turnout. *Electoral studies*, **36**, 281-289.
12. Goodhue, D. L., & Thompson, R. L.: Task-technology fit and individual performance. *MIS Quarterly*, 213-236 (1995).
13. Kamoen, N., Holleman, B.: I don't get it. Response difficulties in answering political attitude statements in Voting Advice Applications. Survey Research Methods **11(2)**, 125-140 (2017). https://doi.org/10.18148/srm/2017.v11i2.6728.
14. Kamoen, N., & Liebrecht, C.: I Need a CAVAA: How Conversational Agent Voting Advice Applications (CAVAAs) Affect Users' Political Knowledge and Tool Experience. Frontiers in Artificial Intelligence, **5** (2022). https://doi.org/10.3389/frai.2022.835505
15. Kamoen, N., McCartan, T., & Liebrecht, C.: Conversational agent voting advice applications: A comparison between a structured, semi-structured, and non-structured chatbot design for communicating with voters about political issues. In International Workshop on Chatbot Research and Design (pp. 160-175). Springer, Cham (2021, November).
16. Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H.: Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, **19(1-2)**, 61-84 (2004).
17. Kim, Y., & Lee, H.: The rise of chatbots in political campaigns: The effects of conversational agents on voting intention. *International Journal of Human-Computer Interaction*, 1-12 (2022).
18. Kocielnik, R., Avrahami, D., Marlow, J., Lu, D., & Hsieh, G.: Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 designing interactive systems conference* (pp. 881-894) (2018, June).
19. Krouwel, A., Vitiello, T., Wall, M.: The practicalities of issuing vote advice: a new methodology for profiling and matching. International Journal of Electronic Governance **5(3-4)**, 223-243 (2012).
20. Le Bigot, L., Jamet, E., Rouet, J.-F., & Amiel, V.: Mode and modal transfer effects on performance and discourse organization with an information retrieval dialogue system in natural language. *Computers in Human Behavior*, **22**, 467-500 (2006).
21. Lee, M. K., Cheung, C. M., & Chen, Z.: Acceptance of Internet-based learning medium: the role of extrinsic and intrinsic motivation. Information & management, **42(8)**, 1095-1104 (2005).
22. Liebrecht, C., & Kamoen, N.: 'Hey Siri, wat is de hondenbelasting?': Voicebots en tekstbots in een politieke context. Tekstblad, **27(1)**, 22-24 (2022).
23. McLean, G., & Osei-Frimpong, K.: Hey Alexa… examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, **99**, 28-37 (2019).
24. Nass, C. I., & Brave, S.: Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. Cambridge, MA: MIT Press (2005).

25. Pal, D., Arpnikanondt, C., Funilkul, S., & Chutimaskul, W.: The Adoption Analysis of Voice-Based Smart Iot Products. *IEEE Internet of Things Journal*, **7(11)**, 10852-10867 (2020).
26. Riefle, L., Brand, A., Mietz, J., Rombach, L., Szekat, C., & Benz, C.: What fits Tim might not fit Tom: Exploring the impact of user characteristics on users' experience with conversational interaction modalities. *Wirtschaftsinformatik 2022 Proceedings*, **13** (2022).
27. Rzepka, C., Berger, B., & Hess, T.: Voice assistant vs. chatbot: examining the fit between conversational agents' interaction modalities and information search tasks. *Information Systems Frontiers*, **24(3)**, 839-856 (2022).
28. Schroeder, J., & Schroeder, M.: Trusting in Machines: How Mode of Interaction Affects Willingness to Share Personal Information with Machines. *Proceedings of the 51st Hawaii International Conference on System Sciences, Hawaii, USA* (2018).
29. Shulman, H. C, & Sweitzer, M. D.: Advancing Framing Theory: Designing an Equivalency Frame to Improve Political Information Processing. Human Communication Research, **44(2)**, 155-175 (2018). https://doi.org/10.1093/hcr/hqx006.
30. Sundar, S. S.: The MAIN Model: A heuristic approach to understanding technology effects on credibility. In: Digital Media, Youth, and Credibility. M.J. Metzger & A.J. Flanagin (Eds.). The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press, 73-100 (2008). doi:10.1162/dmal.9780262562324.073
31. Van Camp, K., Lefevere, J., & Walgrave, S.: "The content and formulation of statements in voting advice applications" in Matching voters with parties and candidates. Voting advice applications in comparative perspective, eds. D. Garzia, & S. Marschall, (ECPR Press, Colchester), 11-32 (2014).
32. Xu, J. D., Benbasat, I., & Cenfetelli, R.T.: The nature and consequences of trade-off transparency in the context of recommendation agents. *MIS Quarterly*, **38**, 379-406 (2014).
33. Yang, H., & Lee, H.: Understanding user behavior of virtual personal assistant devices. *Information Systems and e-Business Management*, **17(1)**, 65-87 (2019).