It’s Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks

LEON REICHERTS

University College London

YVONNE ROGERS

University College London

LICIA CAPRA

University College London

ETHAN WOOD

University College London

TU DINH DUONG

University College London

NEIL SEBIRE

University College London

Voice assistants have become hugely popular in the home as domestic and entertainment devices. Recently, there has been a move towards developing them for work settings. For example, *Alexa for Business* or *IBM Watson for Business* were designed to improve productivity, by assisting with various tasks, such as scheduling meetings and taking minutes. However, this kind of assistance is largely limited to planning and managing user’s work. How might they be developed to do more by way of empowering people at work? Our research is concerned with achieving this by developing an agent with the role of a facilitator that assists users during an ongoing task. Specifically, we were interested in whether the modality in which the agent interacts with users makes a difference: how does a voice versus screen-based agent interaction affect user behavior? We hypothesized that voice would be more immediate and emotive, resulting in more fluid conversations and interactions. Here, we describe a user study that compared the benefits of using voice versus screen-based interactions when interacting with a system incorporating an agent, involving pairs of participants doing an exploratory data analysis task that required them to make sense of a series of data visualizations. The findings from the study show marked differences between the two conditions, with voice resulting in more turn-taking in discussions, questions asked, more interactions with the system and a tendency towards more immediate, faster-paced discussions following agent prompts. We discuss the possible reasons for why talking and being prompted by a voice assistant may be preferable and more effective at mediating human-human conversations and we translate some of the key insights of this research into design implications.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI • Human-centered computing~Human computer interaction (HCI)~Interaction paradigms~Natural language

interfaces • Human-centered computing~Human computer interaction (HCI)~Interaction paradigms~Collaborative interaction • Human-centered computing~Human computer interaction (HCI)~Interaction paradigms~Graphical user interfaces • Human-centered computing~Human computer interaction (HCI)~Interaction techniques

Additional Keywords and Phrases: Voice user interface; voice assistant; interaction modality; proactive agent; agent intervention; question-asking; sensemaking; exploratory data analysis; conversation; collaboration; GUI

ACM Reference Format:

Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tudinh Duong, and Neil J. Sebire. 2021. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. ACM Trans. Comput.-Hum. Interact. XX, X, Article XX (2021), 47 pages.

1 INTRODUCTION

Digital voice assistants, such as Amazon's Alexa or Google Assistant have become mainstream technologies in the home, commonly used for tasks such as playing music, providing news updates or letting users control various smart home devices. They are now also being developed to help businesses, intended to improve productiveness through arranging meetings, taking notes, providing specific information and data, such as business-related metrics, and so on. Likewise, interfaces incorporating chatbots or text-based interactions with agents are gaining popularity in a variety of business contexts, providing similar uses, but through written language and screen-based interactions. The question this raises is: As agents become more "intelligent" and take on more diverse roles in various environments, what is the most effective way of interacting with them, especially for work settings when there is more than one human present, as is the case with meetings? While talking with other humans, is talking with an agent more effective compared with interacting with a chatbot/text-based agent by selecting options and reading its prompts on a screen? If so, is it because talking can enhance the flow of conversation while facilitating collaboration? Our research is concerned with investigating the potential differences between the two and their benefits.

In particular, we are interested in how people interact with either voice or screen-based agent interfaces when carrying out another activity, during meetings – where there is more than one person present, who will interact with it. Clearly, there are different affordances of interacting with graphical/screen-based interfaces compared with voice interfaces. Thus, our interest here is less on the difference in screen versus voice modality per se, but rather the *type of interactions* these interfaces require or enable at a general level. A screen-based interface requires reading information (e.g., the agent's text messages/prompts) and performing manual interactions (typing on a keyboard, using a mouse or a touchscreen). It has the advantage that users can decide when they want to process/read the agent's prompt. On the other hand, if the interface is voice-based, people in the meeting do not have to switch modalities when addressing the agent or being prompted by the agent during an ongoing conversation with other people. With a voice-based interaction, requests to the agent can be made as part of the conversation, and agent prompts will need to be processed/listened to immediately when they are provided, since speech requires immediate attention. Furthermore, talking and listening are well-honed skills that we employ when holding a conversation in the company of others. In contrast, selecting from menus or typing text in at an interface, via a touchscreen or keyboard, is more indirect and something usually done by

one user. The text as it appears on the screen may or may not be read and may require being read aloud by one in a meeting to let the other know they are reading it or for them to listen. Hence, it could be that a voice-based agent is a better match when the agent is intended to be part of a group setting, as it may be more effectively embedded into an ongoing conversation between multiple users.

In co-located multi-user scenarios, users can be engaged with the system in different ways. There are situations where only one person is interacting with a device (e.g., a computer) and the others are only observing the device's output, such as in a presentation, talk or demo. Then there are scenarios, where users are collocated around a device and interact with it simultaneously, depending on the interface, the interactions can be in parallel (e.g., tabletop) or need to be coordinated sequentially (e.g., computer). In the present study, our focus is on co-located simultaneous interactions with an agent-enabled data analytics interface comparing two modalities – screen and voice-based interactions.

While there is a variety of ways the user-agent interaction could be configured (e.g., the agent's output is through voice while the users' input is screen-based or vice versa), the aim of our research is to focus on two combinations that we consider most natural and appropriate for a variety settings and applications – voice-based input *and* output or screen-based input *and* output. In certain situations, outside the scope of this research, it may be more suitable to use other combinations of modalities, such as in an environment where it is not possible for users to speak in front of their computer or other device (e.g., an open office) but possible to hear the agent speaking (e.g., through headphones).

We report here on a user study we conducted to investigate the differences in the way conversations unfold and progress in a group setting when having voice versus screen-based interactions with a software tool, which "incorporates" an agent. The agent was designed with a specific role to play, namely, to "scaffold" and facilitate the participants' activities, prompting them and making suggestions for what to do next when progressing a task. In particular, we were interested in how pairs of users interacted with and responded to an agent for the two conditions when trying to understand and make inferences in a data analysis task where they had to explore a set of time series graphs. Exploratory data analysis was chosen as the task, as it is a key activity of knowledge workers in today's increasingly data-driven economy. While the modality of interacting with the agent-enabled system varied between the conditions, we kept it constant for the main task materials, so that in both conditions participants analyzed time series graphs. The reasons for this were three-fold, (i) our main research focus was the interactions with the agent and not the task material itself, (ii) it was considered more ecologically valid as most computer-supported tasks, including data analysis are usually visual, and (iii) it appears challenging to design a task for which the modality could also be varied (in addition to varying the modality of agent interaction) without introducing additional confounds.

While there has been considerable research investigating the effects on human performance of using speech versus text/screen-based input at the interface (e.g., [6,14,17,32,43,44,46,87]), and the use of speech versus text/screen-based output (e.g., [9,45,53,60,81], little is known as to whether using voice or screen-based interactions, when conducting a cognitive task, impacts upon (i) the way human-human interactions progress and (ii) how the humans interact with the agent. Furthermore, most of this research has focused on single-user scenarios. Here, we propose, from the users' perspective, that voice can be more engaging, sparking more curiosity and sensemaking, by triggering more questioning and hypothesizing during (collaborative) sensemaking. Separating the input/output modality from the ongoing activity taking place at the graphical user interface (i.e., voice-based input and output combined with a visual data analysis task), could also enable a

more natural and free-flowing conversation about the activity in hand. However, there is also the possibility that the users may ignore (or miss) what a voice agent has said or find it irritating when it interrupts their interactions. Conversely, interacting with an agent in the same modality as the ongoing sensemaking task (i.e., using visual screen-based input and output combined with a visual data analysis task) means that the users can decide when to read the agent's messages, although it is a more indirect form of interaction that may integrate less well with ongoing human conversations. When considering tasks with different modalities (e.g., an auditory task), other interface modality combinations may promise to be more effective. However, since most computer-supported tasks are (still) mostly visual in nature, our focus is on this modality (i.e., a visual data analysis task).

To investigate our hypotheses about the benefits of using a voice versus screen-based agent when carrying out a collaborative activity at the user interface, we conducted a user study that compared the two conditions. Pairs of participants were asked to discuss and make inferences about a series of data visualizations that were presented on a shared digital display. For this purpose, we developed a prototype called *Vizzy Analytics*, which uses a Wizard of Oz paradigm [68], to "mimic" an agent that proactively intervenes and makes suggestions. It does this by prompting the pairs at certain times, as to what they can look at in the visualizations, thus playing a role similar to a facilitator. The prompts were in the form of questions about certain trends and patterns in the data, which pairs can then discuss and try to answer together. The agent also acts upon their voice or screen-based requests/commands to select different visualizations for them to look at and compare. To assess the differences in the interactions that took place between the participants and between participants and the agent, both quantitative and qualitative data were collected and analyzed. Quantitative analysis (section 5.1 and 5.2) included the amount of interaction with the system as well as interactions between participants themselves, where the latter included conversational turn-taking and questions participants asked each other. Qualitative analysis involved investigating how participants responded to agent prompts, which was done via conversation analysis of transcribed segments (section 5.3), as well as an analysis of participants' reflections on the experience of doing the analysis task with the tool/agent, which were collected in post-study interviews (section 5.4). A main finding was that there was a significant difference between the two conditions in certain kinds of task-related behaviors and interactions: in the voice condition participants interacted more with the system and requested more visualizations, took more turns in the conversation and asked each other more questions related to the data visualizations. We discuss the possible reasons for why this is the case in relation to the cognitive benefits of voice's immediacy and directness, suggesting how employing voice versus screen-based agent interactions may be beneficial for scaffolding users' thinking, who are working together when carrying out a problem-solving/sensemaking task at the interface.

This paper makes three main contributions: it provides (i) empirical evidence for the effects of interaction modality with an agent-enabled system on users' conversations, interactions and their task-related behavior (section 5), (ii) evidence from qualitative analyses that a system, which prompts users through questions, can enable in-depth discussions and (collaborative) sensemaking (section 5.3 and 5.4) and (iii) design implications for similar systems based on the insights of this study (section 6.2).

2 BACKGROUND

Besides being standalone systems, conversational agents are now beginning to be integrated with other technologies and software applications. Voice assistants, in particular, are seen to offer much potential to guide and support users in conducting complex tasks at the user interface (e.g., [83]), such as data analytics. For

example, commercial applications include *SearchIQ* which is part of *ThoughtSpot* [88] and *Einstein Voice* which is embedded in the software tool *Salesforce* [89]. This combination of agent plus tool is thought to aid task efficiency. In particular, a putative benefit of being able to talk and listen to voice assistants while using data analytics tools is to make it easier for users to understand what the outputs and visualizations mean or to intuitively express data-related questions [66].

More generally, voice assistants have been found to be effective in supporting various types of domestic tasks [2], everyday tasks of people with impairments [58] and also collaborative tasks [83] to name a few. Voice as an interaction modality has been found to be effective at promoting problem-solving and reflection [51] and influencing users' understanding of their role in an interactive narrative task [25]. With the prospect of future advances in AI, there is also the potential for voice assistants to become more intelligent, capable of supporting and guiding users through ever more complex tasks. However, it is not straightforward as how best to support users when provided with more intelligent interfaces. How can we ensure that the next generation of agents do not simply automate what humans currently do, but instead amplify and empower them in their activities (see also Shneiderman [70])? For example, should the agent provide the user with solutions, or should it rather assist users in finding a solution themselves by probing them and asking them questions?

When considering how a conversational agent should appear and how it should interact with users, it also raises the question of what might be the optimal way for users to interact with it, especially when it is intended to be used in a particular setting, such as a meeting, and in combination with using a software tool to perform a given task. What parts of human communication should the agent be programmed to mediate and how best should an agent reply – in terms of suggesting, augmenting, or other ways of responding that help towards achieving an outcome [31]? Most research so far has focused on single-user interactions, although voice interfaces also seem particularly amenable to multi-user scenarios. In the following sections we review research on conversational agents (section 2.1), natural language interfaces for software tools (section 2.2), studies comparing different modalities (section 2.3), and interfaces for supporting collaboration (section 2.4).

2.1 Conversational Agents: Voice Assistants and Chatbots

Conversational agents refer to software that users interact with via typed or spoken natural language. Various types have emerged including chatbots, virtual assistants, and voice assistants. There has been much hype about the potential of the various kinds of conversational agents for transforming user interactions (see [36]). More recently, voice assistants, such as Alexa or Siri, have become popular in everyday life, embedded in various smart devices (e.g., smartphones, smart speakers, wearables), designed to provide help, but so far, only have been designed to be reactive (i.e., responding to user requests).

Recent research into how these kinds of voice assistants are appropriated in domestic settings – where they have become popular in the form of smart speakers – have shown how they can mediate social situations, facilitating various kinds of social bonding and family interactions [7]. For example, Porcheron *et al.* [56] observed how interactions with an *Amazon Echo* in a family setting were seamlessly interwoven with other ongoing activities at family mealtimes where parents were at the same time trying to get their child to eat their food. They also point out how our conversations with each other and voice-assisted technologies interleave in nuanced ways, rather than being separate conversations within the family or between the family and the device, that switch smoothly from one to another. However, another study investigating how virtual assistants were used in multi-party conversations showed that sometimes interactions with Alexa in a social setting can be

awkward; disrupting the flow of normal human social interaction [59]. Family members sometimes needed to repeat and refine queries, which were not understood by the voice assistants; other times, they had to enforce silence so that the assistants could better understand their queries.

Conversational agents have also been developed for more specific applications, such as educational settings (e.g., [22,28,63,69]), to motivate and help students learn or to provide scaffolding. For example, the conversation tutoring system *AutoTutor* and its variations were found to produce significant learning gains [27]. The systems consist of an avatar (the "tutor") that speaks; a graphical interface related to the tasks and a chatbot-like interface that shows what the avatar has said and where the users can provide their input. Some of the key "moves" which the agent supports are: asking questions about the topic at hand, providing hints (until the learner provides a correct or acceptable answer), correcting students' answers and providing feedback. The AutoTutor systems have been mostly designed to support individual user's learning of topics like computer literacy or physics. The studies of it being used by students have shown how it can help them learn about a specific topic by motivating and guiding them [27].

Asking learners questions is key to spark curiosity and scaffold sensemaking [8,77]. Furthermore, enabling students to formulate their own questions can increase their learning performance [15,41,42]. Research has thus investigated how question-asking agents could support learners. For example, Alaimi *et al.* [1] investigated how different types of agents can encourage children to formulate questions and Ceha *et al.* [10] found that question-asking robots can be successful at enhancing students' curiosity about a topic. Questions asked included: *"I am curious. Do the holes form when gas bubbles get trapped when the lava cools? Do you have any idea?*"; The findings from their study also showed that curiosity can be "contagious" as the robot which verbally expressed curiosity was able to influence the participants' curiosity.

More controversially, Winkler *et al.* [83] found that a voice assistant-based tutor had a more positive effect on task outcome and collaboration among learners in a problem-solving task compared with a human tutor. The reason being is thought to be that users feel more confident and comfortable as the voice assistant does not judge them nor build up any pressure. There has also been increasing interest in using chatbots in different learning contexts, as part of online learning (e.g., [47]). These kinds of educational chatbots have been found to improve communication while simplifying learning interactions (e.g., [62]). Winkler et al. [82] showed that in the context of online lectures, a conversational agent, which scaffolded learners' understanding, had more positive effects on learning compared with an agent that did not. Similarly, Song *et al.* [71] developed a conversational agent that was successful at getting learners to think about their progress. Tegos *et al.* [76] found that conversational agents can trigger dialogs between students in online discussion by intervening in a conversation, which substantially improved both individual and group learning outcomes.

This research suggests that conversational agents are effective at stimulating reflection and curiosity as well as guiding users and providing scaffolds for different types of tasks, such as problem solving or learning – while also seeming to put less pressure on users than human tutors/facilitators. Voice assistants, in particular, have the additional potential of being able to be naturally embedded in multi-user interactions and conversations. Next, we investigate how natural language interfaces and conversational agents can be developed as part of software tools for specific tasks, such as data analytics.

2.2 Natural Language Interfaces for Software Tools

There has been much interest in combining natural language interfaces with software tools, such as data analytics and visualization tools [5,16,21,24,34,39,65–67,72–75,78]. One reason is that natural language input can make it easier for users to express data-related questions or queries, since they can use everyday language, which they are familiar with. A recent commercial tool is *Tableau's Ask Data* (see [78]), which enables users to formulate queries in natural language to generate and modify visualizations.

One of the earlier systems that allowed users to plot data using speech or text queries was *Articulate* by Sun *et al.* [75]. Their evaluation showed that when participants had to plot the same data in Microsoft Excel, they were significantly slower and found the steps required more complex and more confusing, despite the majority of them being familiar with Excel and its charting features. Another data analysis tool, *Ava*, that provided a chatbot interface, was designed to allow data scientists to assemble data analytics pipelines [38]. Computer scientists, who were knowledgeable about data science, were able to build machine learning models faster than when using Python. Another example is *Eviza* by Setlur *et al.* [66], which allowed users to interact with and modify visualizations of geospatial data via typed natural language queries. In a user study comparing *Eviza* with Tableau (without any natural language features), participants found *Eviza* to be more natural to use and completed the analysis tasks significantly faster, however, some users experienced a loss of empowerment and ownership.

Overall, the user studies and evaluations of these systems have found that they can improve task performance and that users are generally positive about using natural language interactions. However, they also highlight the challenge of having to deal with users' often underspecified and ambiguous natural language queries, which are often colloquial and have less semantic and syntactic rigor when compared to conventional query languages such as SQL [29]. To address this challenge, various approaches have been suggested. One that has been incorporated in different systems is to display additional interface elements, after an ambiguous query has been made, allowing the user to provide further specifications (e.g., [24,67,72]). An example for this is *DataTone* [24] by Gao et al., which allows the user to type or speak queries when exploring databases or spreadsheets; when a query is ambiguous, such as "Show medals for hockey" the system displays a widget where the user needs to select "Ice Hockey" or "Field Hockey". Another approach is to consider the current "interface context" (e.g., data points, which the user has selected before their query) to infer what the user refers to in their natural language query, which *Orko* by Srinivasan et al. supports [72].

In summary, this research suggests that natural language as an interaction modality, with or without an agent, can help both lay and experienced users perform data analysis tasks more efficiently and enable them to ask questions about the data in a more familiar way. Most of this body of research, however, has focused on how to speed up task completion by using (spoken) natural language or by combining it with other modalities (multimodal interactions). However, little is known as to whether speaking versus screen-based interactions when using such software tools makes a difference in how users engage with and make sense of the data. Furthermore, most of this research has focused on single-user scenarios and it is unclear how the findings would apply to multi-user scenarios. Compared to most previous research the focus of the present study differs in three ways, (i) we consider a multi-user scenario, (ii) the system incorporates an agent that prompts participants, and (iii) rather than focusing on efficiency our aim is to better understand the main differences in how participants progress with their discussion and interact with the agent-enabled system, depending on the interaction modality.

2.3 Using Different Modalities

Clark et al. [12] have summarized empirical research comparing speech versus graphical interfaces on user performance and experience, which has shown mixed results. In some studies, the use of voice was more beneficial than in others. Le Bigot et al. conducted two studies [43,44] investigating written text versus speech input with information retrieval systems, one of them for a restaurant search and the other for travel planning. In their first study [43], they found no difference in transfer effects when switching from one modality to the other. Subsequently, they found people were faster when working in written mode than in spoken mode, although the latter was considered to be easier. They also found spoken interaction led to more *collaboration* with the system – in terms of users matching their utterances to the system's utterances – while written interaction was more *efficient* – in terms turns being required to complete the tasks [44]. In another study, Begany *et al.* [6] investigated users' perceptions of spoken versus written text input for a search interface. Written input was preferred compared to voice input because it was easier to learn and to use. Limerick *et al.* [46] studied pressing keys versus using voice commands and found that speech leads to a diminished sense of agency in users – which is in this case defined as the experience of controlling one's own actions and their outcomes. There have also been modality comparisons of the AutoTutor system; D'Mello *et al.* [17] found no difference in learning outcomes if students made system input via keyboard or speech. Similarly, Litman *et al.* [48] found no difference in students' learning gains for spoken versus typed modalities using a computer tutoring system.

These mixed findings suggest that whether speaking or typing is more effective depends on the context and task. However, most of the research on using different modalities has focused on how task completion performance varies or on exploring users' perceptions of using each modality. There has been less concern with how they impact on behavioral aspects, such as reflection, sensemaking or collaboration with others. One study that investigated the use of speech input on learner reflection found young children engaged in more thinking aloud when the interface supported speech input [51]. Gonzalez and Gordon [25] compared how using speech versus text as input affected the player experience and user understanding of their fictional role in an interactive narrative (which has similarities with playing a game). They found in the text condition, participants were more likely to adopt the role of the narrator (speaking in past tense), and in the voice condition more likely to speak directly to the narrator (or the computer), saying what was happening or what should happen (speaking in present tense). This suggests that speaking at the interface can enable users to step more into character, see themselves more as part of the story and in doing so change how they feel, think and experience. However, the research on how using speech can engender different user experiences, as opposed to improving task performance, is limited.

Hence, while the use of keyboard and GUI to type or select commands often seem preferable for many tasks because of their ease of use as well as increased agency and efficiency, voice user interfaces may prove to be more "natural", immediate or engaging, enabling users to "collaborate" more with the system and to have a different and possibly more interactive experience. In particular, users may also be able to draw upon familiar conversational practices and social norms when speaking with an agent in a group setting, such as a meeting or a classroom, where people are already talking with each other, possibly enabling the agent to naturally be perceived as a facilitator. Next, we consider research that has explored how technology can support human-human conversations and collaboration.

2.4 Supporting Conversation and Collaboration

One of the benefits of having voice assistants playing a role in a group setting is their potential for supporting human collaboration – through all being able to speak and listen to the agent at the same time – as the family studies with Alexa attested to. In particular, employing a voice assistant offers much potential for supporting mediation, prompting and facilitation in a collaborative setting. To understand which role voice assistants might play needs an understanding of how collaboration takes place between humans, and how other technologies have been designed to enhance this.

Collaboration usually involves two or more people working together to carry out a task, typically being collocated and engaging in conversation with each other, mostly in the form of verbal and non-verbal communication. When talking, each person usually takes turns, with each turn consisting of one or more *turn-constructional unit*s, which is a "complete" utterance that possibly leads to a *transition relevance place* [64] where another speaker may take a turn. For example, such *transition relevance places* can be found after questions, which often indicate an "invitation" to another speaker to respond or to take a turn. Thus, questions are an important "motor" of conversational interaction and turn-taking. One way to promote collaboration, is to design a system that supports turn-taking in a group setting. In particular, the question here is, if it is possible to design a system that actively promotes turn-taking (e.g., by question prompts) which in turn encourages more collaboration, for example, enabling ideas to be generated, a problem to be solved or to learn new forms of cooperative play [85]. In many scenarios increasing turn-taking in a conversation is desirable, as it is often related to a higher degree of interactivity in the conversation. This is also the case in the collaborative data analytics scenario – which we investigate here – where ideas and hypotheses need to be generated. The amount of turn-taking or speaker alternation rate per minute has been used as a proxy measure for interactivity of a conversation in previous work [18,19,30].

Yuill and Rogers [84] discuss a variety of methods and design considerations that can be used to promote collaboration, including constraining multi-user interactions through the design of the software. Marshall *et al.* [50] demonstrated that constraining the number of users at a shared interface that can interact at the same time can lead to more turn-taking, collaboration and articulation of ideas on solving problems. Voice interfaces may, likewise, be designed to encourage turn-taking behavior, where an agent asks a user a question and vice-versa. Hence, similar to how the voice modality can be more immediate in the way it encourages people to step into an interactive narrative role [25], a voice agent may have a direct impact on how turn-taking takes place in a group: A voice agent can take an active turn in the ongoing conversation, which a text-based agent displaying prompts on a screen cannot. However, little is known about the effect of this kind of directness and immediacy at the interface. On the one hand, they could facilitate and encourage new forms of collaboration, by prompting or encouraging users to take turns. On the other, they could interrupt the flow of an ongoing conversation (e.g., [52,57]) in ways that using a screen-based interface does not. Our research is concerned with exploring what the effects are on human collaboration when groups interact with an agent, using screen-based versus voice interactions, when engaged in an exploratory sensemaking task using a software tool. Can voice interactions with the system also result in an increase in turn-taking between users compared to screen-based interaction? Here, the challenge is deciding how many turns (of question asking) the agent should take in order to promote more turn-taking among the human group members.

2.5 Summary

Voice assistants and chatbots have been found to be effective at stimulating users' reflection as well as guiding and supporting them in different types of tasks, such as collaborative learning or problem-solving tasks (section 2.1). Another line of research has shown (so far mostly for single-user scenarios) that combining natural language interfaces with software tools can improve the efficiency and experience when completing specific tasks using the tools, such as making it easier for users to express queries in data analytics tools (section 2.2). For the design of conversational agents, an important question is, how the experience and efficiency are affected by the modality of human-agent communication, which has been investigated in different single-user contexts for different types of tasks with mixed findings (section 2.3). There is a paucity of research that addresses this question in situations of collaborative action, where an agent "takes part in" human-human conversations. In such collaborative settings, previous research has mainly investigated how the interfaces of collaborative systems can be designed to promote turn-taking, without the use of conversational agents (section 2.4). However, there seems to be much potential in using conversational agents to this end, due to their capability of actively prompting users by intervening in the ongoing conversation and asking them questions.

Building upon and bringing these strands of research together, the present work addresses the following questions: (i) whether a system containing a question-asking agent can support and facilitate sensemaking, ideation and exploratory thinking in a collaborative analysis task, and (ii) whether the modality of interaction, and more specifically, the modality through which the agent's prompts are provided to users – voice or screen-based – makes a difference to the users' question-asking and turn-taking behavior as well their engagement with the system and the task.

3 AIMS AND HYPOTHESES

The aim of our study is to investigate the effects of voice versus screen-based modality on interactions between the users and the system and between users themselves, in particular in terms of the turn-taking and question-asking that took place, when using an agent-enabled software tool for a sensemaking task. The collaborative task involved exploring a set of data visualizations, which were presented on a large display in front of the participants.

The two conditions compared were: (i) using voice requests alongside an agent that prompts participants by speaking, and (ii) using a screen-based menu to make requests (providing familiar GUI elements, such as check boxes, buttons etc.) alongside an agent that prompts participants through text messages (being displayed in a "chatbot-like" way). In other words, the two conditions differ in how participants request visualizations and how they receive agent prompts (see Figure 1): in the first condition, both system input and output are based on voice, while in the second condition they are both screen-based. Thus, the second condition (the "screen condition") mimics more common GUI-based systems. The way data visualizations were presented was the same in both conditions (also see Figure 1).

Of course, there are other possible permutations of modality combinations, such as voice input + screen output and voice output + screen input (even the task modality could be varied, for example by choosing an auditory instead of a visual one). However, we considered the two combinations that we selected (where both input and output are in the same modality) to be most important to initially explore, since we hypothesized them to be the most meaningful and natural combinations in most situations. Furthermore, we decided to use a visual task (i.e., visual data analysis) as most computer-supported work is mainly visual.

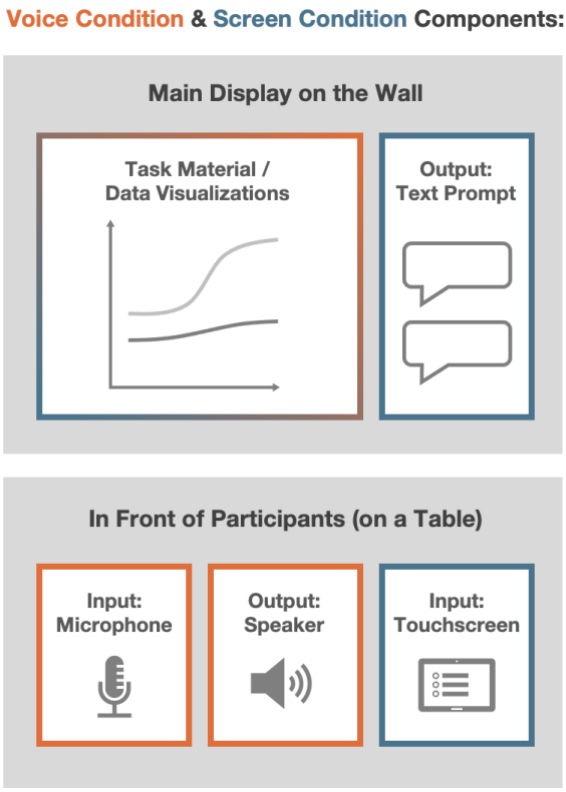

Figure 1. An overview of the modalities used for user input and system/agent output in the two conditions; the task material presented in the same way in both conditions.

We propose that making voice requests integrates better with the participants' ongoing conversation and is more fluid than switching to making requests through the graphical user interface. It is thus expected that voice interaction leads to users interacting more with the interface and exploring more of the data which results in our first hypothesis:

H1 *– human-computer interactions: The voice condition will encourage (a) more interactions with the software tool and (b) more of the available data visualizations being looked at.*

The metrics of (a) how much the interface is interacted with and (b) to what extent the task material (the available visualizations) is explored are proxy metrics for the users' engagement with both the interface and the task.

Regarding the impact of the modalities on *human-human interactions*, we further expect there to be more turn-taking in the voice condition between participants, on the one hand, since previous research showed voice can be more "activating" (see Mavrikis *et al.* [51]) and on the other hand, since the agent proactively speaks and "takes turns" in the conversation itself, which may stimulate/motivate participants to also take more turns.

We also predict that when a speaking agent prompts users with questions, it is likely to spark the users' curiosity more compared with reading its prompts from a screen (see Ceha *et al.* [10] and Gonzalez and Gordon [25]), which leads to our second hypothesis:

***H2** – human-human interactions: The voice condition will encourage (a) more turn-taking and (b) question-asking between participants.*

The amount of turn-taking (or the speaker alternation rate) in a conversation – apart from reflecting the *"rapidity"* of a conversation – has been proposed as a proxy measure for the *interactivity* of a conversation [18,19,30], representing a relevant metric in the present collaborative sensemaking scenario. The reason why number of questions (related to the task/the data visualizations) participants ask each other is used as a metric is because question-asking is key in an exploratory analytical task in which hypotheses should be generated [29,40], and often *"finding the right questions is often more important than finding the answer"* (Tukey [79]). More generally, question asking is fundamental for scientific thinking and intellectual exploration [11,80,86]. The number of questions thus also serves as a proxy for the extent to which participants' try to make sense of the data and for how exploratory and how curious they are [1,8,10].

4 USER STUDY DESIGN

We designed our study so that pairs of participants could take part together. This enabled us to analyze the conversations and turn-taking that took place between the participants themselves, and also between the participants and the agent. Previous research on "pair analytics" by Arias-Hernandez *et al.* [3] has shown how this approach also offers a natural way of making explicit and capturing reasoning processes (in contrast to single-user scenarios) while also enabling a variety of metrics to be used to assess collaboration.

A Wizard of Oz [68] paradigm was used to test the two hypotheses. This set-up allows us to both simulate and control the agent interventions for both conditions. Wizard of Oz studies have frequently been used to simulate and test novel systems with users, in particular, "intelligent interfaces" (e.g., Dahlbäck *et al.* [13] or Porcheron et al. [55]). Our agent, which we called Vizzy, was simulated by a human experimenter, who was tasked with triggering the prompts at certain times during the study.

In the voice condition, participants were asked to change the visualizations through voice requests. In the screen condition, they could make the same requests (i.e., through selecting variables/filters) from a menu shown on a tablet. Vizzy was designed to provide prompts in the form of questions in both conditions, either through synthesized speech (see Figure 2) or through chatbot-like text messages that appeared on the main screen next to the area where visualizations were displayed (see Figure 3). Thus, both conditions were based on examining data visualizations on a large screen; the difference between conditions was just in how requests were given to the system and how the system provided prompts (also see Figure 1), which were both either voice-based (voice requests & synthesized speech prompts) or screen-based (requests based on menu selection & text message prompts). The same set of visualizations were available in both conditions. To control for equivalence across the two conditions, Vizzy's prompts spoken aloud in the voice condition or displayed as text messages in the screen condition were selected from the same set of prompts.

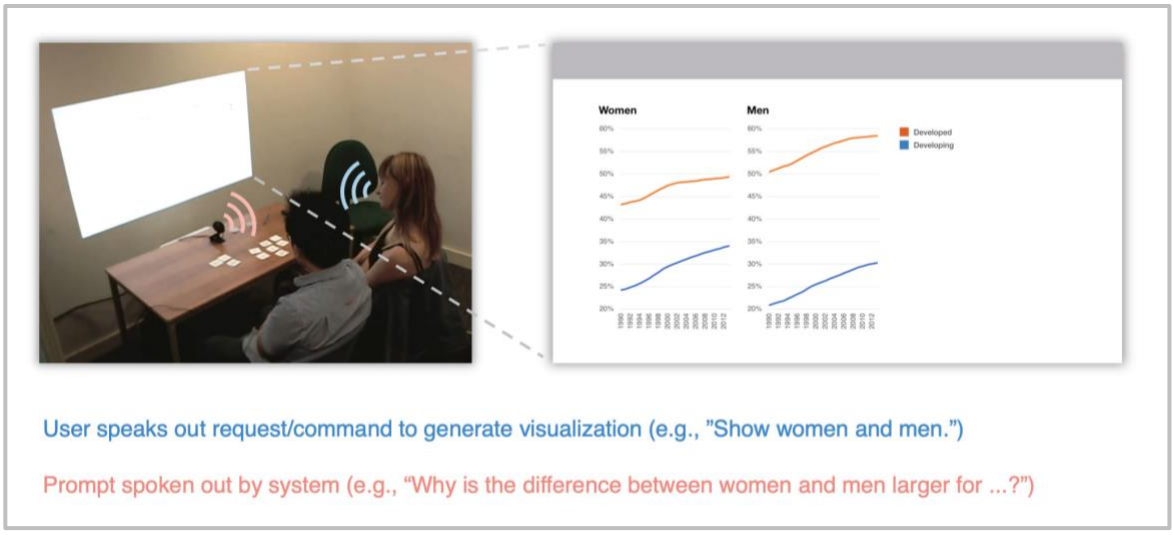

Figure 2. The interfaces for the voice condition with a microphone positioned in the middle of the table. The cards on the table show the set of available visualizations that could be generated.

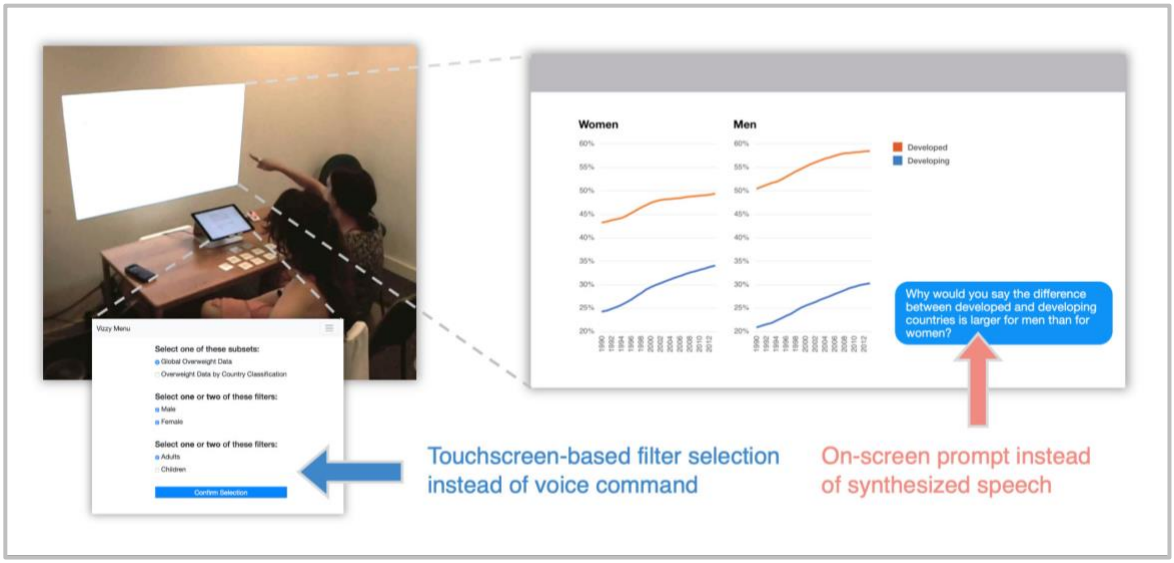

Figure 3. The interfaces for the screen condition with a tablet situated in the middle of the table for user input. The cards on the table show the set of available visualizations that could be generated.

The task. In an initial pilot study, we began by testing well-defined tasks with specific outcomes, by asking participants to find specific patterns in the data. However, participants' conversations were short; they focused on searching for the patterns they were asked to discover. We chose instead for our main study to design a more open-ended task that would require looking for differences and trends in visualizations and inferring the

possible reasons for them. Thus, the sensemaking task we designed involved interpreting visualizations for a given dataset to infer what might be behind the trends depicted over various time series – which reflects common activities of exploratory (time series) data analysis. The aim was to enable participants to try to make sense of a set of visualizations by hypothesizing about and questioning the underlying data, without the need to have a data analytics background. The domain chosen was health, in particular the prevalence of obesity throughout time, a topic that participants would have some understanding and familiarity with. Specifically, the data represented how the prevalence of obesity has increased in recent decades for different populations. The data used for the visualizations was derived from Marinez [49]. It is publicly available and comes from the Global Burden of Disease Study 2013 [54]. We chose a dataset that covers 24 years from 1990-2013. This period was considered sufficiently historical to enable participants to discuss about past developments that could have led to the trends and patterns in the graphs.

The visualizations could be generated from the obesity data by combining time series graphs by age (children/adults), gender (male/female) and "country type" (developed/developing and global). 22 visualizations could be generated, which showed the graphs for adults, children, adults *and* children, men, women, boys, girls, men *and* women, men *and* boys, women *and* girls, boys *and* girls, which could all be displayed as global average or split up into developed and developing countries; see for example Figure 4 which shows the averages of developed and developing countries for boys and girls. In the voice condition this visualization would be generated by saying ("*Vizzy, show boys and girls, developed and developing countries."* or *"Vizzy, show developed and developing for boys and girls."* or similar) and in the screen condition by using the menu on the tablet. The set of available time series to display meant that visualizations which could be generated were simple enough to understand but also sufficiently complex to show interactions and trends. They comprised a range of level differences as well as different types of growth that could be explored at a general level (e.g., the overall increase in the time series for boys and girls) or a more detailed level (e.g., how the speed of growth/growth rates changed throughout time) as can be seen in Figure 4. A larger set of variables and visualizations (based on the chosen dataset [54]) were tested in a pilot study. Since we found that the number of available variables was too large for sufficient experimental control, we decided to limit the set of visualizations to 22. In particular, we wanted to make sure that there would be sufficient overlap in the data the pairs look at together to reduce possible confounds. In a subsequent pilot study, the chosen set of 22 available visualizations proved to be sufficient to allow participants to explore and discover different aspects. The possible interactions between the different time series made the problem-space sufficiently complex while not being too overwhelming.

Agent prompts. In total 19 prompts were composed for Vizzy consisting of open-ended and more well-defined questions, many of which were applicable to more than one of the available visualizations. Examples of the prompts include:

- If I would say one of them is slowing down in recent years, which one would you say it is?
- Is the increase of one more significant than the other?
- What might have caused the sudden spike?
- So, if you look at this, would you say that the increase is slowing down in the number of overweight people for all four groups?
- Why would you say the difference between developed and developing countries is larger for men than for women?

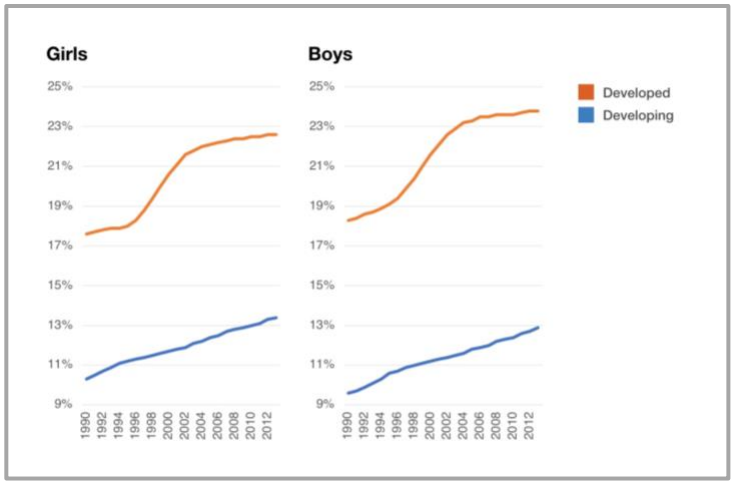

Figure 4. An example visualization of two graphs generated from the task dataset showing the change in prevalence of obese girls and boys for developing and developed countries from 1990 to 2013.

4.1 Participants

A between-subjects design was used with two conditions: *screen* versus *voice*. 36 participants took part in the study; 9 pairs for each condition. The pairs were randomly assigned to either the voice or screen condition. Instead of matching them up as stranger pairs, we asked the 18 participants we initially recruited to bring someone they knew and who they felt comfortable doing a collaborative task with. This enabled the pairs to feel at ease collaborating with each other during the study. Participants were recruited from our university campus and were between 18 and 35 years of age. 18 were female (10 in voice, 8 in screen). In 8 pairs, genders were mixed (4 in voice, 4 in screen). All participants were fluent in English with normal or corrected-to normal vision and hearing capability.

4.2 Experimental Set-Up

Physical Room Set-up. The visualizations were projected onto a screen on a wall. A desk and two chairs were positioned in front of it (also see Figure 2 and Figure 3). This enabled each pair to be able to readily see the projected visualizations while also being able to face and speak to each other. On the desk was placed a set of small cardboard cards (5x5cm) indicating the labels or “title” of each visualization that could be created (e.g., “boys and girls, developed and developing countries”). The cards could be used by participants in both conditions to help them keep track of what they had already looked at or what to look at next (also see Figure 2 where participants grouped the cards). In the voice condition the cards could also be used by participants to help them formulate their requests to Vizzy to show a visualization (e.g., “Vizzy, show [card label]”). In the screen condition, a touchscreen display was placed on the desk showing a menu to generate the visualizations. In the voice condition, two loudspeakers were positioned behind the desk that Vizzy could be heard through. The voice used for speech output of Vizzy was based on Amazon Polly’s [90] voice *Joanna*.

Two webcams were installed in the room to record the participants: one facing down from the ceiling the other one located behind. The former allowed to capture interactions with the tablet (in the screen condition) and cardboard cards, the latter allowed to capture participants along with the main screen to see what they point

and look at. Each participant was also asked to wear a lavalier microphone to record what they said. The video of the participants and the conversations were recorded using *OBS Studio* [91].

Wizard of Oz Set-up. The visualizations and agent's prompts were controlled through a dedicated computer in an adjacent room which was connected to the projector in the experimental room. During the study, the Wizard (a second experimenter) listened to the audio stream and observed the video feeds from the two webcams. The Wizard controlled the interface for the two conditions to present the requested visualizations on the screen and select the Vizzy prompts to be played (voice condition) or displayed on the screen next to the visualization (screen condition).

Vizzy Interface: User Control and System Prompting. In the voice condition, a conference microphone was positioned in the center of the desk (described to participants to be the microphone through which Vizzy listens to their requests). In the screen condition, the microphone was replaced with a tablet showing the GUI from which they were to select their choices. In order to produce consistent recordings between the conditions, the conference microphone was also present in this condition but hidden under the tablet, not visible for participants.

Vizzy was designed to occasionally prompt the participants, intended to encourage them to explore further what was causing the trends and the rise in different obesity levels in the displayed visualization. For each visualization prompts could be selected by the experimenter from a predefined set which were applicable to the specific visualization (see Figure 5). Each visualization was assigned between 1-3 applicable prompts; some prompts were applicable to more than one visualization. For example, Vizzy (or rather the wizard) could select the prompt *"Would you say that the increase is slowing down for all four groups?"* for all visualizations where there were four time series appearing (e.g., "adults and children" of "developed and developing countries"). Prompts were only provided if (i) participants had not yet discussed about the pattern/trend/difference the specific prompt referred to and if then (ii) there was a silence of approximately 3 seconds or more in their conversation to avoid interruptions of the pair's discussion. This threshold was set based on our iterative design process of the system, where after testing different durations, approximately 3 seconds was considered most appropriate. The reasons for this were two-fold, it was (a) long enough that in most cases there was no direct interference with an ongoing conversational turn by one of the participants and (b) short enough that there were still sufficient opportunities for the agent to intervene. However, both conditions mentioned above (i & ii) had to be met for a prompt to be provided, a silence itself did not lead to a prompt to be triggered. Hence, which prompts could be provided for which visualization depended on what pairs discussed until there was the first silence in their conversation about a specific visualization. Furthermore, we aimed at providing prompts only about every two minutes on average as we found in our design process that when the intervals between the prompts are too short, they can become annoying and disruptive to the flow of the discussion. The frequency of agent prompts was kept as similar as possible in both conditions. In the screen condition there was a "clicking" notification sound (similar to sounds used in messaging apps) so that participants would not miss a prompt. To ensure consistency across the two conditions, the Wizard spent considerable time familiarizing themselves with the set of prompts, practicing selecting different ones for the different stages of the task and the types of visualizations being looked at before commencing the study, following the above rules and guidelines.

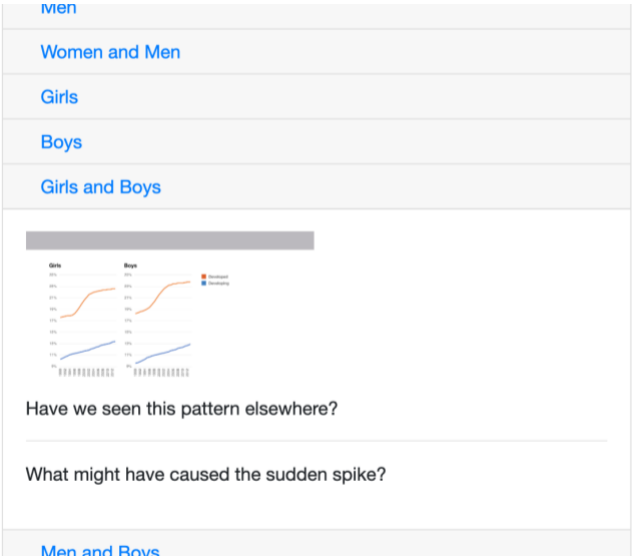

Figure 5. Screenshot of a part of the Wizard of Oz interface showing two of the prompts, which could be selected for one of the available visualizations ("girls and boys, developed and developing countries").

4.3 Procedure

Ethics approval was obtained from our university prior to the study. Pairs of participants were informed about the purpose of the study and asked to fill in a consent form agreeing to being audio and video recorded during the study for subsequent analysis.

The participants were informed that they would be asked to collaborate in an exploratory data analysis task. They were told that there was no right or wrong way to do the task and that they should just try to reflect on and deduce what was causing the trends in the obesity levels over the time period of 1990-2013. They were further instructed that they could press a button on a remote control on the table when they had completed the task or if they had a problem during the task. To begin, the experimenter explained to the participants how to ask Vizzy to generate the visualizations. They were informed that Vizzy would prompt them at certain times during the study. Participants were told that they could decide for themselves if they wanted to respond to Vizzy's prompts and to use them as prompts to guide their thinking. They were then given a set of cards that showed the possible visualizations they could generate. They were told that cards have the purpose of giving them an overview of the available visualizations that could be requested. In the voice condition they were also shown how they could request a visualization (wake word + visualization request) and how they could use the cards to help them formulate their requests (i.e., wake word + label of the card). Finally, they were informed that the task would normally take about 15-20 minutes. The instructions were identical for both conditions except from when describing how to interact with the voice and screen-based interfaces. After the introduction, the experimenter left the room, and the participants commenced the task. After completing the task, the experimenter returned to the room and conducted a semi-structured interview with the pair asking them to reflect on their experiences during the study. Participants were then debriefed about the aim of the study and that it was a Wizard of Oz design. Then, the Wizard, who was controlling Vizzy, came into the experimental room and introduced himself. The participants were each compensated with a £15 Amazon voucher.

4.4 Data Analysis

The video and audio data collected were analyzed using a combination of *automated* speaker diarization and transcription as well as *manual* transcription and analysis of conversations, questions asked and turns taken, providing the basis for the subsequent quantitative and qualitative analyses.

4.4.1 Quantitative Analysis: Interactions Between Participants and the System

To test our H1 regarding the human-computer interactions *(The voice condition will encourage (a) more interactions with the software tool and (b) more of the available data visualizations being looked at.)* the corresponding user interactions were broken down into *(a) visualization requests,* measuring the total number of requests that were made for visualizations, *(b) visualizations explored*, which was determined by how many *unique* visualizations (of the set of available visualizations) the participants looked at together. For example, out of four possible visualizations (A, B, C, D) if three are looked at (e.g., A, B, C), it would provide a measure of ¾ or 75%.

When considering the following sample sequence of requesting visualizations (A, B, A, B, C, A), this would result in *visualization requests* = 6 but *visualizations explored* = 75%, since only 3 out of the 4 available visualizations were looked at (i.e., A, B, C).

In the present scenario, *visualizations explored* was chosen as a metric, since it was considered a good proxy for how extensively the participants examined and discussed the set of available visualizations. The reason for this is that as part of the task, participants were asked to think about and discuss each visualization they requested, which participants also did in most cases. The aim of both measures, therefore, was to capture how much users interact with the system and the task material.

4.4.2 Quantitative Analysis: Interactions Between the Participants

To test our H2 regarding the human-human interactions *(The voice condition will encourage (a) more turn-taking and (b) question-asking between participants.)* we measured (a) the number of speaker changes made during a conversation and (b) the number of questions participants asked each other. Turn-taking was approximated through quantifying the number of speaker changes (see [18,30]). As these metrics intend to capture participants' behavior to compare it between conditions, turns by Vizzy (when it provided a prompt) were not counted towards these metrics.

4.4.3 Qualitative Analysis: Patterns of Collaboration and Sensemaking

To investigate how the conversations after Vizzy's interventions unfolded, a randomly selected set of segments was transcribed. From these, we examined in more detail the content of the conversations and the extent of turn-taking and how initial ideas and inferences about the data visualizations were followed up. Excerpts are provided to illustrate the patterns of discussions and interactions that took place in section 5.3. In addition, interviews were conducted at the end of the experiment to find out more about the participants' experience of using the system. An analysis of the interviews is provided in section 5.4.

4.4.4 Data Analysis Methods

To achieve sufficiently accurate identification of the active speaker with the given set-up we developed our own speaker diarization model, which took the three audio streams (left speaker microphone, center microphone, right speaker microphone) and compared the intensity values using *Parselmouth* [37]. Based on manually diarized recordings, threshold values for each microphone were defined using *Sequential Model-based Algorithm Configuration (SMAC)* [35]. Based on the thresholds, absolute differences between the microphones

were defined to identify the active speaker, which was used to identify *speaker changes*. Each audio segment of the speaker diarization model was then transcribed using *Azure Cognitive Speech Services*. While the accuracy was not perfect, it was sufficient to then also quantify the *number of words spoken* for each participant.

Vizzy's utterances, along with the visualizations that were requested by the participants, were automatically tracked using a log file, from which the timestamps were extracted. Timestamps were manually recorded after each prompt by Vizzy from the beginning to the end of the participants' discussion about that prompt to record the discussion lengths (if a prompt was ignored by participants, the discussion duration was set to 0).

The number of *questions* asked by each participant pair as part of their discussion were not based on the automatically generated transcription files but manually tagged and coded, since their questions were not always identified with the required accuracy by *Azure Cognitive Speech Services*. To reduce confounds we aimed to exclude the questions that were not related to the data analysis task in our quantitative analyses, for example questions about Vizzy and its capabilities (e.g., "Do you think Vizzy can do this?"). We did this as we hypothesized that the voice modality could lead to more questions about the interface capabilities compared to the screen modality (post-hoc analysis showed that this was indeed the case). Furthermore, in those cases where a participant requested a visualization by asking a question, which occasionally happened (e.g., *"Vizzy, could you show developed and developing countries?"* instead of "Vizzy, please show developed and developing countries."), this was also not considered for the *participants' questions* metric to avoid confounds.

The duration of the individual sessions varied across participant pairs when exploring the dataset. To control for this, averages per minute were calculated instead of totals for the metrics *requests* (for visualizations)*, speaker turns* and *questions* by participants.

5 FINDINGS

5.1 Main Quantitative Results

Overall, participants in both conditions looked at most of the available 22 visualizations. A main finding was that the participants in the voice condition interacted more with the system, explored and discussed more of the available visualizations and asked more questions about the visualizations. As our hypotheses were one-tailed, significance was tested using an alpha level of 0.05 for t-tests and U-tests[1].

We conducted (i) a U-test on the number of visualizations that were looked at and (ii) a t-test on the number of requests made per minute. Both null hypotheses were able to be rejected as the results were found to be significant. In support of H1 (U_{18} = 18, p = .017), the percentage of *visualizations explored* was higher in the voice condition (M = 96.97%, SD = 6.43%) compared with the screen condition (M = 89.39%, SD = 7.54%) and the difference in *requests* (per minute) to change the visualizations was also found to be significant ($t(16) = 2.75$, p = .007), where more requests were made in the voice condition (M = 1.08, SD = .15) than in the screen condition (M = .80, SD = .26). These two significant findings therefore support our first hypothesis (H1) that participants in the voice condition would interact more with the system and look at more of the available visualizations. Figure 6 shows the percentage of visualizations looked at for both conditions and Figure 7 the number of requests per minute for both conditions. See Table 1 below for an overview of all results.

[1] Based on an examination of box plots, no significant outliers were identified. The Shapiro-Wilk statistic indicated that the data was distributed normally ($p > .05$), which was confirmed by examination of histograms as well as skewness and kurtosis values; only for the metric *exploration* this was not the case, which is why a Mann-Whitney U test was conducted here instead of a t-test.

Table 1. Summary of findings: Inferential statistics are *t-tests*, and effect sizes (ES) are *Cohen's d* except *exploration*, which is based on a *Mann-Whitney U test* and *eta squared*; all *p* values are significant.

Hypothesis/Analysis Category	Metric	Condition	Mean	*SD*	Statistic	*p*	ES
Human-Computer Interactions (Hypothesis 1)	***exploration***	voice	96.97	6.43	18.00	**.017**	.27
		screen	89.39	7.54			
	requests	voice	1.08	.15	2.75	**.007**	-1.30
	(per minute)	screen	.80	.26			
Human-Human Interactions (Hypothesis 2)	***questions***	voice	.98	.22	3.51	**.002**	-1.65
	(per minute)	screen	.55	.29			
	turns taken	voice	6.39	1.11	2.10	**.026**	-0.99
	(per minute)	screen	5.52	.57			

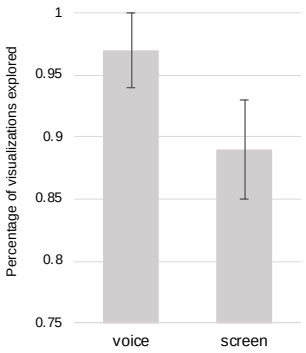

Figure 6. Percentage of the visualizations looked at for voice and screen conditions.

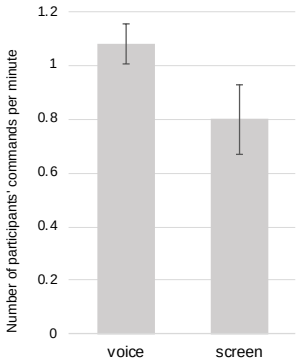

Figure 7. Requests made by participants per minute for voice and screen conditions.

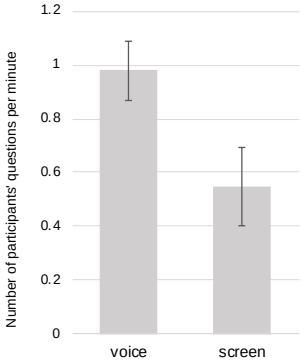

Figure 8. Questions asked by participants per minute for voice and screen conditions.

The t-test was also found to be significant for H2 ($t(16) = 3.51$, $p = .002$); participants asked more questions per minute in the voice condition ($M = .98$, $SD = .22$) than in the screen condition ($M = .55$, $SD = .29$) as can also be seen in Figure 8. Specifically, participants in the voice condition asked 78% (0.98/0.55) more questions than in the screen condition. We also found a significant difference between the two conditions for the number of changes of who spoke at any given time ($t(16) = 2.10$, $p = .026$) which we refer to as *turns taken*; this happened more often in the voice condition ($M = 6.39$, $SD = 1.11$) than in the screen condition ($M = 5.52$, $SD = .57$).

Post-hoc analyses revealed that there were no relevant differences between conditions regarding how many of the 19 available agent prompts were triggered per minute[2].

5.2 Quantitative Analysis of Types and Patterns of Reponses

In addition to testing our two hypotheses we examined the patterns of responses across the two conditions. In particular, we looked at the levels of participation between the pairs in the two conditions in terms of what each participant contributed. Overall, there was a tendency towards more equal participation in the voice condition. However, the differences were found to be not statistically significant. For this we used two measures:

(1) Interactions between each participant and Vizzy Analytics: We calculated the average deviation from *equal contribution* (i.e., that both participants would make 50% of requests). The deviation in percentage points was found to be smaller in the voice condition (*M* = 18.07, *SD* = 13.41) than in the screen condition (*M* = 22.28, *SD* = 15.42), suggesting that the pairs interacted with the system more equally in the voice condition.

(2) Interactions between participants: For *total words spoken*, the deviation in percentage points from *equal contribution* in the voice condition (*M* = 4.38, *SD* = 2.78) was also found to be smaller compared to the screen condition (M = 7.52, SD = 8.51). Similarly, for *total duration of speech*, the deviation from *equal contribution* in voice was (*M* = 5.36, *SD* = 5.37) slightly smaller than in the screen condition (*M* = 6.63, *SD* = 3.49). It is worth noting that there may be multiple factors related to the tendency towards more balanced interactions in the voice condition. For example, in the screen condition it often was the same participant, who read the prompt out loud, which may have somewhat affected the above metrics. However, these factors will not be considered in more detail here, since both (1) and (2) were not of primary interest in this study and rather an adjunct to our main analyses.

The duration of the discussions in response to Vizzy's prompts was found to be somewhat shorter in the voice condition (*M* = 35.10s, *SD* = 17.39) than in the screen condition (*M* = 47.67s, *SD* = 24.85). However, there was no significant difference. To examine further the patterns of conversation, we subsequently analyzed how the pairs responded to Vizzy's prompts in terms of (i) the percentage of prompts that were responded to versus ignored by the pairs in the two conditions, (ii) how long they took before responding, and (iii) the length of their conversations. These analyses were not conducted with the aim to discover specific differences between conditions but to better understand the general process/pattern of how participants interacted with the system and responded to its prompts regardless of the condition.

(i) Prompts responded to versus those ignored across the two conditions. Nearly all pairs responded to Vizzy's prompts in both conditions. Only 6.90% of its prompts were ignored. As expected, pairs in the voice condition ignored fewer prompts (*M* = 4.56%, *SD* = 9.91) than in the screen condition (*M* = 9.30%, *SD* = 11.39).

(ii) Time to respond to Vizzy. The average time taken by a pair to react/respond to a Vizzy prompt was roughly 4 seconds for both conditions (*M* = 3.85, *SD* = 1.70). As the screen-based prompts did not have to be processed by pairs immediately, we were interested in whether the time to respond was longer. It was found to be slightly longer in the screen condition (*M* = 4.19, *SD* = 1.69) than in the voice condition (*M* = 3.52, *SD* = 1.73).

[2] Although the metrics were defined so that they would not be affected if there are differences in how often Vizzy asked questions, we aimed to provide agent questions in a similar way in both conditions (see also section 4.5). To assure this was the case we did a manipulation check, which showed that the number of questions asked per minute were indeed very similar in the voice condition (*M* = .44, *SD* = .15) and text condition (*M* = .45, *SD* = .16).

(iii) Duration of responses after Vizzy prompts. We classified the discussions in response to Vizzy's prompts in terms of whether they were "short" or "long". Short response usually lasted 5-20 seconds and made up 20.90% of the total conversations – in most of these cases pairs just agreed on an answer without discussing it further. However, there were far more longer responses across both conditions, comprising 72.20% (*SD* = 19.01). They generally lasted between 21-90 seconds, and in a few cases even more. The long responses usually consisted of pairs talking about the patterns they saw being depicted in the graph data, followed by hypothesizing about the possible reasons for this. There was only a small difference in the long responses across the two conditions with an average of 74.46%, (*SD* = 18.99) in the voice condition and of 69.95%, (*SD* = 19.90) in the screen condition.

5.3 Qualitative Analysis of the Participants' Discussions Following Vizzy's Interventions

The quantitative findings have shown significant differences between the two conditions in terms of the number and kinds of interactions among participants themselves and with the agent. In both conditions Vizzy's prompts acted as facilitators for participants' conversations, triggering them to talk about the possible reasons behind the changes in the obesity data for the different demographics. Here, we are interested in examining further the types of sensemaking that occurred following Vizzy's prompts and to see if there were any differences, in terms of what the pair said and did next. To do this, we carried out an in-depth conversation analysis of 4 conversation segments before and after one of Vizzy's prompts. We present here two randomly selected prompts by VIzzy for which we transcribed the participants' conversations before and after Vizzy asked each prompt in both conditions (voice and screen). Each of the transcribed segments were then analyzed with respect to the patterns of how the conversation was structured and what was spoken about.

To examine the interactions and sensemaking that took place we used an adapted form of conversational analysis that focused on the turn-taking between the participants following an intervention from Vizzy. We draw from Porcheron ***et al.'s*** method [56], who used it to describe in detail the various methods families use to organize their talk with and around their smart speaker in their everyday conversations. The method has become an accepted method in HCI, where a small number of segments are chosen to illustrate aspects of everyday conversations and social conduct. When transcribing the segments we follow some of the standard transcription conventions [4,33,57] in conversation analysis used in HCI. For reference, we indicate where pauses take place (e.g., (1.7) for 1.7 seconds), where an utterance is <faster> than usual, or where it is elong:::ated, where talk is LOUD or °quiet°. Empty parentheses ( ) are used where spoken words could not be recognized. Where speech overlaps indentation and [square brackets] are used and ((unspoken actions)) are given in double parentheses, which can be either actions of speakers or the system. Speakers are indicated by P1 and P2, the synthesized speech produced by Vizzy is identified by the label "VZ".

The segments start slightly before Vizzy's prompt and end after the participants either request another visualization, start discussing about another topic or agree/conclude on their answer. After transcribing and analyzing four segments, the recordings of all 18 pairs were listened to several times in full length by two of the researchers and analyzed for patterns in participants' behaviors – such as how Vizzy's prompts were responded to in both conditions. After that, the identified patterns were discussed between the two researchers. The insights from this preliminary analysis were used to describe if the behaviors that were observed in the four segments below were found to be typical for the respective condition. Overall, for both conditions, the pairs seemed to be at ease with each other, taking turns and engaging in a level of banter. Furthermore, participants

seemed to quickly get used to the system not following up on their questions; they took on board the agent prompts and included them in their conversations, usually discussed them until they came to a conclusion or felt they have sufficiently discussed about the prompt and then moved on to another visualization.

With respect to the two conditions, it was found that the same prompt in voice or screen condition elicited similar levels and types of sensemaking. However, there were also specific patterns in each condition regarding how the conversations and interactions took place. In particular, the segments reveal that a typical conversational pattern for the voice condition was for a pair to start or continue to discuss relatively quickly after an agent prompt by "bouncing off" ideas of each other, asking each other questions, generating hypotheses about possible reasons behind the patterns in the graphs and then moving on to another topic and/or visualization. The typical pattern in the screen condition was that one or both participants initially read the prompt out loud when it appeared on the screen, which was then followed by a discussion similar to those which took place in the voice condition. However, the conversation often resumed and progressed with a "slower pace" and there seemed to be less "thinking aloud" or bouncing off ideas.

5.3.1 Responses to the Prompt "What Might Have Caused the Sudden Spike?"

First, we present examples of the conversations in the voice and screen condition following the prompt *"What might have caused the sudden spike?"* asked by Vizzy when the visualization "girls and boys, developed and developing countries" was being displayed (see Figure 9). As the visualization shows, there is a marked increase ("spike") in developed countries between 1996 and 2002 which the prompt refers to. This open-ended prompt aims to trigger the participants to look at the data and generate their own hypotheses when responding.

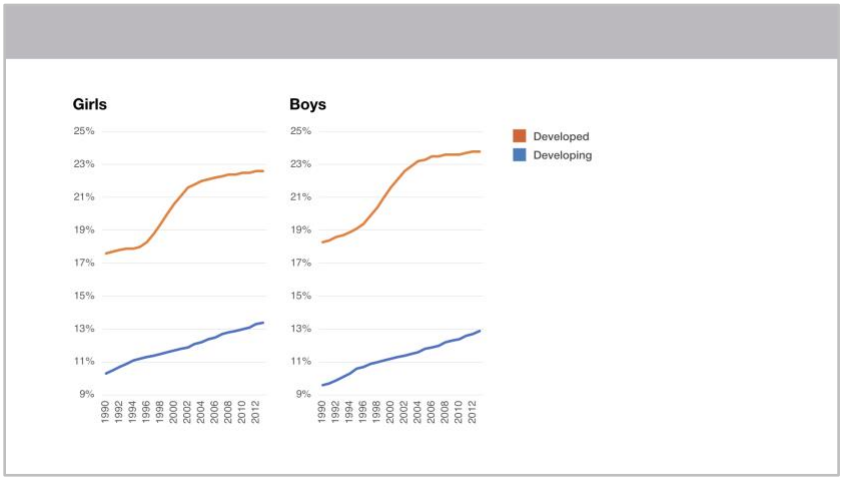

Figure 9. Visualization "girls and boys, developed and developing countries".

The first segment (1-VOI) illustrates that there is a high number of speaker changes comprising a quite rapid back and forth of suggestions between the participants in the voice condition. They seem to iteratively construct their hypotheses about possible reasons, as to what may have caused the sharp increase in the time series graphs they are looking at, by building upon (or contrasting) what the other person says. Often, they seem to just "think out loud" while generating ideas (e.g., line 1-2, 16, 19-23).

01	P2	It's the same pattern ((points at the visualization and traces the line in
02		the air)) as the global time series.
03	P1	Ah, you mean that ((points at visualization)) [the orange line] is the
04		same as…
05	P2	[Yeah.]
06	P1	Yeah. The global one?
07	P2	… The global increase in uhh…
08	P1	Yeah.
09	P2	… in the overweight.
10	P1	Yeah.
11	P2	But here it's an...
12		(3.6)
13	VZ	**What might have caused the sudden spike?**
14	P2	°Sudden spike°, ah…
15	P1	It is around the 2000s, shortly before 2000.
16	P2	I don't know, like (0.3) globalization?
17		(2.2)
18	P1	I have no idea.
19	P2	And possibility of getting a lot of different foods (2.4) or could also
20		be...
21	P1	Oh, I think, obviously, electro::nics – compu::ters, PlayStations.
22	P2	Ah, yeah!
23	P1	So, children play less outside and get fat.
24	P2	Yeah (0.6) yeah, true.
25		(2.2)
26	P1	Uhm, but it’s actually quite surprising. That's an (0.8) 5% increase
27		(0.4) °almost°.
28	P2	Yeah, even a bit more.
29	P1	Yeah, °or a little bit less°. (0.4) CRAZY!
30	P2	People stay less outside and play inside.
31	P1	And look, the other one is just linear.
32	P2	Yeah, this is developing, yeah.
33	P1	This is a cra:azy increase. We should maybe look if it’s the same, (1.6)
34		uhmm…
35	P2	Men, like, uhhm ((points at one of the cardboard cards on the table))
36	P1	Yeah, Vizzy, show (1.3) men and boys.
37		(2.3)
38		((Vizzy shows requested visualization and participants continue discussing
39		the newly opened visualization.))

Segment 1-VOI: Pair 3, Visualization on display: “girls and boys, developed and developing countries”.

In Segment 1-VOI presented above, P1 begins by describing the part of the graph which Vizzy’s prompt is related to and then the two participants alternate between hypothesizing about the reasons as to why this happened (also see Figure 9) and which historical events could be related to it. Before Vizzy triggered the prompt, the pair were involved in a discussion about how the pattern they are seeing compares to patterns in other visualizations they have previously explored. After a silence of more than three seconds, Vizzy then triggers its prompt. When it plays via the speaker, it appears to scaffold participants’ thinking around when and why there was a spike in the developed countries (see orange line in Figure 9). Vizzy does not contribute any

further to the conversation. Instead, the participants engage in a discussion about possible reasons for the increase. As mentioned previously, it can also be seen here that participants quickly got used to Vizzy not following up after asking a question, and thus usually just continued their conversation.

After Vizzy's prompt, P2 makes two suggestions for possible answers (on line 16 and 19), which is then followed by P1 also making a suggestion, namely that new technologies were the main reason (line 21), which P2 then agrees to (line 22 and 24). The segment finishes with participants taking a closer look at the increase and an attempt to quantify it in a percentage increase (line 26-29). After that, on line 30, P2 provides a "conclusion". This leads to a new train of thought; to check if this increase can also be found in the graphs for *men and boys*. At which point they then request Vizzy to take a look at the visualization with the corresponding time series shown side by side. Hence, the role the pair see Vizzy playing is essentially "someone", who follows their requests for showing new visualizations on the display, and who will occasionally prompt them with a question.

The participants sometimes ask questions (e.g., see line 3), which could be directed to Vizzy or each other. As they have learnt to understand when Vizzy will intervene and what Vizzy will say, their question-asking is more of a way of thinking aloud to clarify what they are looking at in the data or to ascertain what the other participant meant when they said something.

In this segment it is worth noting that P2 may have continued their thought after having paused mid-sentence ("But here it's an..."). It is possible that the silence considered by Vizzy/the Wizard to be an opportune time to trigger a prompt, was in fact P2 (and P1) thinking about a possible argument/reasoning, and not because they were "lost" or stuck. Interestingly, though, from our analysis of the transcripts across both conditions, pairs often did not seem to mind when Vizzy's prompts were not perfectly "aligned" with their ongoing conversation and they just "reoriented" their conversation towards the prompt, as it was also the case in this segment. In other cases, the pairs occasionally just ignored an imperfectly triggered prompt or answered it briefly to then carry on with another topic or with what they discussed prior to the prompt.

The second segment contains the same prompt but was asked in the screen condition (1-SCR). In this segment participants also had an extensive discussion in response to Vizzy's prompt. However, in contrast to the first segment (1-VOI), the discussion unfolds at a somewhat slower pace; there are several pauses in the conversation (e.g., line 13, 15, 35), the turns are longer and there is less of the rapid and iterative generation of ideas and hypotheses seen in the first segment.

```
01  P2   Maybe there is not that much of a difference in terms of lifestyle for
02       children.
03  P1   Yeah.
04  P2   Than for adults.
05  P1   Yeah.
06  P2   (              ) I don't know, run around and playing.
07  P1   So basically, this can't show that there is no differences in (1.9)
08       hormonal stuff...
09       (3.2)
10  VZ   ((Vizzy displays prompt: What might have caused the sudden spike?))
11  P1   ((Reads prompt and mumbles part of what she is reading.))
12       Around 96…
13       (7.9)
14       °I don't know°
```

```
15        (6.3)
16  P2    Mmmmh, (1.7) there might be a lot of (0.3) things. I mean, maybe there was
17        some (0.8) ehmm (1.8) maybe around 2000 there were lots of companies like
18        (      ) companies
19  P1    Yeah, that's, that's possible or maybe more ehm women, more MOMS started
20        working and not cooking so much or something like that. Or, (1.6) uhmm.
21  P2    Yeah (2.7), it's very hard to tell to be honest.
22  P1    Yeah.
23  P2    Uhmm.
24  P1    So, there is a sudden spike. (1.3) Yeah, but I think this shows that it's
25        mostly something abou::ut (0.4) the lifestyle, right?
26  P2    Yeah. And also, it's something that people realize is wrong, otherwise
27        (0.6) it wouldn't slow down.
28  P1    Yeah, and you need to take (0.8) care of it. (0.6) Yeah, that's what I
29        mean that (0.4) how much more can it go from 60%? (0.8) At some point you
30        will start (1.7) someho::ow taking (     ) from the government or from I
31        don't know.
32  P2    Mmhm.
33  P1    They will start taking (1.7) initiatives to slow it down.
34  P2    Mhm. (1.4) OK, so.
35        (3.2)
36  P1    So, OK, let's recap. (0.8) Uhmm, (switches to other visualization: Women
37        and Men, Global) So, basically (0.4) there are more women (2.2) but this
38        because it depends on (1.2) developed and (2.7) developing classification
39        (switches to other visualization: Women and Men, Developed and Developing
40        Countries) this may be because of a lifestyle change or because of
41        lifestyle differences.
43        ((Subsequently P1 continues to summarize further findings which they have
44        made and they both discuss and conclude what the main patterns were.))
```

Segment 1-SCR: Pair 12, Visualization on display: “girls and boys, developed and developing countries”.

Segment 1-SCR shows how the participants begin with a discussion about how the differences in children between developed and developing countries are less pronounced than for adults, which the pair explains is most likely due to their different lifestyles, in the sense that children may have more similar levels of physical activity between developed and developing countries than adults have. After a silence (line 9), Vizzy displays the prompt which both participants read while P1 mumbles it aloud to the other. This behavior was found to be typical in the screen condition; either one or both of the participants often read Vizzy’s prompt aloud or mumbled while reading it from the screen. In doing so, they let the other person know that they are currently reading it while drawing their joint attention to it.

It seems as if the pair struggles to know what to think or say at the beginning, which is also reflected by the long silence after Vizzy’s intervention (line 13/15). However, then Vizzy’s prompt triggers a long discussion between the two, where there is some back and forth in the way they come up with different hypotheses as to why the trend they are looking at occurred. Sometimes the hypotheses are generated as questions for the other to consider, for example P1 poses one of their hypotheses as: *“Yeah, but I think this shows that it's mostly something about the lifestyle, right?”* followed later by another one: *“Yeah, that's what I mean that how much more can it go from 60%?”*.

In this segment, both P1 and P2 make suggestions although P2 tends to agree with what P1 is suggesting following Vizzy's prompt with one-word answers or questions. This difference in who contributes the most was generally more marked in the screen condition, corroborating the quantitative findings. Why this should be the case may be due to one person taking the baton in steering the discussion as a result of being the person, who has implicitly decided to be the one who reads out Vizzy's prompt, or it could also be as a result of more vocal partners taking the lead. However, as previously described in the quantitative analyses, this difference was not found to be significant.

5.3.2 Responses to the Prompt "Is the Increase of One More Significant Than the Other?"

Next, we will look at two examples of conversations where Vizzy triggered the prompt *"Is the increase of one more significant than the other*?" in the voice condition (segment 2-VOI) and the screen condition (segment 2-SCR) for the visualization "women and girls, developed and developing countries" (see Figure 10). To literally answer this prompt, participants would just need to compare the increase in the different time series graphs. However, the question could be viewed as a "leading" one, implicitly asking them to consider why this might be the case. Indeed, in both conditions, the participant pairs had extensive discussions following this prompt rather than simply answering "yes" or "no" before moving onto the next visualization.

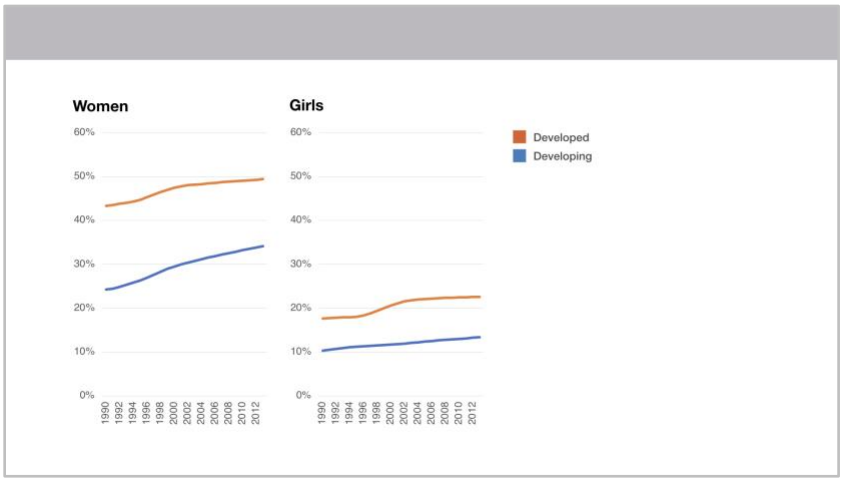

Figure 10. Visualization "women and girls, developed and developing countries".

Similar to the previous conversations, segment 2-VOI illustrates how P1 and P2 take turns to bounce their ideas off each other in an exploratory way (e.g., lines 1-10); how they generate hypotheses and what types of questions they ask each other while doing so (e.g., lines 8-10, 29). The conversation was also relatively fast paced, and the pair often appear to be thinking "on their feet" and out loud (e.g., lines 3-6, 8-10, 15-18, 23-26).

```
01  P2    I like the "developed girls". (0.8) It is FLAT. (0.4) It stopped (0.3)
02        rising.
03  P1    Yeah, it's (0.4) <it's really weird> it's like, w::e went from what, 18%
04        to 22% <then (    ) it>.((laughs))
05  P2    ((laughs)) (1.5) Ha. It hasn't changed at all.
```

```
06  P1    °Yeah°, but in developing countries it's steady:::y something.
07  P2    Hmm.
08  P1    We don't really know how it will evolve. (1.2) Like will it keep rising
09        slowly::y ((shows with hands)? Will it ((shows with hands)) jump u::up?
10        Will it stabili::ize?
11        (2.6)
12  VZ    Is the increase of one more significant than the other?
13        (1.4)
14  P2    No.
15  P1    Ehm, yeah, but it...
16  P2    Which one?
17  P1    Adults are, the...
18  P2    OK, ehm...Yeah.
19  P1    The increase is more significant.
20  P2    If you look between those two ((points at the graphs)), yeah...
21  P1    Yeah, the children ARE SOMEWHAT protected, I guess. ((laughs))
22  P2    ((laughs))
23  P1    It's like it's influencing them less, <but>, (0.3) <in the meantime> I
24        mean they are more ehhm (0.4) checked up by doctors, by everybody (    )
25        (0.4) which is their height their weight (       ). (0.5) Yeah, even in
26        schools their food is checked and basically...
27  P2    Mhm.
28  P1    And adults they can do basically whatever they want, so...
29  P2    Vizzy can you show global? (0.7) °What happens if we put them together?°
30        (3.5)
31        ((new visualization shows up))
32        (8.6)
33  P1    Yeah, it doesn't add much I guess.
34  P2    All four of them were almost steady right so if you add them you get
35        something almost steady. ((laughs))
36  P1    ((laughs)) Yeah, it makes sense ((laughs))
37  P2    ((laughs))
38        ((They subsequently continue discussing the new visualization on screen
39        and comparing it with the previous one.))
```

Segment 2-VOI: Pair 8, Visualization on display: “women and girls, developed and developing countries”.

In this segment, after Vizzy triggered the prompt *“Is the increase of one more significant than the other?”* P2 answers immediately “No.” However, P1 does not seem to agree with P2. Instead, P1 provides a reason why there is a more significant rise for adults compared with children: because they *“can do whatever they want”* whereas children are monitored much more when growing up. P2 listens to his explanation occasionally interjecting with disfluencies, suggesting she is considering P1’s explanation but still does not seem to fully agree. Then, after they discussed it for a certain amount of time, they move on to another visualization (line 29).

This segment shows how the participants generate their hypotheses by asking questions of the graph, connecting what they are seeing with possible reasons for the different trends but also hypothesizing what might happen. For example, P1 asks three questions in succession when looking at a line graph: *“Like will it keep rising slowly? Will it jump up? Will it stabilize?”*. Furthermore, it can be seen how the pair seems to use humor when formulating certain tentative or more tentative or “daring” hypotheses (e.g., line 21 where P1 says *“Yeah, the children ARE SOMEWHAT protected I guess.”* which then both laugh at). Taken together, the segment

shows how Vizzy's interjection led the pair to having a relatively fast-paced discussion, bouncing ideas off each other, asking each other questions and hypothesizing about the differences.

The 4th segment below (2-SCR) illustrates the patterns of conversation before and after Vizzy's prompt *"Is the increase of one more significant than the other*?" in the screen condition. In this segment, Vizzy triggered the prompt after a pause when the participants were deciding which visualization to look at next (lines 6-10). This may have given the impression that Vizzy was not helping them choose but instead providing a prompt about the current visualization they were looking at. The pair appeared not to mind Vizzy's intervention as they change tack to think about what the answer might be.

As can be seen from the segment, there is a long discussion between the participants which evolves rather slowly over time, consisting of relatively long speaker turns. Following Vizzy's prompt "*Is the increase of one more significant than the other?"* the pair work out an analysis (including calculations). Similar to the previous voice segment (1-VOI), participants discussed their answer quite extensively despite the prompt/question being a "yes/no" one. The participants also appeared to engage in a process of thinking aloud (lines 1-3, 22-25, 28-32) when hypothesizing about possible reasons for the increase.

In contrast to the previous segment (2-VOI), the pair spent considerable time looking at the visualization without speaking, before discussing it in more detail (line 19). There were several other instances of long silences when they were reading and appearing to figure out what the change in the slope of lines in the graph meant. However, they spent more time examining the graphs; reading off and inferring the percentages that helped them work through why there might be a significant difference – which did not happen in the previous segment.

```
01  P1   But I think maybe it just looks like a straight line because the increase
02       is too, [      (2.6)      it's too slow maybe we can't,      (2.2)       ]
03       yeah, maybe we cannot really identify this from this figure.
04  P2           [(0.6) Mmmh (1.3) The difference or the percentage is too narrow.]
05  P2   Mhm. (4.2) OK.
06  P1   OK. So, what to do next.
07       ((Both start to look at the cardboard cards.))
08  P2   Up to you. (3.1) One we haven't done.
09  P1   Yeah.
10       ((Both continue to look at the cardboard cards.))
11  VZ   ((Vizzy displays prompt: Is the increase of one more significant than the
12       other?))
13       ((Both read the prompt and mumble part of they read.))
14  P2   (4.1) What does it mean "one than the other"? It means the developed and
15       developing or women and girls?
16       (1.6)
17  P1   Mmh, (1.5) maybe (            ) the question.
18  P2   Hmm, I'm a bit confused.
19  P1   ((Both look at different parts of the visualizations for 21 seconds))
20  P2   More signi::ifi::icant.
21       ((Both look at the visualizations for another 14 seconds))
22  P1   Yeah, for me personally I would say the increase of the, of the women
23       [are more significant], because... But, we don't, we just know the
24       absolute increase is more significant, we don't know the corresponding
25       difference. So...
26  P2   [ Yeah, I think so. ]
```

```
27  P2    Mmh.
28  P1    So, () I cannot really answer this question (0.8) cause [ (0.9) ]... Maybe
29        to a simple conclusion, from about 25% to about 35% so it's, it's 10 over
30        25 is about 40% (1.5) and for the GIRLS (2.7) it seems a little bit less
31        than 40% so I, yeah, so I think the increase of the women is more
32        significant than the girls, yeah. So, do you see what I discovered?
33  P2                                                          [  Mhm. ]
34  P2    Mmh.
35  P1    Yeah, (1.4) yeah, under the condition that my calculation is correct.
36        ((laughs))
37  P1    ((Both mumble something and look at the cardboard cards and the tablet
38  P2    with the filter menu.))
39        (7.0)
40  P1    ((Starts making a selection on the tablet.)) Would we get the same result
41        for "male"?
42        ((Vizzy shows requested visualization.))
43        ((Both look at the visualization for about 8 seconds and then continue to
44        discuss the differences.))
```

Segment 2-SCR: Pair 17, Visualization on display: “women and girls, developed and developing countries”.

The discussion in this segment was initially more focused on making sense of the prompt (lines 17-21) followed by reading the data from the visualizations (lines 22-35). Here, they provide an answer to Vizzy’s prompt but do not suggest hypotheses or come up with ideas as to why this might be the case. Instead, they elaborate on Vizzy’s prompt, focusing on the visualization itself and the details in the change of the curves in the graphs. In doing so, they attempt to quantify the increase in percentage points. After exploring the data in this way, they conclude by conferring that there is a more significant increase in women’s obesity as compared to girls.

Overall, the above four segments illustrate the varied types of conversations that followed after Vizzy’s prompts, guiding the participants to think about possible reasons for the trends in the data while orienting them towards paying attention to particular aspects of the data. The main behaviors and characteristics identified in the qualitative analysis are summarized in Table 2. Some of the findings of the qualitative analysis corroborate those found in the quantitative analysis. For example, the pace of the discussion in regards of qualitative aspects (“thinking out loud” and quickly start exploring possible answers/ideas) corroborates the quantitative aspects described in section 5.1 (more turn-taking). The kinds of ensuing conversations for both the voice and screen condition had a similar pattern: after receiving a prompt by Vizzy, participants suggested different reasons, asked questions of each other and generated hypotheses. In terms of whether there were any marked differences between the screen and voice condition, we observed how participants in the voice condition tended to recommence the conversation relatively quickly after Vizzy provided the prompt and did so at a fast pace by bouncing ideas off each other, as if they were brainstorming. In contrast, in the screen condition the participants often took their time to start up the conversation again after having read the prompt on the screen. Stopping to read the prompt, therefore, had the effect of slowing down the conversation; when it resumed, there appeared to be less of the rapid and exploratory idea generation found in the voice condition. This finding corroborates the differences in turn-taking found in the quantitative analysis in section 5.1.

Table 2. Summarized findings of the qualitative analyses of the discussions following agent prompts showing key similarities and differences between both conditions.

Voice Condition		Screen Condition
Extended discussions exploring various hypotheses and possible explanations	=	Extended discussions exploring various hypotheses and possible explanations
After a prompt the discussion recommences quickly	≠	After a prompt the discussion slows down/pauses
After a prompt quickly starting to explore possible or tentative ideas/answers (by "thinking out loud")	≠	After a prompt discussing more about what the prompt means and how to answer it before discussing possible answers

5.4 Participants' Reflections on Vizzy

In the semi-structured interviews following the study, participants reflected upon the role Vizzy played in the task. However, as the experiment was designed to be between subjects, each pair only experienced one setting, so their reflections only refer to the experience they had. Thus, the interviews provided insights into how participants perceived and experienced Vizzy, its prompts and the role it plays – on a general level, independent of the condition/modality. Overall, the interviews showed that most of the participants found Vizzy and its prompts useful. The section is split into two subsections reflecting two general aspects of how Vizzy's role was understood by participants – how Vizzy's prompts (i) scaffolded and (ii) "slowed down" their thinking.

5.4.1 Scaffolding Participants' Thinking

About half of the pairs mentioned that it felt like Vizzy is *part of the discussion*, like a facilitator or even a collaborator, that helps them when they needed some input. For example, participant 1 in pair 4 in the voice condition (hereafter these identifiers will be abbreviated, in this case VOI-4-1) mentioned: *"It's like talking to a colleague, like Vizzy, could you please check…"* Furthermore, it also seemed as if most participants understood and appreciated Vizzy's behavior of prompting them to consider when there was a silence, without subsequently following up on it, for example participant 2 in pair 5 in the screen condition (SCR-5-2): *"The questions were interesting, they were all pointing to something that we have missed, for example the steadiness, I wouldn't have analyzed the steadiness myself."* Similarly, VOI-9-2 said: *"The suggestions were useful when we were having a break; it would help us see what else was there. It was waiting for us."*

The majority of participants pointed out that the assistant helped them to not get lost or stuck on a particular data visualization. It also allowed them to find additional differences or trends in the data when they thought that they had already discovered everything or couldn't find any other patterns, for example VOI-8-1: *"I think one thing that helped was that when we were kind of stuck and we were not saying anything, it would just generate a suggestion. I found that useful."* Another comment by SCR-5-2 illustrates how the prompts helped them "open up" their thinking and discussion: *"I like the fact that the questions were on finding out more about the data so asking like 'which was more steady, the men or the women?', by answering the question or even by looking at it you would think about the consequences of a steadier line and then you would ask yourself WHY is this more steady than the other which wouldn't necessarily happen without the assistant."*

About one third of the pairs mentioned that they would rather not have Vizzy interject, when performing a specific, well-defined task/analysis or when they knew what to look for. VOI-8-1 also thought Vizzy could help to generate hypotheses (i.e., exploratory analysis) instead of testing existing hypotheses (i.e., confirmatory analysis): *"If you have a lot of variables and you are not really sure what you are looking for or if you are training someone it might be a good thing to use. If I know what I am looking for, I probably won't use it. (…) I would use it to GENERATE hypotheses instead of TESTING my hypotheses."* And similarly, SCR-1-1: *"It depends on if the data that I am working on is something that I am familiar with. If it is something that I have been working on for the past five years probably not [use such a system] – if I am using new data, sure."* These comments also suggest that participants consider the system to be more suitable for working with new datasets rather than with familiar ones.

A couple of groups also reflected on how Vizzy helped them remember the data by making them talk about it. For instance, SCR-9-1 said: *"I am really impressed what we all remember from that, so maybe it is also the thing of remembering data by talking about it and having a facilitator."* And similarly, VOI-5-2: *"Because it is so interactive, I think it stays in my memory as well."*

However, there were a few of the participants who did not like the way Vizzy prompted them. SCR-3-1 commented that they were not always helpful: *"Sometimes it asked something that we already discussed or that we were in the middle of discussing."* (Even if the experimenter tried to provide prompts about aspects that participants have not previously discussed, it was not always possible to find prompts without any overlap.) SCR-9-2 commented that *"it was more like an examiner as we need to find an answer to the question it asks. While the guidance is quite minimal."* SCR-7-2 also reflected: *"it is like they [Vizzy] are joining the conversation and immediately leaving it".* One participant, SCR-7-1 mentioned it would be better to *"have it help only when we want help"* rather than it being proactive. However, this was rather an exception – most of the other groups said Vizzy's interjections were helpful, probing them, steering their thinking and guiding them to know what to look for in the data.

More than half of the pairs in the voice condition mentioned positive aspects related to the shareability of a voice interface and its suitability for collaborative situations, which was not the case in the screen condition. This corroborates the tendency towards more balanced interactions in the voice condition that was found in the quantitative analysis (section 5.2). For example, VOI-7-2 mentioned: "*One big advantage is that we are both in control, whereas in a typical laptop or tablet scenario it would either be my computer, or his computer and he says let's look at 'women and girls' and then I would have to change it. It is a nice interaction when we are both in control. We are exploring more actively."*

5.4.2 Slowing Down Participants' Thinking

About a third of the pairs mentioned that they had the impression Vizzy made them do the data analysis task more slowly than if they were just doing it by themselves or with common analytics tools. However, most of them acknowledged that this slowing down effect can also have benefits, for example, in situations where they are exploring a new dataset/topic or getting a new perspective on one they may already be familiar with, such as SCR-9-1: *"I think it is good [to use this system] if you have time and you are trying to figure out things."* Similarly, SCR-2-2: *"I mean it was more time-consuming than traditional tools but that also has benefits if you are not in a rush."* Related to this, SCR-5-1 also described how they understood the concept of the agent that does not follow up: *"Instead of an assistant I would say it's like a tutor, so he or she has the answer already and*

he or she is trying to guide me." This illustrates the different purposes an agent can have, one that helps to get things done (i.e., an assistant) versus one that guides the users in doing the thinking themselves (i.e., a tutor), which was our aim in the present scenario.

Two pairs mentioned that this kind of slower interaction (i.e., "tutor model") would be most suitable for users who are getting familiar with a dataset (or data analysis more generally). Once they have become familiarized, they may then prefer for the system to become "faster" (i.e., "assistant model"), by making suggestions about what visualizations to look at first and in what sequence or by having additional commands/controls for switching between visualizations more quickly. For example, VOI-6-1 mentioned: *"If you are looking at the same data for an extended period of time, you mostly want it to be very fast to get data out. This isn't exactly fast. I guess this is more suitable if you are introducing a new topic or if you are trying to get a new perspective on the same data."*

This comment summarizes well the main purpose of Vizzy, which is to support users in exploring a (new) dataset/topic from different angles. Most participants understood that the purpose of Vizzy is less to enable users to quickly and efficiently conduct specific (confirmatory) analyses. They understood that it is mainly designed for users, who are not (yet) experts in the dataset and/or data analytics – or it can allow those who are experts to approach familiar datasets from a different angle (in the words of VOI-6-1: *"to get a new perspective"*).

Furthermore, the participants' comments above illustrate the trade-off between providing an essentially proactive agent that probes (i.e., "tutor model") versus a largely automated reactive one (i.e., "assistant model"), that does more of the work. The former can help users to think more for themselves and see connections, sequences and use their commonsense knowledge, whereas wanting something more intelligent is based on a desire for an agent that can "do the thinking" on their behalf, generating hypotheses they could then concur with and accept. Here, our interest was in how agents could act more as facilitators or tutors, probing the users so that they get a better understanding and are engaged in the sensemaking activity; in other contexts, it may be more desirable if the agent takes more the role of an assistant, doing more of the computation, making suggestions and drawing conclusions, rather than prompting the user to do the reasoning, learning or decision-making themselves. Which kind of agent to model will depend on the role desired of an agent in a given setting. Depending on the setting, it may be more important to support users in learning and acquiring new knowledge (and in enabling them to transfer that knowledge) or to help them become more effective at solving a specific problem. In other settings the goal may be to "just" make users complete a specific task as quickly as possible.

6 DISCUSSION

Our study has shown how voice versus screen-based human-agent interaction can affect users' conversation and collaborative sensemaking as well as users' interactions with the system incorporating an agent. Supporting our first hypothesis, we found participant pairs in the voice condition made significantly more requests to Vizzy and explored more of the available data visualizations. Supporting our second hypothesis we found participant pairs took more turns and asked each other more questions when interacting with the system in the voice condition compared with the screen condition. When analyzing the conversations to determine why this was the case, we observed the interactions in the voice condition to be at a faster pace with more bouncing off ideas between the participants.

One possible reason for these differences is the use of voice is more seamless and better aligned with the human-human interaction/conversations that took place, in the sense that participants could just "embed" the voice requests into their ongoing conversation while keeping their eyes on the screen depicting the visualizations. The finding that participants took more turns and asked more questions in the voice condition can be further explained by the use of voice being more immediate and direct compared to screen. This may have stimulated more discussion and encouraged more turn-taking as participants explored different hypotheses and ideas. Furthermore, in the voice condition Vizzy "takes turns" itself by intervening in participants' conversations, which was not the case in the screen condition – where the participants instead had to stop their conversation to start reading from the screen. Vizzy actively intervening in the conversation by speaking aloud may have motivated participants to proactively take turns and generate ideas or questions like Vizzy did (as it can also be seen in segment 1-VOI). However, in the follow-up interviews, some of the participants also mentioned challenges of the system and the modality of voice, namely that it was sometimes awkward if the system just chimed in, in particular, if it did so in unexpected ways. This could be, for example, when Vizzy provided a prompt, which overlapped with what they previously discussed.

Our findings resonate with those found by Gonzalez and Gordon [25] on the effects of different modalities in interactive narratives, where the voice modality resulted in participants understanding their "role" differently and behaving in a different way compared to text-based interaction. Also, it concurs with findings from previous studies, where participants had more interactions with each other after the system spoke aloud following certain actions they made when using a tangible interface (e.g., [20]). Furthermore, voice requires immediate joint attention when it occurs, whereas reading text from a screen together requires paying attention in a different way. For the latter, there may be a slight delay as one waits for the other to finish reading and knowing when it is appropriate to start the conversation again.

The qualitative analyses also indicated that there were nuanced differences between the conditions in how engaged and interactive the discussions were and the extent to which ideas and hypotheses were generated. In particular, participants in the screen condition often needed a bit more time to start the discussion and there were more silent pauses. In this condition it seemed as if the participants were thinking more about what the prompts mean and how they should answer before discussing possible answers, while in the voice condition they often immediately started discussing possible answers as if they were "thinking out loud". It seemed as if they were willing to be exploratory in their discussion and less focused on providing a "correct" answer to Vizzy's prompt.

There could be a number of reasons for the differences mentioned above. Firstly, if a prompt is asked via voice, people may feel more compelled to answer it, as it feels more similar to interacting with another human being. Thus, the users may adhere more to the rules of human-human conversations where it would be awkward or inappropriate to wait more than a few seconds before responding. Text-messages on the other hand, even if represented as chat bubbles, feel less human-like and people may feel less compulsion to answer them immediately. This suggests that the agent's modality impacts on how users conceptualize, think of, respond to and "treat" the agent. Thus, the difference between conditions in how participants responded to an agent prompt, is most likely not just due to how the human mind processes one or the other modality, but also due to how people conceptualize the agent depending on the modality. In other words, when an agent feels more human-like (i.e., because it speaks), people may not want to "let it wait" for too long, when it asks them something. This may partly explain the more immediate responses to agent prompts observed in the voice condition, where

participants tended to just start thinking out loud and discussing possible answers. Secondly, since in the voice condition Vizzy actively intervened in the conversation, participants may have also been more proactive and just saying what they think. Text messages, on the other hand, are less of an active intervention in the ongoing conversation, which possibly resulted in participants being less proactive themselves. Thirdly, another reason may be that when participants were in the middle of a conversation and they were prompted through voice, it may have been more natural for them to integrate Vizzy's prompt into the flow of their conversation. In other words, as there is no change in modality (from reading text to speaking and vice versa, as it is the case in the screen condition) they could just continue with the discussion.

However, if the agent is perceived to be "butting in" too often it could become annoying (in particular, in the voice condition). Given that we had designed Vizzy to have a minimal level of interaction (i.e., one prompt approximately every two minutes), the pairs were usually forgiving on the occasion when being interrupted in their ongoing conversation. Moreover, we deliberately designed Vizzy to prompt the participant pairs at opportune times rather than "joining in" their ongoing conversation. The pairs quickly understood this underlying user model and it seemed from the interviews they were happy with it in general – not wanting or expecting Vizzy to be an equal partner in the conversation. This limited form of "proactive agency", therefore, may in the long run be more effective than trying to design the agent to be a human-like conversationalist, at least for group settings where it may be undesirable to have a system that intervenes too often in an ongoing conversation between humans.

Taken together, the quantitative and qualitative findings suggest that voice interfaces can enable a faster "pace" in the conversation in terms of its structure (turn-taking) but also in terms of its content (responding more immediately and "bouncing off "ideas). Furthermore, we found a tendency towards more balanced human-human and human-agent interactions in the voice conditions. In addition, voice modality seems to lead to more exploratory behavior and curiosity in terms of how many questions are being asked but also in how Vizzy's prompts are responded to (coming up with different ideas/hypotheses and "thinking out loud"). Finally, the voice modality also showed a higher engagement in terms of interactions with the system (number of visualizations *requested* and *explored*).

6.1 Limitations

The study investigated how an agent could support users in an exploratory task via prompts, which it provided proactively at opportune times. It is worth noting that (i) these prompts were prepared prior to the study by the research team and (ii) that they were triggered by a human experimenter/the "wizard" (approximately every two minutes) based on simple rules, which were (a) the topic of the prompt was not previously discussed by participants and if (b) there was a silence of at least three seconds. Hence, there is the aspect of (i) appropriateness of the prompt itself given the context (i.e., the ongoing discussion between participants) and (ii) appropriateness of its specific timing. If (i) the creation of the prompts would be implemented and automated, they may not always reach the same accuracy and quality. However, several prompts in this study were in fact not highly complex and usually focused on a distinct pattern in the data (e.g., "Is the increase of one more significant than the other?"). Hence, it is likely that a system that scans for certain distinct patterns in the (time series) data would be able to come up with prompts similar to those designed in this study. If (ii) the timing of prompts would be fully implemented based on monitoring the ongoing conversation (using natural language processing), it is likely that the system would not be as accurate in identifying appropriate moments to trigger a

prompt – despite the relatively simple rules that were used. However, it is worth noting that in the present scenario the effect of potentially inaccurate timing may be smaller compared to others, since the agent prompts were designed to be rather infrequent, and our study showed that the majority participants did not mind occasional imperfect timing in the given exploratory, open-ended task as seen in the qualitative analyses.

The study compared two conditions: voice-based agent input and output versus screen-based agent input and output (along with the screen-based presentation of visualizations in both conditions). It is likely that some of the effects observed were more strongly related to the output than to the input in the specific modality and vice versa. With the present study design, it is not possible to fully disambiguate the effects of input and output in the respective modalities, given the inclusion of only two modality combinations (for the agent interactions). Other possible combinations/permutations may be investigated in future research to understand in more detail for which behaviors input and for which output in the respective modality is more relevant. In the present study, we intentionally focused on voice-based versus screen-based agent-enabled systems – reflecting the two combinations we considered most natural and appropriate for a variety settings and applications.

With the chosen approach there are also certain limitations concerning the chosen metrics. As such, the number of visualizations participants *requested* and *explored* need to be understood and interpreted only as proxies of participants' engagement in the data exploration task. It is thus important to interpret these metrics in conjunction with the other quantitative and qualitative analyses to get a better understanding of participants' engagement, sensemaking and curiosity. However, since in the present scenario participants were asked to discuss the visualizations, which they chose to look at together, it is worth noting that there was always some engagement for each requested visualization, and both metrics can thus be considered meaningful proxies for task engagement in the given scenario/task.

Taken together, the metrics of this study were focused on how much participant pairs interact with the system and the task material as well as how they interact with each other (in terms of turns taken and questions asked), which we triangulated with the qualitative analysis of pairs' interactions. Although these analyses cover multiple aspects that are relevant for the present scenario, there are other behaviors that can be of interest depending on the research motivation or approach. Our focus was on how the human-computer and human-human interactions were affected by the modality on a general level. However, the data visualization field, for example, may also be interested in understanding users' task-related performance, such as what they infer or conclude and how quickly they are able to do so. Although such analyses were outside the scope of this work, future research could address the effects of different interaction modalities at a more granular level (e.g., specific to the task domain).

In summary, the findings from our study suggest the use of a limited form of proactive *voice* agent/interface could be preferable compared with a *screen-based* one for (multi-user) settings where a high level of interaction with the tool/system is desired, along with a fast-paced conversation, more questioning, as well as "think aloud" and rapid/more immediate answers and idea generation. However, this does not mean that voice will be better in every situation; in some cases, it may be desirable to slow down or stop the current discussion by the use of screen-based prompts to trigger different types of thinking or to get participants discuss the prompts and build a shared understanding before answering.

6.2 Design Implications

Below we present a set of design recommendations, based on the findings from our study, for designing screen or voice-based agent interfaces that are specifically intended for supporting groups interacting with a software tool for open-ended and exploratory tasks. It assumes that the agent is there to support and facilitate the human activity rather than to provide answers. It may be that some of the recommendations are also generalizable to other settings, involving different tasks, contexts and individual users.

Having a voice interface may elicit more (rapid) questioning, idea generation and exploration of the provided problem space if the task is open-ended* and *does not have a specific/well-defined outcome.

Our study showed that there are differences in how the pairs carry out their task if they are speaking and listening to an agent versus selecting commands and reading agent's prompts from the screen. Participants in the voice condition interacted more with the system, explored more of the available visualizations and asked more questions. However, differences in the conversation patterns were rather nuanced. Participants in the screen condition often needed a bit more time to start off with the discussion after a prompt and had more silent pauses. In contrast, in the voice condition the pairs spoke more as if they were "thinking out loud" and were more willing to brainstorm and be more exploratory in their discussion. Furthermore, voice prompts could directly be responded to by participants in the voice condition, whereas in the screen condition, participants had to first decide when to direct their attention to the text prompt and read it. In addition, in the screen condition participants seemed to spend more time thinking and discussing about the prompt and building a shared understanding of it. In the voice condition, there seemed to be a tendency to avoid longer silences after a prompt and keep the conversation going. Thus, it seemed participants spent less time reflecting on the prompt and instead answering commenced more immediately by just starting to generate hypotheses and to explore possible answers.

Having a limited form of proactive agency, in the form of occasional context-specific prompts can be effective at scaffolding and steering a conversation in group settings.

Our study showed that designing the agent's interventions at a minimal level was generally found not to be disruptive or annoying for this open-ended task. The pairs in both conditions readily understood that the agent was not designed to have a conversation with them or to be an equal partner in the conversation but that it would occasionally prompt them when it appeared that they were in need of help or getting stuck. This meant they had no expectation that the agent would be like a human-like conversationalist and therefore did not ask it questions it could not answer or get "confused" by.

It may not be critical for open-ended tasks that the agent identifies the best moment to intervene in the human conversation for either voice or screen-based interfaces.

When designing an agent that intervenes in human-human conversations, potential interventions need to be more carefully considered for voice interfaces compared with text-based interfaces, since they are more likely to disrupt an ongoing conversation than a text-based one. In our study, sometimes in the voice condition Vizzy interrupted when one participant was speaking. But this turned out to not be a problem, as the participants generally appeared not to mind. Most pairs either carried on speaking or stopped and listened to Vizzy. In many ways this is similar to a human-human conversation, when sometimes two speakers overlap, or another interrupts a conversation. In contrast, in the screen condition: When a prompt popped up on the screen, the participants could read it when they were ready to do so. However, if there is limited interaction from the agent,

interruptions generally do not matter so much at least for open-ended tasks where users do not have a specific idea of how they should or have to proceed with the task.

The interface should be designed to make it easy for all group members to equally interact with the agent – whether it is screen or voice-based – so that everyone has equal control.

Our interactive data analysis tool was designed to provide a familiar style of visualizing data (line graphs) that users could request to be generated for different time series. For the screen condition, widely used graphical menu/filter interfaces and chatbot-like message display were mimicked. For the voice condition, the way requests were made (so-called *wake word* followed by a request: "*Vizzy, show …*") was based on familiar voice interfaces. For both conditions, the groups readily learnt how to interact with the agent and did not expect it to do more than occasionally prompting them. Enabling each participant to interact with the system by making it equally accessible to both of them is important for both interface types though easier to achieve with voice interfaces. One reason for choosing a tablet for the screen condition was that it would be easily accessible to both participants if it is positioned between them. Nevertheless, it occasionally happened that one participant was "taking control" over it (e.g., by bending over the tablet).

6.3 Future Work

Our study was limited to exploring how voice versus screen-based interactions compared for a specific kind of task. For the findings to be generalizable to other contexts, further research is needed to investigate whether similar effects can be found with other tasks (e.g., problem-solving, decision-making, or other types of data analysis tasks) which could either be open-ended or well-defined and which users may be familiar with or not. This could lead to further insights into which types of tasks proactive/probing agents that take the role of a facilitator are most suitable for. For the task it would be interesting to explore different modalities, too. In the present study, a visual task (examining data visualizations) was chosen in combination with an agent-enabled interface that was screen-based (visual) versus a voice-based (auditory) one. Another avenue of research could be to vary the modalities of the task (e.g., an auditory one) to examine the effects different "agent modalities" would have when combined with different "task modalities". Another direction would be to compare how well different agent-enabled interfaces and modalities work for different user groups or types of users (e.g., lay versus expert).

The goal of our study was to investigate the differences between screen-based and voice interaction modalities at a general level – for both input and output. It would be interesting to see in future research the effect of other combinations of modalities – e.g., voice input and screen output or voice output and screen input. This would provide further understanding of how the modalities interact with each other (i.e., interaction effects) and can complement each other for different tasks and behaviors. For example, for certain tasks it may be important that it is easy to make requests to the system while being engaged in a conversation (i.e., voice input) but at the same time it may be important that users can carefully read a prompt (i.e., screen output), as it may be required that participants consider and discuss in more detail what it means/refers to (e.g., for more complex prompts).

The present study investigated how a specific set of relevant user behaviors can be affected by the interaction modality. Future research could examine other behaviors in more detail, such as (non-verbal) communication involved in the coordination and the "micro-interactions" involved in agreeing on joint actions,

making requests, responding to agent prompts and jointly diverting the attention (e.g., from an agent prompt or a visualization and vice-versa). Regarding the design of the interface and agent, future work could investigate the effects of using different types of prompts – such as different types of questions to be asked by the agent [23,26,61:91] – to further explore how and when the agent needs to intervene so that it optimally integrates into and supports users' ongoing conversations and thinking processes.

7 CONCLUSION

Our study has shown how interacting with an interface that incorporates an agent either through a voice or a graphical/screen-based interface with chatbot-type messages provides new opportunities for user interactions: scaffolding and prompting users when completing an open-ended sensemaking task, such as exploring a dataset using a data visualization tool. The agent's role of being a facilitator that occasionally provides a prompt for things to consider and patterns to look at in the data was readily accepted by participants and appears to be a promising approach for how agents can be designed to become "part of" and facilitate human-human conversation and collaboration in the future, without taking away control from the users. Participants using the voice interface were found to be more engaged in the sensemaking task compared with the screen condition and interactions tended to be more balanced. Having an agent that speaks directly to them, led to users asking more questions and taking more turns, resulting in many of their discussions resuming and progressing more rapidly than when reading the same prompt in a chat window on a screen. However, presenting prompts on a screen also has potential benefits; users can decide when to read the prompts leading to fewer interruptions of their ongoing conversations. Furthermore, having to pause the conversation to read a prompt can have the effect of slowing down users' conversation and thinking, which in some contexts may result in users spending more time thinking about what the prompt means and how to answer it. In sum, it can be "good to talk" using an agent-enabled interface in either modality: voice-based interactions may encourage more fast-flowing talk, while screen-based interactions may slow down the conversation. Which is preferable depends on what the activity or task is about.

Acknowledgements

We would like to thank everyone who participated in our study, the pilot studies, and our workshop. We are also grateful for the support by the Great Ormond Street Hospital DRIVE team for their help in running the workshop and pilot studies. Finally, we would like to extend our sincere thanks to our anonymous reviewers as well as Warren Park and Michael Reicherts for their valuable comments, suggestions, and ideas for improving the manuscript.

References

[1] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical Agents for Fostering Question-Asking Skills in Children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), Association for Computing Machinery, Honolulu, HI, USA, 1–13. DOI:https://doi.org/10.1145/3313831.3376776

[2] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3 (April 2019), 1–28. DOI:https://doi.org/10.1145/3311956

[3] Richard Arias-Hernandez, Linda T. Kaastra, Tera M. Green, and Brian Fisher. 2011. Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. In *Proceedings of the 2011 44th*

Hawaii International Conference on System Sciences (HICSS '11), IEEE Computer Society, USA, 1–10. DOI:https://doi.org/10.1109/HICSS.2011.339

[4] J. Maxwell Atkinson and John Heritage. 1999. Transcript Notation - Structures of Social Action: Studies in Conversation Analysis. *Aphasiology* 13, 4–5 (April 1999), 243–249. DOI:https://doi.org/10.1080/026870399402073

[5] Jillian Aurisano, Abhinav Kumar, Alberto Gonzalez, Jason Leigh, Barbara DiEugenio, and Andrew Johnson. 2016. Articulate2: Toward a conversational interface for visual data exploration. In *IEEE Visualization*.

[6] Grace M. Begany, Ning Sa, and Xiaojun Yuan. 2016. Factors Affecting User Perception of a Spoken Language vs. Textual Search Interface: A Content Analysis. *Interact Comput* 28, 2 (March 2016), 170–180. DOI:https://doi.org/10.1093/iwc/iwv029

[7] Diana Beirl, Nicola Yuill, and Yvonne Rogers. 2019. Using Voice Assistant Skills in Family Life. (June 2019). Retrieved August 27, 2020 from https://repository.isls.org//handle/1/1750

[8] Amanda Benedict-Chambers, Sylvie M. Kademian, Elizabeth A. Davis, and Annemarie Sullivan Palincsar. 2017. Guiding students towards sensemaking: teacher questions focused on integrating scientific practices with science content. *International Journal of Science Education* 39, 15 (2017), 1977–2001. DOI:https://doi.org/10.1080/09500693.2017.1366674

[9] Ludovic Le Bigot, Patrice Terrier, Eric Jamet, Valérie Botherel, and Jean-François Rouet. 2010. Does textual feedback hinder spoken interaction in natural language? *Ergonomics* 53, 1 (January 2010), 43–55. DOI:https://doi.org/10.1080/00140130903306666

[10] Jessy Ceha, Nalin Chhibber, Joslin Goh, Corina McDonald, Pierre-Yves Oudeyer, Dana Kulić, and Edith Law. 2019. Expression of Curiosity in Social Robots: Design, Perception, and Effects on Behaviour. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), Association for Computing Machinery, Glasgow, Scotland Uk, 1–12. DOI:https://doi.org/10.1145/3290605.3300636

[11] Christine Chin and Jonathan Osborne. 2008. Students' questions: a potential resource for teaching and learning science. *Studies in Science Education* 44, 1 (March 2008), 1–39. DOI:https://doi.org/10.1080/03057260701828101

[12] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interact Comput* 31, 4 (June 2019), 349–371. DOI:https://doi.org/10.1093/iwc/iwz016

[13] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces* (IUI '93), Association for Computing Machinery, New York, NY, USA, 193–200. DOI:https://doi.org/10.1145/169891.169968

[14] Laurie Damianos, Dan Loehr, Carl Burke, Steve Hansen, and Michael Viszmeg. 2003. The MSIIA Experiment: Using Speech to Enhance Human Performance on a Cognitive Task. *International Journal of Speech Technology* 6, 2 (April 2003), 133–144. DOI:https://doi.org/10.1023/A:1022334530417

[15] Beth Davey and Susan McBride. 1986. Generating Self-Questions after Reading: A Comprehension Assist for Elementary Students. *The Journal of Educational Research* 80, 1 (September 1986), 43–46. DOI:https://doi.org/10.1080/00220671.1986.10885720

[16] Kedar Dhamdhere, Kevin S. McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring Data with Conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI '17*, ACM Press, Limassol, Cyprus, 493–504. DOI:https://doi.org/10.1145/3025171.3025227

[17] Sidney K. D'Mello, Nia Dowell, and Arthur Graesser. 2011. Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied* 17, 1 (2011), 1–17. DOI:https://doi.org/10.1037/a0022674

[18] Sylvie Dubois, Martine Boutin, and David Sankoff. 1996. The Quantitative Analysis of Turntaking in Multiparticipant Conversations. *University of Pennsylvania Working Papers in Linguistics* 3, 1 (1996), 20.

[19] Sebastian Egger, Michal Ries, and Peter Reichl. 2010. Quality-of-experience beyond MOS: experiences with a holistic user test methodology for interactive video services. In *21st ITC Specialist Seminar on Multimedia Applications-Traffic, Performance and QoE*, Citeseer, 13–18.

[20] William Farr, Nicola Yuill, Eric Harris, and Steve Hinske. 2010. In my own words: configuration of tangibles, object interaction and children with autism. In *Proceedings of the 9th International Conference on Interaction Design and Children* (IDC '10), Association for Computing Machinery, New York, NY, USA, 30–38. DOI:https://doi.org/10.1145/1810543.1810548
[21] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein. 2018. Iris: A Conversational Agent for Complex Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, ACM Press, Montreal QC, Canada, 1–12. DOI:https://doi.org/10.1145/3173574.3174047
[22] Luke Fryer and Rollo Carpenter. 2006. Bots as language learning tools. *Language Learning & Technology* 10, 3 (2006), 8–14.
[23] James J. Gallagher and Mary Jane Aschner. 1963. A Preliminary Report on Analyses of Classroom Interaction. *Merrill-Palmer Quarterly of Behavior and Development* 9, 3 (1963), 183–194.
[24] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (UIST '15), ACM, New York, NY, USA, 489–500. DOI:https://doi.org/10.1145/2807442.2807478
[25] Diego Gonzalez and Andrew S. Gordon. 2018. Comparing Speech and Text Input in Interactive Narratives. In *23rd International Conference on Intelligent User Interfaces* (IUI '18), Association for Computing Machinery, Tokyo, Japan, 141–145. DOI:https://doi.org/10.1145/3172944.3172999
[26] Art Graesser, Vasile Rus, and Zhiqiang Cai. 2008. Question classification schemes. In *Proceedings of the Workshop on Question Generation*, 10–17.
[27] Arthur C. Graesser. 2011. Learning, thinking, and emoting with discourse technologies. *Am Psychol* 66, 8 (November 2011), 746–757. DOI:https://doi.org/10.1037/a0024974
[28] Arthur C. Graesser, Kurt VanLehn, Carolyn P. Rosé, Pamela W. Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI Magazine* 22, 4 (2001), 39–39.
[29] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How information visualization novices construct visualizations. *IEEE Trans Vis Comput Graph* 16, 6 (December 2010), 943–952. DOI:https://doi.org/10.1109/TVCG.2010.164
[30] Florian Hammer, Peter Reichl, and Alexander Raake. 2004. Elements of interactivity in telephone conversations. In *Eighth International Conference on Spoken Language Processing*.
[31] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *J Comput Mediat Commun* 25, 1 (March 2020), 89–100. DOI:https://doi.org/10.1093/jcmc/zmz022
[32] Alexander G. Hauptmann and Alexander I. Rudnicky. 1990. A comparison of speech and typed input. In *Proceedings of the workshop on Speech and Natural Language - HLT '90*, Association for Computational Linguistics, Hidden Valley, Pennsylvania, 219–224. DOI:https://doi.org/10.3115/116580.116652
[33] Christian Heath, Jon Hindmarsh, and Paul Luff. 2010. *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. SAGE Publications, Inc., London. DOI:https://doi.org/10.4135/9781526435385
[34] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (January 2018), 309–318. DOI:https://doi.org/10.1109/TVCG.2017.2744684
[35] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2011. Sequential Model-based Optimization for General Algorithm Configuration. In *Proceedings of the 5th International Conference on Learning and Intelligent Optimization* (LION'05), Springer-Verlag, Berlin, Heidelberg, 507–523. DOI:https://doi.org/10.1007/978-3-642-25566-3_40
[36] Richard Jacques, Asbjørn Følstad, Elizabeth Gerber, Jonathan Grudin, Ewa Luger, Andrés Monroy-Hernández, and Dakuo Wang. 2019. Conversational Agents: Acting on the Wave of Research and Development. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19), Association for Computing Machinery, Glasgow, Scotland Uk, 1–8. DOI:https://doi.org/10.1145/3290607.3299034

[37] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71, (November 2018), 1–15. DOI:https://doi.org/10.1016/j.wocn.2018.07.001

[38] Rogers Jeffrey Leo John, Navneet Potti, and Jignesh M. Patel. 2017. Ava: From Data to Insights Through Conversations. In *CIDR 2017*.

[39] Jan-Frederik Kassel and Michael Rohs. 2018. Valletto: A Multimodal Interface for Ubiquitous Visual Analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6. Retrieved February 19, 2021 from https://doi.org/10.1145/3170427.3188445

[40] Kim A. Kastens, Melissa Zrada, and Margie Turrin. 2019. What kinds of questions do students ask while exploring data visualizations? *Journal of Geoscience Education* 0, 0 (October 2019), 1–21. DOI:https://doi.org/10.1080/10899995.2019.1675447

[41] Alison King. 1989. Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology* 14, 4 (October 1989), 366–381. DOI:https://doi.org/10.1016/0361-476X(89)90022-2

[42] Alison King. 1994. Autonomy and question asking: The role of personal control in guided student-generated questioning. *Learning and Individual Differences* 6, 2 (January 1994), 163–185. DOI:https://doi.org/10.1016/1041-6080(94)90008-6

[43] Ludovic Le Bigot, Eric Jamet, Jean-François Rouet, and Virginie Amiel. 2006. Mode and modal transfer effects on performance and discourse organization with an information retrieval dialogue system in natural language. *Computers in Human Behavior* 22, 3 (May 2006), 467–500. DOI:https://doi.org/10.1016/j.chb.2004.10.006

[44] Ludovic Le Bigot, Patrice Terrier, Virginie Amiel, Gérard Poulain, Eric Jamet, and Jean-François Rouet. 2007. Effect of modality on collaboration with a dialogue system. *International Journal of Human-Computer Studies* 65, 12 (December 2007), 983–991. DOI:https://doi.org/10.1016/j.ijhcs.2007.07.002

[45] Eun-Ju Lee. 2008. Flattery may get computers somewhere, sometimes: The moderating role of output modality, computer gender, and user gender. *Int. J. Hum.-Comput. Stud.* 66, 11 (November 2008), 789–800. DOI:https://doi.org/10.1016/j.ijhcs.2008.07.009

[46] Hannah Limerick, James W. Moore, and David Coyle. 2015. Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), ACM, New York, NY, USA, 3967–3970. DOI:https://doi.org/10.1145/2702123.2702379

[47] Lijia Lin, Paul Ginns, Tianhui Wang, and Peilin Zhang. 2020. Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education* 143, (January 2020), 103658. DOI:https://doi.org/10.1016/j.compedu.2019.103658

[48] Diane J. Litman, Carolyn P. Rosé, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2004. Spoken Versus Typed Human and Computer Dialogue Tutoring. In *Intelligent Tutoring Systems*, James C. Lester, Rosa Maria Vicari and Fábio Paraguaçu (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 368–379. DOI:https://doi.org/10.1007/978-3-540-30139-4_35

[49] Ramon Marinez. 2015. Level and Trends of Overweight and Obesity. *Tableau Public*. Retrieved August 27, 2020 from https://public.tableau.com/profile/ramon.martinez#!/vizhome/LevelandTrendsofOverweightandObesity/Overweightandobesitylevel

[50] Paul Marshall, Eva Hornecker, Richard Morris, Nick Sheep Dalton, and Yvonne Rogers. 2008. When the fingers do the talking: A study of group participation with varying constraints to a tabletop interface. In *2008 3rd IEEE International Workshop on Horizontal Interactive Human Computer Systems*, 33–40. DOI:https://doi.org/10.1109/TABLETOP.2008.4660181

[51] Manolis Mavrikis, Beate Grawemeyer, Alice Hansen, and Sergio Gutierrez-Santos. 2014. Exploring the Potential of Speech Recognition to Support Problem Solving and Reflection. In *Open Learning and Teaching in Educational Communities* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 263–276. DOI:https://doi.org/10.1007/978-3-319-11200-8_20

[52] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported*

Cooperative Work and Social Computing (CSCW '17), Association for Computing Machinery, Portland, Oregon, USA, 2208–2220. DOI:https://doi.org/10.1145/2998181.2998335

[53] Clifford Nass, Erica Robles, Charles Heenan, Hilary Bienstock, and Marissa Treinen. 2003. Speech-Based Disclosure Systems: Effects of Modality, Gender of Prompt, and Gender of User. *International Journal of Speech Technology* 6, 2 (April 2003), 113–121. DOI:https://doi.org/10.1023/A:1022378312670

[54] Marie Ng, Tom Fleming, Margaret Robinson, Blake Thomson, Nicholas Graetz, Christopher Margono, Erin C Mullany, Stan Biryukov, Cristiana Abbafati, Semaw Ferede Abera, Jerry P Abraham, Niveen M E Abu-Rmeileh, Tom Achoki, Fadia S AlBuhairan, Zewdie A Alemu, Rafael Alfonso, Mohammed K Ali, Raghib Ali, Nelson Alvis Guzman, Walid Ammar, Palwasha Anwari, Amitava Banerjee, Simon Barquera, Sanjay Basu, Derrick A Bennett, Zulfiqar Bhutta, Jed Blore, Norberto Cabral, Ismael Campos Nonato, Jung-Chen Chang, Rajiv Chowdhury, Karen J Courville, Michael H Criqui, David K Cundiff, Kaustubh C Dabhadkar, Lalit Dandona, Adrian Davis, Anand Dayama, Samath D Dharmaratne, Eric L Ding, Adnan M Durrani, Alireza Esteghamati, Farshad Farzadfar, Derek F J Fay, Valery L Feigin, Abraham Flaxman, Mohammad H Forouzanfar, Atsushi Goto, Mark A Green, Rajeev Gupta, Nima Hafezi-Nejad, Graeme J Hankey, Heather C Harewood, Rasmus Havmoeller, Simon Hay, Lucia Hernandez, Abdullatif Husseini, Bulat T Idrisov, Nayu Ikeda, Farhad Islami, Eiman Jahangir, Simerjot K Jassal, Sun Ha Jee, Mona Jeffreys, Jost B Jonas, Edmond K Kabagambe, Shams Eldin Ali Hassan Khalifa, Andre Pascal Kengne, Yousef Saleh Khader, Young-Ho Khang, Daniel Kim, Ruth W Kimokoti, Jonas M Kinge, Yoshihiro Kokubo, Soewarta Kosen, Gene Kwan, Taavi Lai, Mall Leinsalu, Yichong Li, Xiaofeng Liang, Shiwei Liu, Giancarlo Logroscino, Paulo A Lotufo, Yuan Lu, Jixiang Ma, Nana Kwaku Mainoo, George A Mensah, Tony R Merriman, Ali H Mokdad, Joanna Moschandreas, Mohsen Naghavi, Aliya Naheed, Devina Nand, K M Venkat Narayan, Erica Leigh Nelson, Marian L Neuhouser, Muhammad Imran Nisar, Takayoshi Ohkubo, Samuel O Oti, Andrea Pedroza, Dorairaj Prabhakaran, Nobhojit Roy, Uchechukwu Sampson, Hyeyoung Seo, Sadaf G Sepanlou, Kenji Shibuya, Rahman Shiri, Ivy Shiue, Gitanjali M Singh, Jasvinder A Singh, Vegard Skirbekk, Nicolas J C Stapelberg, Lela Sturua, Bryan L Sykes, Martin Tobias, Bach X Tran, Leonardo Trasande, Hideaki Toyoshima, Steven van de Vijver, Tommi J Vasankari, J Lennert Veerman, Gustavo Velasquez-Melendez, Vasiliy Victorovich Vlassov, Stein Emil Vollset, Theo Vos, Claire Wang, XiaoRong Wang, Elisabete Weiderpass, Andrea Werdecker, Jonathan L Wright, Y Claire Yang, Hiroshi Yatsuya, Jihyun Yoon, Seok-Jun Yoon, Yong Zhao, Maigeng Zhou, Shankuan Zhu, Alan D Lopez, Christopher J L Murray, and Emmanuela Gakidou. 2014. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 384, 9945 (August 2014), 766–781. DOI:https://doi.org/10.1016/S0140-6736(14)60460-8

[55] Martin Porcheron, Joel E. Fischer, and Stuart Reeves. 2021. Pulling Back the Curtain on the Wizards of Oz. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (January 2021), 243:1-243:22. DOI:https://doi.org/10.1145/3432942

[56] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3173574.3174214

[57] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), Association for Computing Machinery, Portland, Oregon, USA, 207–219. DOI:https://doi.org/10.1145/2998181.2998298

[58] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3173574.3174033

[59] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing*

Systems (CHI EA ’17), Association for Computing Machinery, Denver, Colorado, USA, 2853–2859. DOI:https://doi.org/10.1145/3027063.3053246

[60] Pernilla Qvarfordt, Arne Jönsson, and Nils Dahlbäck. 2003. The role of spoken feedback in experiencing multimodal interfaces as human-like. In *Proceedings of the 5th international conference on Multimodal interfaces* (ICMI ’03), Association for Computing Machinery, Vancouver, British Columbia, Canada, 250–257. DOI:https://doi.org/10.1145/958432.958478

[61] Michael Reicherts. 2015. *L’entretien psychologique et le counselling. De l’approche centrée sur la personne aux interventions ciblées*. Edition ZKS-Verlag, Coburg.

[62] Donya Rooein. 2019. Data-Driven Edu Chatbots. In *Companion Proceedings of The 2019 World Wide Web Conference* (WWW ’19), Association for Computing Machinery, San Francisco, USA, 46–49. DOI:https://doi.org/10.1145/3308560.3314191

[63] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI ’19), Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. DOI:https://doi.org/10.1145/3290605.3300587

[64] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (1974), 696–735. DOI:https://doi.org/10.2307/412243

[65] Ayshwarya Saktheeswaran, A. Srinivasan, and J. Stasko. 2020. Touch? Speech? or Touch and Speech? Investigating Multimodal Interaction for Visual Network Exploration and Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2020). DOI:https://doi.org/10.1109/TVCG.2020.2970512

[66] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (UIST ’16), Association for Computing Machinery, New York, NY, USA, 365–377. DOI:https://doi.org/10.1145/2984511.2984588

[67] Vidya Setlur, Melanie Tory, and Alex Djalali. 2019. Inferencing underspecified natural language utterances in visual analysis. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI ’19), Association for Computing Machinery, New York, NY, USA, 40–51. DOI:https://doi.org/10.1145/3301275.3302270

[68] Helen Sharp, Yvonne Rogers, and Jenny Preece. 2019. *Interaction Design: Beyond Human-Computer Interaction. Fifth Edition*. John Wiley, UK.

[69] Bayan Abu Shawar and Eric Atwell. 2007. Fostering language learner autonomy through adaptive conversation tutors. In *Proceedings of the The fourth Corpus Linguistics conference*.

[70] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (April 2020), 495–504. DOI:https://doi.org/10.1080/10447318.2020.1741118

[71] Donggil Song, Eun Young Oh, and Marilyn Rice. 2017. Interacting with a conversational agent system for educational purposes in online courses. In *2017 10th International Conference on Human System Interactions (HSI)*, 78–82. DOI:https://doi.org/10.1109/HSI.2017.8005002

[72] A. Srinivasan and J. Stasko. 2018. Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks. *IEEE Transactions on Visualization and Computer Graphics* (2018). DOI:https://doi.org/10.1109/TVCG.2017.2745219

[73] Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M. Drucker, and Ken Hinckley. 2020. InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI ’20), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3313831.3376782

[74] Arjun Srinivasan and John Stasko. 2017. Natural language interfaces for data analysis with visualization: considering what has and could be asked. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers* (EuroVis ’17), Eurographics Association, Goslar, DEU, 55–59. DOI:https://doi.org/10.2312/eurovisshort.20171133

[75] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A Semi-automated Model for Translating Natural Language Queries into Meaningful Visualizations. In *Smart Graphics*, Springer Berlin Heidelberg, Berlin, Heidelberg, 184–195.
[76] Stergios Tegos and Stavros Demetriadis. 2017. Conversational Agents Improve Peer Learning through Building on Prior Knowledge. *Journal of Educational Technology & Society* 20, 1 (2017), 99–111.
[77] Toyin Tofade, Jamie Elsner, and Stuart T. Haines. 2013. Best Practice Strategies for Effective Use of Questions as a Teaching Tool. *Am J Pharm Educ* 77, 7 (September 2013). DOI:https://doi.org/10.5688/ajpe777155
[78] Melanie Tory and Vidya Setlur. 2019. Do What I Mean, Not What I Say! Design Considerations for Supporting Intent and Context in Analytical Conversation. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 93–103. DOI:https://doi.org/10.1109/VAST47406.2019.8986918
[79] John W. Tukey. 1980. We Need Both Exploratory and Confirmatory. *The American Statistician* 34, 1 (1980), 23–25. DOI:https://doi.org/10.2307/2682991
[80] Ronald D. Vale. 2013. The value of asking questions. *Mol Biol Cell* 24, 6 (March 2013), 680–682. DOI:https://doi.org/10.1091/mbc.E12-09-0660
[81] QianYing Wang and Clifford Nass. 2005. Less visible and wireless: two experiments on the effects of microphone type on users' performance and perception. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '05), Association for Computing Machinery, Portland, Oregon, USA, 809–818. DOI:https://doi.org/10.1145/1054972.1055086
[82] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), Association for Computing Machinery, Honolulu, HI, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376781
[83] Rainer Winkler, Matthias Söllner, Maya Lisa Neuweiler, Flavia Conti Rossini, and Jan Marco Leimeister. 2019. Alexa, Can You Help Us Solve This Problem?: How Conversations With Smart Personal Assistant Tutors Increase Task Group Outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*, ACM Press, Glasgow, Scotland Uk, 1–6. DOI:https://doi.org/10.1145/3290607.3313090
[84] Nicola Yuill and Yvonne Rogers. 2012. Mechanisms for collaboration: A design and evaluation framework for multi-user interfaces. *ACM Trans. Comput.-Hum. Interact.* 19, 1 (May 2012), 1:1-1:25. DOI:https://doi.org/10.1145/2147783.2147784
[85] Nicola Yuill, Yvonne Rogers, and Jochen Rick. 2013. Pass the iPad: collaborative creating and sharing in family groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), Association for Computing Machinery, Paris, France, 941–950. DOI:https://doi.org/10.1145/2470654.2466120
[86] Emily H. van Zee, Marletta Iwasyk, Akiko Kurose, Dorothy Simpson, and Judy Wild. 2001. Student and teacher questioning during conversations about science. *Journal of Research in Science Teaching* 38, 2 (2001), 159–190. DOI:https://doi.org/10.1002/1098-2736(200102)38:2<159::AID-TEA1002>3.0.CO;2-J
[87] Rui Zhang, Stephen North, and Eleftherios Koutsofios. 2010. A comparison of speech and GUI input for navigation in complex visualizations on mobile devices. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services* (MobileHCI '10), Association for Computing Machinery, Lisbon, Portugal, 357–360. DOI:https://doi.org/10.1145/1851600.1851665
[88] SearchIQ by ThoughtSpot. Retrieved August 27, 2020 from https://www.thoughtspot.com/thoughtspot-5-0-searchiq
[89] Einstein Voice. Retrieved August 27, 2020 from https://www.salesforce.com/ca/products/einstein/einstein-voice/
[90] Amazon Polly. Retrieved August 27, 2020 from https://aws.amazon.com/polly/
[91] OBS Studio. Retrieved August 27, 2020 from https://obsproject.com/

Authorship Statement / Statement of Previous Research

I hereby confirm that this manuscript contains original work done by myself and my co-authors that has not been previously published at or submitted to any other venues. We also confirm that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, and that we have no conflict of interest to declare. The paper has no relation to prior papers.

Leon Reicherts