The Effects of an Agent's Clarifying Questioning Behavior During Task-guidance and Collaboration in Virtual Reality

Rohith Venkatakrishnan* University of Central Florida
Roshan Venkatakrishnan† University of Central Florida
Ahmed Rageeb Ahsan‡ University of Florida
Andrew W. Tompkins§ University of Florida
Eric D. Ragan¶ University of Florida
Jaime Ruiz‖ University of Florida

Figure 1: First person perspective of the sorting task. The left image depicts the user grasping the object to be sorted and the right image shows the agent's placement instructions. These images together illustrate the specifics of the task described in Section 4.2.

Abstract

With the rapid growth of artificial intelligence (AI), intelligent agents are becoming increasingly prevalent, guiding people through various tasks. When coupled with virtual and augmented reality devices, these task-guiding agents offer significant potential to train users and assist with performing real-world procedures. Inspired by their promise and imminent ubiquity, we conducted a study focusing on scenarios where an agent may have to ask clarifying questions to its human partner to resolve its uncertainty and proffer correct guidance accordingly. In a mixed factorial experiment, we manipulated the difficulty of the agent's questions and how frequently it asked questions, assessing how task performance levels, subjective perceptions of the agent and workload, and user behaviors were affected. Results suggested that the impact of the agent's question-difficulty was influenced by how often the questions were posed, with counterproductive effects on task performance levels and perceptions emerging, most notably when difficult questions were asked frequently. While frequent questioning can degrade task-performance, it appears to be somewhat tolerated when the demands involved in providing clarifications are low. We discuss how task-guiding virtual agents can question users when uncertain and need clarifying information.

Index Terms: Virtual reality, task guidance, agents, clarifications

1 Introduction

As virtual reality (VR) and augmented reality (AR) technologies continue to mature, their use as platforms for training, assistance, and task support has gained significant traction across domains such as defense [15], aerospace [44, 58], medicine [37], and the Air Force [64]. In these contexts, users are often guided through complex procedures by intelligent agents that instruct, assist, and collaborate as teammates rather than passive tools. Such systems rely on mixed-initiative interaction where both the human and agent proactively contribute by initiating interactions, requests, or information exchange as needed to successfully complete a task [1]. With recent advances in computer vision through artificial intelligence (AI), task-guiding agents are expected to become ubiquitous in extended reality (XR) devices, supporting users during real-world activities such as maintenance, inspection, and repair [7]. These developments position intelligent agents as central components in task-guidance, fundamentally shaping how complex tasks are performed with immersive AR and VR devices.

In task-guided collaboration, the effectiveness of an intelligent agent depends not only on what information it provides, but also how and when that information is exchanged. Prior research has shown that factors such as the accuracy, relevance, modality, and timing of an agent's guidance play a critical role in determining task performance and users' perceptions of the agent [10, 29, 32, 33, 55]. In contexts where poor guidance can lead to costly errors, it is highly important for the agent to ask users clarifying questions when uncertain or in need of additional information to safely guide the execution of the task [28, 56]. In XR-based task-guidance, an agent's uncertainty can arise due to a variety of reasons, including limitations in the device's perception systems where camera vision models fail to accurately detect the state of the world, objects, actions, or steps involved in procedural tasks like mechanical repair, object-sorting, and assembly [69]. In these scenarios (i.e., in uncertainty states), agents must proactively ask users clarifying questions to resolve missing information and adapt their guidance accordingly [56].

Agent-initiated questioning during uncertainty constitutes an important interaction design consideration, as the cognitive effort required to answer an agent's questions and the frequency with which those questions are posed can fundamentally shape how users experience the interaction. While guide agents must ask questions to resolve uncertainty [28], designers may intentionally limit the difficulty or frequency of these clarifications to reduce interruptions,

*e-mail: rohith.venkatakrishnan@ucf.edu
†e-mail: roshan.venkatakrishnan@ucf.edu
‡e-mail: ahmedrageebahsan@ufl.edu
§e-mail: andrewtomp14@gmail.com
¶e-mail: eragan@ufl.edu
‖e-mail: jaime.ruiz@ufl.edu

minimize cognitive load, or preserve users' sense of control during task execution. However, such conservative questioning strategies may leave agents with insufficient information to provide timely or accurate guidance while more demanding or frequent questioning may slow task execution, increase frustration, or discourage collaboration, prompting users to disengage from the agent altogether and rely on alternative forms of task support [31]. Despite the centrality of these trade-offs, no empirical work has evaluated the effects of different questioning behaviors during task-guidance in XR, leaving designers with little guidance on how agents should seek information from users when uncertain. This highlights the need to systematically examine how agent-questioning behavior across varying levels of difficulty and frequency interact to influence user perceptions, task performance, and tendencies to resort to alternative forms of task support during partnered tasks performed with XR guide agents.

In this work, we contribute to the aforementioned problem space, studying how an agent's clarifying questioning behavior affects guided tasks performed in VR. We focused on two aspects of questioning behavior, namely difficulty (i.e., how difficult the agent's questions are) and frequency (i.e., how often the agent asks questions), seeking to understand how they affect perceptions, behaviors, and task performance levels during collaborative tasks that users perform under an agent's guidance. Apropos of these pursuits, we conducted a multifactorial mixed-design study wherein users performed an object-sorting task based on guidance from a disembodied agent whose questioning behaviors were systematically manipulated. We present results and discuss insights derived, explicating how clarifying questions posed by guide agents affect user experience during partnered tasks performed in immersive XR settings.

2 Related Work

2.1 Collaborating with Agents

Collaboration involves working in teams towards a common goal. When the team consists of both humans and AI systems, it can be considered collaboration between humans and what are commonly called agents [30]. In this type of collaboration, agents usually have a certain degree of autonomy, reasoning abilities, awareness of the task's status, and an ability to direct their own behaviors and act, making them effective partners [59]. In human-agent collaborations, agents can vary based on the setting, embodiment, and other interacting characteristics (e.g., speech, gaze, animations, etc.). In digital settings (i.e, virtual worlds, computers) they are usually referred to as virtual or software agents. When agents physically inhabit the real world and are embodied, they can be considered robotic agents [3].

A number of researchers have studied collaborations between humans and robotic agents. In an investigation involving a collaborative construction task [24], users worked with agents that would either instructively guide them or passively support them by assisting with building-pieces only when they were about to make a mistake. It was found that users did not have a preference and adapted to the agent that they were collaborating with. It has also been shown that robotic guide agents verbally assisting with jigsaw puzzle tasks significantly improve performance when performing pointing gestures along with eye gaze [25]. Similarly, agents that gesticulate their hands and end-effectors are evaluated more positively when guiding users to sort items into kitchen cupboards [52].

Several investigations have attempted to study humans collaborating with virtual agents [4, 36, 51]. With respect to the representation of virtual guide agents in VR, it has been shown that realism may have a negligible effect on content comprehension [51], and audio-only guidance without representation can improve task performance but cause poor social presence with the guide [23, 51]. For embodied VR guide agents assisting with maintenance operations on turbines in naval ships, adding non verbal animations has been shown to be more effective than guidance with only verbal instructions [49]. This importance of virtual agents communicating non verbally through gestures and facial expressions was further highlighted in a VR guide agent assisting with property purchase where these behaviors were identified as necessary to fulfill conversational functions during guidance [9]. Researchers have also explored how users collaborate with virtual agents that either exhibit leading or following behaviors in performing an object transportation task in VR [36]. It was found that having to follow agents that exhibit leading behavior can lead to higher levels of perceived workload because users have to adapt to the guidance of the agent rather than vice versa. More recently, it was shown that the intelligence level of a guide agent affects users' perceptions of the agent, social presence, and their performance when being guided by the agent to complete a jigsaw puzzle [11]. Researchers have even studied humans guiding virtual agents to evacuate buildings during fire drills by issuing commands to the agents, finding that the difficulty of the evacuation did not change the number of commands users issued to the agents [35]. Overall, research suggests that when visually embodied, agents must be non-verbally animated to improve guidance-efficacy and that aural guidance without embodiment is preferable for task-performance.

2.2 Task Guidance and Mixed Initiative Interaction

Task guidance can be broadly defined as the provision of information to aid in the performance of a task [2]. Task guidance is effective when it improves the efficiency, accuracy, and ease with which tasks can be completed [17, 43]. Task guidance systems have been studied and applied to a number of contexts that range from simple activities like cooking [18, 19, 47] and household equipment operation [48], to complex commercial operations like aircraft inspection [43] and surgical training [20]. In these systems, humans and agents collaborate with the latter guiding the former towards achieving certain goals.

In the context of task-guidance, collaboration strategies and team cognition are of relevance because they dictate how effectively, efficiently, coordinately, and cooperatively a task is performed [13, 39]. Mixed-initiative interactions refer to a flexible collaboration strategy between human users and agents (or any combination thereof) to perform a task, where each party contributes what they are best suited for at the appropriate time [1]. Control is shared and information can be passed back and forth between members, allowing the team as a whole to perform the task more efficiently [42]. While most definitions of the term vary and are borrowed from the human-computer interaction literature [6], this interaction strategy is largely targeted towards creating helpful problem-solving agents [42]. Researchers have outlined several important characteristics that virtual agent teammates must exhibit in effective mixed-initiative human-agent interactions. These include, but are not limited to, the agent's ability to understand the user's goals and needs, contribute at the right time, engage in dialog with the user when uncertain, and minimize poor actions and incorrect contributions [28]. To exhibit such qualities, agents have to be proactive and initiate interactions with users, especially when seeking information necessary to assist with a task [56]. While intelligent tutoring systems research has evaluated how agents should question when uncertain (supplemental material), the focus is more on assisting learning rather than actually executing tasks.

2.3 Agent-initiated Interactions and Questioning

When interactions between users and agents are initiated by the agent, they are considered agent-initiated interactions. Agents proactively greeting users and enquiring about temperature suitability have been shown to foster more natural interactions with users [46]. LifePod, a smart speaker agent reminds older adults about medication and appointments [63]. Google Home provides timely alerts from users' calendars. Alexa also initiates interactions by offering reminders about events [45]. As such, agent-initiated interactions are common but their effects during task-guidance remains unexplored.

Notwithstanding the usefulness of agents capable of behaving proactively such as those aforementioned, interactions initi-

ated by them during collaboration can be perceived as frustrating, time-wasting, and unnecessary when task performance is disrupted [27, 50]. Despite these concerns, intervening or asking questions during guided collaborative tasks may be necessary to be able to assist users, thus outweighing apprehensions about being interruptive [50]. Researchers have studied agents asking clarifying questions before.Earlier studies [16, 38] directed learner agents to ask questions to human oracles for computer vision models to locate unknown objects in rich image scenes. These works did not focus on how users perceive this questioning during collaboration but rather collected a dataset of question-answer pairs that aids in developing vision model algorithms for object-discovery through dialogue. The closest work studying effects of an agent's questioning behavior lies in online product search where it was found that users may be unwilling to answer more than 11-20 of an agent's clarifying questions, leaving them unanswered when perceived as irrelevant [71], and performing poorly when they are perceived as low quality [70]. Despite such investigations pertaining to web search, there have been no human-centered studies on the effects of a guide-agent's clarifying questioning behaviors during collaborative mixed initiative tasks that users perform with intelligent agents in immersive XR systems.

While embodied conversational agents have been widely used to guide users through tasks in immersive virtual environments (as discussed in Section 2.1), prior work has largely focused on agent realism, interactive animations, social presence, or knowledge depth rather than systematically examining how the characteristics of an agent's clarification questioning affects perceptions during partnered tasks in XR systems. Moreover, research on agents asking clarifying questions remain confined to literature on assisted web search or the system-level implementation of modules for computer vision systems to learn from questions rather than studies focusing on how they affect user experience in immersive systems. Our work directly contributes towards this end, ascertaining how a guide-agent's questioning behaviors affect users' perceptions, behaviors, and performance of partnered tasks performed in XR. We focus on disembodied task-guiding (aural & textual) virtual agents to isolate the effects of clarifying questioning behavior without confounds from embodiment and to reflect smart XR guidance where the agent's visual representation is not always present, feasible, or required.

3 System Description

3.1 Apparatus

The virtual environment (VE) used for this study was built using the Unity 2019.4.28 game engine. The simulation was rendered on an HTC Vive Pro Eye Head Mounted Display (HMD) attached to a computer equipped with an Intel I7-11700KF 3.6 GZ 8 core, 32 GB Ram, and a Nvidia Geforce RTX 3070 graphics card. The trigger button on the Vive Pro's handheld controllers was used to facilitate object interactions in the VE. During pilots, the simulation's framerate was measured and found to be stable and approximately equal to the device's maximum refresh rate (90Hz).

3.2 Virtual Environment

A rectangular virtual room (9.3m by 10m by 5m) was designed to house the object-sorting task described in Section 4.2. The room contained a wooden table, a cubical storage organizer, and a large wall-mounted cork board (5m wide, 2.5m tall) with the task's instruction manual (instructions with legends) pinned to it. The manual was made large to increase the readability of the textual instructions on it. The organizer had eight equally sized shelves (0.9m tall, 0.9m wide, 1m deep), spread evenly over two rows and four columns. Each shelf was assigned a number and its numerical label was positioned at its top-left corner, allowing users to view it as needed while performing the task. The objects to be sorted in the organizer would sequentially spawn atop the table, standing 0.63m beside the organizer and 3.3m from the cork board. These objects were modeled to be primitive cubes (edge 40cm) whose faces featured a randomized and patterned combination of shapes that were used to realize the different levels of the agent's question-difficulty investigated in this study (see Section 4.3.1 and Table 1). Each face contained between zero and five shapes, including triangles, circles, squares, hearts, pentagons, stars, crosses, and arrows. Three different hues were used to color each set of parallel sides of the cube, making the shapes appear more salient. In the scene, participants interacted with the objects using a visual representation of their dominant hands, rendered as a human-hand wearing a glove. Cladding the virtual hand with a glove allowed to prevent feelings of eeriness that could arise from choosing skin textures that are not gender and racially matched [54]. The guiding virtual agent was disembodied and communicated both aurally through voice and visually through textual annotation of the words it spoke. The textual transcription of the agent was anchored to the center of the user's FOV. When asked clarifying questions, users responded verbally and the experimenter determined the correctness of these responses to avoid recognition errors and latency arising from the use of speech-recognition systems, as employed in [53]. The agent's voice was gender-matched to participants and was pre-generated using a text-to-speech engine.

4 Experiment

4.1 Research Question

This work's overarching aim was to answer the question: **"How does a guide agent's clarifying questioning behavior (in uncertainty states) affect perceptions, performance, and behaviors in a collaborative mixed-initiative task in VR?"** We focused on two aspects of the agent's clarifying behavior, relating to how often it asks questions and how difficult these questions are. We utilized an object-sorting task (Sections 4.2 & 4.3) to realize these aims.

4.2 Task

For this experiment, an agent-guided sorting task was conceptualized wherein users had to place thirty objects onto designated shelves of a cubical storage organizer with guidance from a disembodied virtual agent (representing routine tasks encountered in operational sortation workflows). The objects that needed to be placed in the organizer were modeled as primitive cubes, each featuring a different set of shapes on their faces (i.e., some combination of hearts, triangles, arrows, etc.). These objects would sequentially appear on a table in front of the user from where they would have to be picked up and placed in the organizer, one at a time. The agent's role was to simply help users perform this sorting task by informing them of the organizer's shelf number onto which each object had to be placed. Users therefore placed these objects on the different shelves of the organizer, one after another, based on where the agent would tell them to. In order to be helpful, the agent would sometimes need information about the current object that needed sorting, representing a state of uncertainty (e.g., camera vision failing to detect the state of the world, actions, or steps during a task) where additional information is needed before the agent can offer appropriate guidance. Whenever the agent needed this information (i.e, only when it didn't already have all the information it needed about the object), it would ask a task-relevant clarifying question pertaining to the shapes on the object. The difficulty of the question and the frequency with which the agent asked such questions were the two manipulations in the study (Section 4.3). Users could either answer the agent's clarifying questions and receive its guidance on where to place the object or alternatively ignore answering its questions and successfully perform the task without its guidance by consulting a large instruction manual 3.3m away in terms of viewing distance (Figure 1). This sorting task was performed across three phases, one for each of the agent's questioning-frequencies tested (Section 4.3). Overall, every user sorted 30 objects in each phase, accruing a total of 90 object-placements in the study.

Table 1: Examples of questions for each question-difficulty level.

Question-difficulty	Example
Easy	*Is there a star on the object?*
Medium	*How many stars are there on the object?*
Hard	*How many more squares than stars are on the object?*

Each object (trial) to be sorted would first appear on the table from where it could be picked up and placed in the organizer. The agent would promptly inform users about the shelf number onto which the object belonged, simulating contexts where guidance is offered as soon as the agent has the information it needs. Following this guidance, users would place the object in its correct location, initiating the start of the next trial (i.e., spawning the next object on the table). If the object was incorrectly placed on the wrong shelf, the agent would repeat the instruction, allowing users to correct their error. Whenever the agent asked clarifying questions about the object (communicated both aurally and visually as text), users would respond verbally. There was no defined response window (i.e, no task time out), meaning that users could take as long as needed to answer the agent's question or opt out of answering, relying on the instruction manual instead. Thus, if users did not respond to the agent's question, the agent did not re-prompt the question (which was still visible as text in users' FOV) unless explicitly asked to repeat the question again. If the response to the question was correct, the agent would reveal the object's designated shelf number. If the response was incorrect, the agent would ask another question of the same difficulty. We consciously decided against making the agent provide inaccurate guidance based on incorrect responses to its questions because we wanted to ensure that users were made aware of the fact that their verbal response was incorrect. If instead, users were misguided based on an incorrect verbal response, it would not have been obvious that the agent's question was answered incorrectly because the source of error (indicated by the trial not completing) could have also been attributed to the idea of them incorrectly following the agent's guidance (e.g., misunderstanding the shelf number and placing the object based on this misunderstanding). Additionally, the agent did not provide correct guidance to an incorrect response to ensure that there were no avenues for simulation-based exploits that users could employ in attempts to game the system and simply proceed through the experiment, answering every question arbitrarily. The agent always communicated both aurally (through voice) and visually (through textual annotation in the center of the FOV) to make it more accessible and representative of XR scenarios.

Multimodal feedback and visualizations were designed to make for intuitive interaction design. When users' virtual hand was in proximity of the object, a yellow highlight would appear around the object along with haptic feedback on the controller, implying the affordance of graspability. Upon grasping the object, the virtual hand disappeared, indicating that the object could be manipulated. This grasping visualization technique was chosen to make the content of the cube's faces salient when the object was being held. As some trials required answering questions about the object, visualizing the grasping hand would have obscured its contents (i.e., the shapes on its faces), making such questions cumbersome to answer. When an object was correctly placed in its designated location, a chiming sound was deployed and the object would disappear after two seconds, indicating the successful completion of the trial and the initiation of the next. When an object was placed in an incorrect shelf, a ding sound was played, the agent repeated its instruction/question, and the object remained in its location, indicating that the trial was not completed successfully. The visual and auditory feedback pertaining to any event was presented simultaneously, providing rich multi-modal feedback that was clearly indicative of the different events that transpired during the task.

Figure 2: Example of an object that users had to sort. Objects only varied in terms of the patterned combination of shapes on their sides.

4.3 Study Design

To empirically evaluate how a virtual agent's information-seeking behavior affects users' performance, perceptions and behaviors during task-guided collaboration in VR, we conceptualized a study, focusing on aspects related to the agent's questioning behavior. Specifically, we employed a 3 (question difficulty) $\times$ 3 (questioning-frequency) mixed multi-factorial design, manipulating the former as a between-participants factor across three experimental conditions: (1) Easy (agent's questions are relatively easy to answer); (2) Medium (agent's questions are of relatively medium difficulty); (3) Hard (agent's questions are relatively hard to answer). Users in each condition performed the object-sorting task described in Section 4.2 across three phases of the agent's questioning-frequency (i.e., how frequently the agent asked questions), constituting a within-participants manipulation. To control for potential order effects, a Balanced Latin Square design was employed, counterbalancing the order of these questioning-frequency phases. Participants hence experienced one possible order (out of a total of 6 possible orders) of these frequencies. In each phase, users performed 30 trials (sorting 30 objects), thus accruing a total of 90 trials across the entire experiment. The specifics of the manipulations are detailed subsequently.

4.3.1 Question-difficulty

To manipulate the difficulty of the questions posed by the agent, we created three categories of questions about the object that needed sorting, each differing in terms of the intrinsic load required to answer the question correctly [60]. Along these lines, easy questions involved simple yes or no responses (e.g, *"Is there a star on the object?"*) while medium difficulty questions were more load-inducing, requiring users to count the number of shapes (e.g, *"how many stars are there on the object?"*). Hard questions involved the greatest difficulty with users having to count shapes and then perform an addition or subtraction operation (e.g., *"how many more squares than stars are on the object?"*). Every question had a predetermined correct answer based on the object's shape pattern, requiring users to visually inspect the object to determine this answer that they could then provide verbally. Any other answer was deemed incorrect. Figure 2 and Table 1 together illustrate the different levels of question-difficulty for an object (primitive cube) used in the study.

4.3.2 Questioning-frequency

The questioning-frequency represents how frequently the agent would need and ask for information about the object that needed sorting. The three levels of questioning-frequency tested were namely 'Less', 'Somewhat', and 'More Frequent' questioning. Numerically, the frequency was defined as the number of questions the agent asked per phase, each comprising 30 trials (object-placements). The values of questioning-frequency studied were 30, 10, and 6. A frequency of 30 meant that a question was asked on every trial (30 trials / 30 questions = 1) while a frequency of 6 meant that a question was asked on every fifth trial (30 trials / 6 questions = 5). Similarly, a frequency of 10 meant that the agent asked the user a question on every third trial. Thus a higher questioning-frequency implies that the agent asked questions more frequently.

4.4 Measures

Duration: The amount of time taken to sort all thirty objects on their respective shelves was measured in seconds for each phase.
Resorting to the Alternative: In each phase, the number of times users ignored answering the agent's questions, instead performing the task successfully with the instruction manual was measured.
Perceived Workload: After each phase, the perceived levels of workload associated with performing the task was measured using the NASA TLX questionnaire [26]. This questionnaire measures workload across six dimensions, namely mental demand, physical demand, temporal demand, performance, effort, and frustration.
Perceptions of the Agent: After each phase, the Agent Rating Questionnaire (ARQ) [66] was administered to evaluate subjective perceptions of the agent. With ordinal items, the ARQ measures perceptions across five dimensions: helpfulness, reliability, likability, willingness to interact again, and appropriateness of behavior.

4.5 Hypotheses

We developed the following hypotheses for this study based on work discussed in Section 2 and subsequent portions of this section.

- **H1:** The duration for task-completion will increase with increases in the agent's (a) question-difficulty & (b) questioning-frequency.
- **H2:** The levels of workload perceived will increase with increases in the agent's (a) question-difficulty & (b) questioning-frequency.
- **H3:** The perceptions of the agent will deteriorate with increases in its (a) question-difficulty & (b) questioning-frequency.
- **H4:** Higher question-difficulty will increase the number of times users resort to using the instruction manual to perform the task.

Prior work and intuition suggest that difficult questions generally take longer to answer correctly and impose greater cognitive demand [57]. More frequent questions create more opportunities for delays and cognitive strain to accumulate. It is hence reasonable to expect that increased question-difficulty and questioning-frequency will prolong task duration and elevate perceived workload. Frequent questioning, especially difficult ones requiring effortful responses can also frustrate users, leading to less favorable perceptions of the agent [8, 12, 62]. Additionally, the negative effects of increased question-difficulty are likely to be more pronounced when questioning occurs more frequently, suggesting an interaction effect. In terms of the tendency to disengage with the agent and instead use instruction manual, users may do this more often when confronted with hard questions because difficult questions often cause people to abandon attempts to directly provide answers [40, 41].

4.6 Participants

An apriori power analysis using G*Power revealed that for a study of 3 question-difficulty levels manipulated between-subjects and 3 questioning-frequency levels manipulated within-subjects, considering α = 0.05, power (1-β) = 0.95, a medium effect size f= 0.25, correlation among repeated measures of 0.5, and sphericity assumption, the estimated sample size was 54. Thus, 54 participants were recruited for this Institutional Review Board (IRB) approved study (18 per difficulty condition). Participants were recruited from intro level courses from the university through email. Their ages ranged from 18 to 54 (M = 21.86, SD = 5.37). Regarding gender, 20 identified as female, 32 as male, and 2 as gender non-conforming. All participants had normal or corrected-to-normal (20/20) vision.

4.7 Procedure

Upon arrival at the laboratory, participants read and signed an informed consent form. After consenting, participants filled out a demographics questionnaire that included information about their backgrounds. The experimenter then explained the task (Section 4.2), detailing the logistics involved in using the hardware devices (controllers and HMD). Following this, users donned the HMD (adjusted for their IPD) and engaged in a tutorial wherein they learned to perform the task both with and without the agent's guidance over 9 trials. In this tutorial, they began by simply following the agent's instructions and sorting 3 objects into the organizer, familiarizing them with the task mechanics and the agent's guidance behavior. They further encountered 3 occasions where the agent was uncertain and asked clarifying questions (difficulty matched to users' assigned condition), learning to verbally answer with the information that the agent needed to continue its guidance. Next, participants performed the task without the agent's guidance until they correctly sorted 3 objects with only the manual after first confirming the readability and legibility of the instruction manual from their vantage point. This marked the end of the tutorial phase at which point users demonstrated and reported proficiency in performing the task both with and without the agent's guidance. Participants then began the experiment, performing the sorting task across the three questioning-frequency phases of the study. At the end of each phase, they responded to a set of questionnaires that included the ARQ [66] and NASA TLX [26]. Before proceeding to the next phase, users took breaks (minimum of five minutes) until they reported being at ease and well-rested. When all three phases of the study were completed, participants engaged in a semi-structured interview with the experimenter, consisting of questions related to their perceptions of the experience, strategies they used, opinions about the agent, and aspects they found challenging. This marked the end of the study and participants were then compensated with $20 USD.

5 Results

We conducted all statistical analyses in R (version 4.4.0) using the packages `dplyr`, `ARTool`, `emmeans`, `tibble`, `zoib`, and `ZIBR`. For dependent variables measured on a continuous scale (duration, workload), we first tested for normality using Shapiro-Wilk's test and found that the data distributions were significantly non-normal. Given that assumptions for parametric analyses were violated, applying aligned rank transformations (ART) to the data is necessary for conducting multifactorial non-parametric analyses including random effects for repeated measures [68]. Unlike standard non-parametric tests that cannot handle interaction effects, or mixed-effects models that assume normality of residuals [72], the ART procedure allows for rigorous analyses of both main and interaction effects in non-normal, repeated-measures data [68]. Furthermore the ordinal and Likert-scaled nature of the ARQ questionnaire necessitates this analysis procedure, as employed in [66]. Thus, we conducted ART-ANOVA analyses on our dependent measures, aiming to explore the effects question-difficulty, questioning-frequency, and their interaction effects. In each analysis, significant effects and their posthoc tests are presented with p values. Post hoc pairwise comparisons between levels of the manipulations were conducted using ART-Contrasts. The test statistic and measures of effect size are presented in Table 2. The ${\eta_p}^2$ represents the proportion of variance uniquely explained by a given variable, relative to the total variance remaining after controlling for other variables in the model.

5.1 Duration

With task duration, we found significant main effects of question-difficulty ($p < .001$) and questioning-frequency ($p < .001$). Table 2 summarizes results for both factors and their interaction. With higher difficulty, it took longer to complete the task. Similar trends were observed with more frequent questioning. There was also a significant interaction, indicating that frequency moderated the effect of difficulty ($p < .001$). The magnitude of differences in duration across the levels of difficulty varied across levels of frequency (see Figures 3a and 4a, which show task duration in minutes).

Table 2: Summary of the ART-ANOVA analyses on the measures described in section 4.4. The table presents the statistics, p-values, and effect sizes (η_p^2) of the main effects of both question-difficulty and questioning-frequency along with their interaction effects. For each of these main effects, the mean and standard deviation of the different levels of the manipulations are listed in increasing order (i.e., Easy, Medium, & Hard for question-difficulty levels; Less, Somewhat, and More Frequent [LF, SF, &MF] for questioning-frequency levels). Two color-blind accessible palettes are used to depict post-hoc tests, one for each manipulated factor. Pairwise comparisons are indicated (in the M(SD) column) in each row wherein cells with different hues imply that levels of a factor are significantly different from each other. For significant interaction effects, post-hocs are depicted in Figures 3 and 4. Note that post-hocs presented are based on the ART ANOVA analyses results detailed in Section 5.

Measure	Dimension	Result			M(SD)		
		Effect	Statistic	η_p^2	Easy/LF	Medium/SF	Hard/MF
NASA TLX [26]	Mental Demand	Difficulty	F(2,51) = 14.56, p <.001	.363	20.81(22.17)	25.57(22.00)	51.63(24.05)
		Frequency	F(2,102) = 17.98, p <.001	.261	26.13(24.29)	31.09(27.22)	40.80(25.89)
		Interaction	F(4,102) = 1.53, p = .19	.057		-	
	Physical Demand	Difficulty	F(2,51) = 2.19, p = .12	.079	11.04(13.53)	18.81(17.33)	17.65(19.39)
		Frequency	F(2,102) = 0.35, p = .70	.007	16.46(18.18)	15.28(16.96)	15.76(16.62)
		Interaction	F(4,102) = 0.46, p = .76	.018		-	
	Temporal Demand	Difficulty	F(2,51) = 2.16, p = .13	.078	31.37(26.44)	31.61(23.6)	44.81(23.51)
		Frequency	F(2,102) = 0.56, p = .57	.010	37.48(26.37)	35.50(25.31)	34.81(24.28)
		Interaction	F(4,102) = 1.23, p = .30	.046		-	
	Performance	Difficulty	F(2,51) = 4.91, p = .01	.161	91.67(9.65)	83.19(15.73)	79.93(16.51)
		Frequency	F(2,102) = 11.27, p <.001	.181	86.72(16.79)	87.41(12.25)	80.65(15.08)
		Interaction	F(4,102) = 0.46, p = .77	.018		-	
	Effort	Difficulty	F(2,51) = 14.84, p <.001	.368	See Figures 3b and 4b		
		Frequency	F(2,102) = 19.49, p <.001	.276			
		Interaction	F(4,102) = 3.92, p<.01	.133			
	Frustration	Difficulty	F(2,51) = 4.79, p = .01	.158	See Figures 3c and 4c		
		Frequency	F(2,102) = 11.27, p <.001	.181			
		Interaction	F(4,102) = 5.16, p <.001	.168			
Duration	Duration in minutes	Difficulty	F(2,51) = 49.94, p <.001	.662	See Figures 3a and 4a		
		Frequency	F(2,102) = 205.91, p <.001	.801			
		Interaction	F(4,102) = 26.22, p <.001	.506			
ARQ [66]	Helpfulness	Difficulty	F(2,51) = 4.06, p = .02	.137	3.78(1.76)	3.91(1.74)	2.83(1.76)
		Frequency	F(2,102) = 2.79, p = .07	.051	3.56(1.87)	3.48(1.86)	3.48(1.71)
		Interaction	F(4,102) = 1.40, p = .24	.052		-	
	Relatability	Difficulty	F(2,51) = 1.38, p = .26	.051	2.54(1.38)	2.48(1.48)	1.83(1.13)
		Frequency	F(2,102) = 1.77, p = .18	.033	2.43(1.39)	2.19(1.35)	2.24(1.37)
		Interaction	F(4,102) = 0.64, p = .64	.024		-	
	Appropriate Behavior	Difficulty	F(2,51) = 0.01, p = .99	.000	No significant posthocs		
		Frequency	F(2,102) = 2.33, p = .10	.043			
		Interaction	F(4,102) = 2.83, p = .02	.099			
	Likeability	Difficulty	F(2,51) = 0.20, p = .82	.008	2.96(1.73)	3.19(1.96)	2.65(1.63)
		Frequency	F(2,102) = 0.20, p = .81	.004	2.93(1.81)	2.89(1.83)	2.98(1.73)
		Interaction	F(4,102) = 0.41, p = .79	.015		-	
	Willingness to Interact Again	Difficulty	F(2,51) = 0.47, p = .63	.018	3.5(1.88)	3.41(1.96)	2.85(1.79)
		Frequency	F(2,102) = 2.99, p = .06	.055	3.35(1.91)	3.28(1.89)	3.13(1.89)
		Interaction	F(4,102) = 1.08, p = .37	.041		-	

5.2 Perceived Workload - NASA-TLX

We analyzed the effects of our manipulations on each of the six dimensions of the NASA-TLX workload questionnaire, namely mental demand, physical demand, temporal demand, performance, effort and frustration. We found no significant main or interaction effects of the manipulations on physical and temporal demand.

In terms of mental demand, we found main effects of both question-difficulty ($p < .001$) and questioning-frequency ($p < .001$). Hard questions were perceived as significantly more mentally demanding than both medium ($p < .001$) and easy-difficulty ($p < .001$) questions. Mental demand was also significantly higher when the agent asked questions more frequently than when it did somewhat ($p < .001$) or less frequently ($p < .001$). No other significant differences were found between levels of the factors on mental demand and there were no significant interaction effects.

In terms of perceived performance, we found main effects of both question-difficulty ($p = .01$) and questioning-frequency ($p < .001$). Perceived performance significantly deteriorated with hard questions compared to easy ones($p = .01$). More frequent questions were associated with lower levels of perceived performance than both somewhat ($p < .001$) and less frequent questioning ($p < .001$). No other differences in perceived performance were found between levels of the factors and there were no significant interaction effects.

Considering the perceived levels of effort required to perform the task, we found significant main effects of both question-difficulty ($p < .001$) and questioning-frequency ($p < .001$). Easy ques-

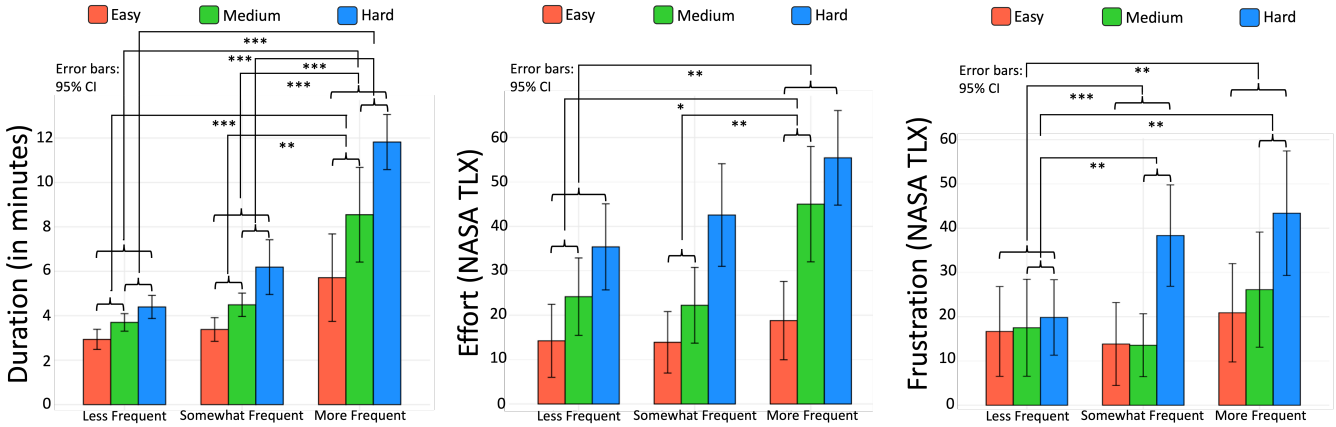

Figure 3: Interaction effects showing how questioning-frequency moderated the effect of question-difficulty on (a) task duration, (b) perceived effort required, and (c) perceived frustration. Each sub-figure shows comparisons of the magnitudes of differences (in a given measure) between question-difficulty levels across the three levels of questioning-frequency. Note that annotations on the figures are based on the ART-ANOVA analyses detailed in Section 5 and Table 2. Significance levels: '*' denotes $p < 0.05$, '**' denotes $p < 0.01$, and '***' denotes $p < 0.001$.

tions were reported to require significantly less effort than both hard ($p < .001$) and medium-difficulty questions ($p = .01$). The agent asking questions more frequently was associated with significantly higher levels of effort than both somewhat ($p < .001$) and less frequent questioning ($p < .001$). More interestingly, we found a significant interaction effect, indicating that the questioning-frequency moderated the effect of question-difficulty on perceived effort ($p = .005$). Along these lines, the magnitude of differences in perceived effort between the different levels of question-difficulty varied across the levels of questioning-frequency (Figure 3b).

In terms of frustration, we found significant main effects of both question-difficulty ($p = .012$) and questioning-frequency ($p < .001$). Users were more frustrated when the agent asked hard questions than when it asked easy or medium-difficulty questions. Frustration levels dropped when the agent asked questions less frequently than when it did somewhat or more frequently. We also found a significant interaction effect, indicating that frequency moderated the effect of difficulty on perceived frustration ($p < .001$). The magnitude of differences in frustration between the different levels of question-difficulty varied across levels of questioning-frequency (Figure 3c).

5.3 Agent Rating Questionnaire (ARQ)

We analyzed effects of the factors on ARQ dimensions (Section 4.4). No significant main or interaction effects were found for relatability, likeability, or willingness to interact with the agent again.

In terms of helpfulness, we found a main effect of question-difficulty ($p = .023$). The helpfulness of the agent was rated significantly higher when it asked easy questions than hard ones ($p = .03$). No other significant differences in helpfulness were found between levels of question-difficulty. There were no other significant main or interaction effects on this measure. In terms of perceptions of the agent behaving appropriately, there was a significant interaction effect, indicating that frequency moderated the effect of difficulty on this measure ($p = .02$). However, posthoc analyses did not show any significant differences across the different pairs of difficulty-frequency combinations. This pattern can arise when interaction effects capture broader trends not reflected in individual pairwise contrasts due to conservative multiple-comparison corrections [22].

5.4 Resorting to the Alternative

Since the dependent variable is a count (i.e., the number of times participants defaulted to the instruction manual), Poisson regression is appropriate [14]. However, directly analyzing this count would be misleading because the number of opportunities to seek the alternative varied with questioning-frequency. More frequent questioning creates more opportunities to use the manual. To account for this, we computed the percentage of times the manual was sought. As this produces values between 0 and 1, beta regression is favored [21]. Since many values were zero, we used a zero-inflated beta regression with a complementary log-log transformation [61]. This model first estimates whether a user sought the manual (zero vs. non-zero), then models the extent of seeking behavior among those who did. We assessed data nesting by calculating the intraclass correlation coefficient (ICC) from a null model. The ICC was 0.116, showing 11.6% of variance was due to user differences, violating independence assumptions. A multilevel model was thus used with a random intercept for Participant ID. The model with main effects (AIC = 153.6, df = 8) fit marginally better than the null model (AIC = 154, df = 4), $\chi^2 = 8.34$. The R^2 change was .057, indicating a 5.7% increase in explained variance. The zero-inflation intercept was significant ($B = 0.74$, $SE = 0.17$, $z = 4.45$, $p < .001$), suggesting a strong baseline tendency to avoid using the manual. In the conditional model, users were significantly more likely to resort to the manual when answering hard questions ($B = 1.17$, $SE = 0.59$, $z = 1.97$, $p = .049$) than easy ones, with the estimated probability rising from 23.8% to 58.6%. Medium-difficulty questions did not affect this likelihood ($B = 0.81$, $SE = 0.64$, $z = 1.28$, $p = .20$), and no significant differences emerged across frequency levels.

6 Discussion

6.1 Task Performance

The statistical analyses of the time taken to complete the task revealed significant main effects of both question-difficulty and questioning-frequency, indicating that the agent asking harder questions or asking questions more often generally led to greater delays in users completing the sorting task (Table 2). These results offer support for hypotheses **H1a** & **H1b** understandably because answering harder questions like those asked by the agent typically takes longer [57], and frequent questioning causes disruptions that compound delays over time [34]. We also found a significant interaction effect between these factors, suggesting that the agent's questioning-frequency moderated the effect of its questions' difficulties on how long it took users to complete the task. As seen in Figure 3a, the differences in task-completion times across the three levels of the agent's question-difficulty become significantly more pronounced as the frequency of its questions increases. When the agent asks questions less frequently, the delays caused by answering harder questions are relatively modest, whereas the cumulative delays from harder questions become more substantial with more frequent questioning. This indicates that increases in the difficulty of an agent's

Duration	Easy-Less Frequent	Easy-Somewhat Frequent	Easy-More Frequent	Medium-Less Frequent	Medium-Somewhat Frequent	Medium-More Frequent	Hard-Less Frequent	Hard-Somewhat Frequent	Hard-More Frequent
Easy-Less Frequent	-	0.44	**2.77** ***	**0.76** *	**1.55** ***	**5.61** ***	**1.46** ***	**3.25** ***	**8.88** ***
Easy-Somewhat Frequent	-0.44	-	**2.33** ***	0.32	**1.11** ***	**5.17** ***	**1.02** **	**2.81** ***	**8.44** ***
Easy-More Frequent	**-2.77** ***	**-2.33** ***	-	**-2.01** **	-1.22	**2.84** ***	-1.31	0.48	**6.11** ***
Medium-Less Frequent	**-0.76** *	-0.32	**2.01** **	-	**0.79** **	**4.85** ***	0.70	**2.49** ***	**8.12** ***
Medium-Somewhat Frequent	**-1.55** ***	**-1.11** ***	1.22	**-0.79** **	-	**4.06** ***	-0.09	**1.70** *	**7.33** ***
Medium-More Frequent	**-5.61** ***	**-5.17** ***	**-2.84** ***	**-4.85** ***	**-4.06** ***	-	**-4.15** ***	-2.36	3.27
Hard-Less Frequent	**-1.46** ***	**-1.02** **	1.31	-0.70	0.09	**4.15** ***	-	**1.79** ***	**7.42** ***
Hard-Somewhat Frequent	**-3.25** ***	**-2.81** ***	-0.48	**-2.49** ***	**-1.70** *	2.36	**-1.79** ***	-	**5.63** ***
Hard-More Frequent	**-8.88** ***	**-8.44** ***	**-6.11** ***	**-8.12** ***	**-7.33** ***	-3.27	**-7.42** ***	**-5.63** ***	-

(a) Duration

Effort	Easy-Less Frequent	Easy-Somewhat Frequent	Easy-More Frequent	Medium-Less Frequent	Medium-Somewhat Frequent	Medium-More Frequent	Hard-Less Frequent	Hard-Somewhat Frequent	Hard-More Frequent
Easy-Less Frequent	-	-0.33	4.56	9.95	8	**30.78** ***	**21.17** **	**28.34** ***	**41.22** ***
Easy-Somewhat Frequent	0.33	-	4.89	10.28	8.33	**31.11** ***	**21.5** **	**28.67** ***	**41.55** ***
Easy-More Frequent	-4.56	-4.89	-	5.39	3.44	**26.22** **	16.61	**23.78** **	**36.66** ***
Medium-Less Frequent	-9.95	-10.28	-5.39	-	-1.95	**20.83** ***	11.22	18.39	**31.27** **
Medium-Somewhat Frequent	-8	-8.33	-3.44	1.95	-	**22.78** ***	13.17	20.34	**33.22** ***
Medium-More Frequent	**-30.78** ***	**-31.11** ***	**-26.22** **	**-20.83** ***	**-22.78** ***	-	-9.61	-2.44	10.44
Hard-Less Frequent	**-21.17** **	**-21.5** **	-16.61	-11.22	-13.17	9.61	-	7.17	**20.05** **
Hard-Somewhat Frequent	**-28.34** ***	**-28.67** ***	**-23.78** **	-18.39	-20.34	2.44	-7.17	-	12.88
Hard-More Frequent	**-41.22** ***	**-41.55** ***	**-36.66** ***	**-31.27** **	**-33.22** ***	-10.44	**-20.05** **	-12.88	-

(b) Effort

Frustration	Easy-Less Frequent	Easy-Somewhat Frequent	Easy-More Frequent	Medium-Less Frequent	Medium-Somewhat Frequent	Medium-More Frequent	Hard-Less Frequent	Hard-Somewhat Frequent	Hard-More Frequent
Easy-Less Frequent	-	-2.84	4.22	0.83	-3.11	9.44	3.16	21.66	26.72
Easy-Somewhat Frequent	2.84	-	7.06	3.67	-0.27	12.28	6.00	**24.5** *	**29.56** *
Easy-More Frequent	-4.22	-7.06	-	-3.39	-7.33	5.22	-1.06	17.44	22.5
Medium-Less Frequent	-0.83	-3.67	3.39	-	-3.94	8.61	2.33	**20.83** *	**25.89** *
Medium-Somewhat Frequent	3.11	0.27	7.33	3.94	-	12.55	6.27	**24.77** *	**29.83** *
Medium-More Frequent	-9.44	-12.28	-5.22	-8.61	-12.55	-	-6.28	12.22	17.28
Hard-Less Frequent	-3.16	-6.00	1.06	-2.33	-6.27	6.28	-	**18.5** *	**23.56** *
Hard-Somewhat Frequent	-21.66	**-24.5** *	-17.44	**-20.83** *	**-24.77** *	-12.22	**-18.5** *	-	5.06
Hard-More Frequent	-26.72	**-29.56** *	-22.5	**-25.89** *	**-29.83** *	-17.28	**-23.56** *	-5.06	-

(c) Frustration

Figure 4: Post-hoc analyses of significant interactions between question-difficulty and questioning-frequency on (a) task duration, (b) perceived effort required, and (c) perceived frustration. Each cell in a table represents the differences (col-row) between the means of corresponding difficulty-frequency configurations for that measure. Note that annotated significance levels (in red) in each table are based on the ART-ANOVA analyses detailed in Section 5 and Table 2. Significance levels: '*' denotes $p < 0.05$, '**' denotes $p < 0.01$, and '***' denotes $p < 0.001$.

questions posed during task-guidance lead to longer task completion times, especially when these questions are asked more frequently.

A noteworthy discovery we made from our analyses was that there were no statistically significant differences in task-durations between conditions where the agent asked hard questions less frequently or easy questions more frequently (Figure 4a). This implies that during collaboration, guide-agents that need information from users can maintain task performance efficiency, either asking low-effort questions frequently or more demanding questions infrequently. Our finding here highlights that task-guiding agents in immersive virtual environments can adopt either of these clarifying questioning behaviors in uncertainty states without differentially affecting task-performance efficiency. This being said, what are considered demanding and low-effort questions for one person need not apply to another, making it important for agents to tailor their questioning strategies based on users' preferences and cognitive loads. Future research should investigate how to use real-time data to classify cognitive load for agents to accordingly adapt their interaction strategies to maintain task performance levels without escalating workload.

6.2 Perceived Workload

We found statistically significant main effects of both question-difficulty and questioning-frequency on several dimensions of the NASA-TLX workload questionnaire (Table 2). The trends observed suggested that harder questions and more frequent questioning were generally associated with weakened perceptions of performance and greater levels of mental demand, frustration, and effort required to perform the task. These results offer support for hypotheses **H2a** & **H2b**, and are intuitive because challenging and frequent interactions generally impose greater demand on cognitive resources, thereby heightening perceived effort and frustration [5, 67]. Given the nature of the task, it was not surprising that perceived levels of physical and temporal demand were relatively unaffected by the manipulations. With users simply sorting virtual objects onto the nearby organizer based on guidance from the agent, the task did not induce any physical or time-related pressure, explaining the lack of statistically significant effects on these dimensions of workload.

Notably, we found significant interaction effects on the dimensions of frustration and effort, indicating that the impact of question-difficulty depended on how frequently the agent asked questions. As can be seen in Figures 3b and 3c, the differences in these aspects across the levels of question-difficulty become more apparent as the agent asks questions more frequently. For instance, when questioned less frequently, users' frustration levels did not significantly increase when confronted by more challenging questions from the agent (Figure 4c). With somewhat and more frequent questioning, however, increases in question complexity tended to escalate frustration. Similarly, increases in perceived levels of effort from having to answer harder questions became more pronounced as the agent asked questions more frequently. While we did not observe these type of interaction effect trends on all dimensions of workload, our results suggest that the effects of more demanding agent interactions become more pronounced with increased interaction-frequency.

Our analyses revealed nuances associated with user-tolerances to providing clarifications. We found that hard questions frustrated users to a significantly smaller extent when asked less frequently than when asked somewhat or more frequently (Figures 3c & 4c). This suggests that users tolerate answering a guide-agent's challenging questions without becoming frustrated if they are well spaced out. With respect to effort, we observed that having to answer hard questions somewhat frequently was rated significantly more demanding than having to answer easy questions more frequently (Figures 3b & 4b). Thus, the effort involved in frequently responding to simple clarification questions appears to be less than the effort needed to answer more demanding questions even if asked less frequently. Taken together with the duration results showing that even simple clarifications prolong task duration when needed frequently (Figure 4a), these findings suggest frequent questioning may be preferable even at the expense of performance. However, the tolerance for performance detriments is likely context-dependent, requiring a balance between optimizing for performance and cognitive tolerance.

6.3 Perceptions of the Agent

Regarding users' perceptions of the guide-agent, the analyses revealed a statistically significant main effect of question-difficulty on how helpful the agent was perceived to be (Table 2). When the agent asked more difficult questions, users generally perceived it as less helpful. Neither factor produced significant main effects on any other dimensions of the ARQ but generally indicated that increases in them could lead to poorer perceptions of the agent. We also found a significant interaction effect between these factors on how appropriately users perceived the agent's behavior to be. However, posthoc analyses revealed no significant differences across the different pairs of difficulty-frequency combinations. This kind of pattern is not uncommon and can arise when an interaction effect captures a broader and complex trend in the data that is not reflected in any single pairwise contrast due to the conservative nature of multiple comparison corrections [22]. Nevertheless, it appears that users' per-

ceptions of a guide-agent's behavioral appropriateness are shaped by some nuanced relationship in its questioning tendencies, an avenue that warrants further investigation. Notwithstanding the collective evidence offering only marginal support for hypothesis **H3a** and no support for **H3b**, we believe that increases in these aspects of an agent's questioning behaviors are likely to affect perceptions based on findings from educational research [62]. While the purpose of questioning in a classroom setting starkly differs from why an agent would ask questions during task-guidance, their effects when highly challenging or excessive in number may be similar, leading to passivity, negatively valenced emotions, and perceptions [8, 12]. There, however, may be mission-critical scenarios (e.g., cockpit repair during battle) in which an AR guide-agent has to resolve crucial uncertainties during collaboration, potentially permitting such extreme behaviors even at the cost of it being perceived poorly.

6.4 Resorting to Alternative Forms of Task Support

Our analyses provided insights into when users chose to ignore the agent's questions and instead perform the task, consulting the instruction manual. As described in Section 5.4, a significant majority of users opted to work with the agent regardless of its questioning behavior, avoiding the instruction manual altogether. For those users who did use the manual, we found a significant effect of question-difficulty, showing that the agent asking hard questions led to an increase in the probability of them seeking the alternative. The estimated probability of resorting to the instruction manual increased from about 24% with easy questions to 59% with hard questions. While such increases with medium-difficulty questions did not reach significance, the results generally indicate that facing more challenging questions from an agent during task-guidance can turn users away from the agent, making them seek other forms of task support. These trends offer partial support for hypothesis **H4**. Interestingly, even among users that defaulted to the manual, there was no significant effect of questioning-frequency, suggesting that frequent clarifications are better tolerated than demanding ones. Guide agents that need information should hence attempt to chunk challenging questions into low-complexity frequent interactions to dissuade users from disengaging during collaboration. Our results must be interpreted, noting the model's modest effect size in explaining users' propensities to resort to the manual. One's natural tendency to disengage from an agent when asked demanding clarifying questions thus still remains an important determinant of whether or not they will seek alternative forms of task support. The difficulty of providing clarifications, however, appears to be a stronger predictor of this behavior than how often clarifications have to be provided.

6.5 Novel Insights and Design Recommendations

Beyond confirming expected detriments from guide agents asking more frequent and difficult clarifying questions, our results reveal several less obvious empirically derived insights into how these factors interact. First, difficult questions posed infrequently cause performance tradeoffs comparable to easy ones posed more frequently, suggesting that these factors should be considered jointly rather than independently. Second, easy clarifications even if needed frequently require less effort than infrequent hard ones, indicating that reducing difficulty may be more effective than merely trying to reduce clarification frequency. Third, compared to frequency, the difficulty of providing clarifications appears to more strongly predict one's behavioral tendency to disengage with a guide agent and seek alternative forms of task support. Overall, the notable theme emerging from these insights is that when task-guiding agents require clarifying information from users to resolve their uncertainties, it appears preferable for them to reduce the complexity of their questions even if doing so means asking more questions. Accordingly, an actionable step that manufacturers of XR devices equipped with task-guiding agents should take is to ensure that their agents prioritize decomposing complex clarification questions into simpler, lower-effort ones, even if doing so requires more frequent interactions. From the standpoint of application environment designers that strategically integrate XR device-based guidance into their workflows, objects should be equipped with additional attributes to allow agents resolve uncertainties using simpler clarifications. In warehouse settings, for example, adding supplementary labels, markings, or metadata on the objects being sorted can create more ways for XR guide agents to resolve uncertainty through easier clarifications.

6.6 Scope and Limitations

This work on studying the consequential effects of a guide agent's questioning-behavior was conducted in VR. Despite this, our findings should largely apply to settings involving virtual task-guiding agents in general (e.g., in AR & MR) because the core interaction dynamics in human-agent teams are largely modality-agnostic [65], making our insights generally applicable to immersive mediums. This being said, embodied agents (e.g., virtual characters or robots) can introduce social cues that can change the dynamics of collaboration. It is therefore appropriate to limit the scope of this work to scenarios where guide agents are disembodied. Furthermore, our task was designed to reflect routine guidance workflows (e.g., warehouse object sortation), and our findings may not fully generalize to more complex guidance scenarios such as repair, inspection, or assembly. Our study may also be limited by a design asymmetry wherein performing the task with the manual, despite being readable, may have been markedly more inconvenient than working with the agent. While such asymmetries justify the use of guide agents, its magnitude being large likely explains the manual's limited use, requiring research reducing this disparity. It's also worth noting that the agent's behaviors operated with certain regularities. For instance, at any level of questioning-frequency, subsequent questions were always asked in regular intervals after a fixed number of object placements. Similarly, question types did not vary and were standardized. These behavioral presets compromise our study's ecological validity because agents assisting users during real world tasks rarely exhibit such systematized behavior, instead interacting in more situationally driven ways that vary with context. It should hence be noted that our findings may be limited to guidance contexts involving scripted rather than fully autonomous or adaptive agent-interactions, an avenue warranting future investigations.

7 Conclusion and Future Work

In this work, we investigated how an agent's information-seeking behavior affects task-guidance in VR. Users sorted objects into an organizer based on guidance from an agent that asked clarifying questions when uncertain. We manipulated the difficulty of the questions and their frequency, studying how task performance, behaviors, and perceptions were affected. Results indicated that the effects of difficulty depended on frequency, with hard questions producing unfavorable outcomes, especially when asked frequently. However, users appear tolerant of frequent clarifications even at the expense of task-performance, provided the demands involved are low.

This work raises several interesting follow up questions. Would similar trends hold for different types of questions? Would sporadic questioning change perceptions? Are questions from embodied agents perceived differently? In future work, we wish to answer such questions and understand how an agent's questioning behavior should adapt based on real-time physiological data of users' workload, and how interactions should evolve over longitudinal time-frames where both agents and users have adapted to each other.

Acknowledgments

This work was partially funded by the U.S. Defense Advanced Research Projects Agency under contract #HR00112220004. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect these agencies' views.

REFERENCES

[1] J. E. Allen, C. I. Guinn, and E. Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999. 1, 2

[2] A. Barquero, R. L. Calvo, D. A. Delgado, I. Wang, L. Anthony, and J. Ruiz. Understanding user needs for task guidance systems through the lens of cooking. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 2006–2018, 2024. 2

[3] G. A. Bekey. On autonomous robots. *The Knowledge Engineering Review*, 13(2):143–146, 1998. 2

[4] R. K. Bellamy, S. Andrist, T. Bickmore, E. F. Churchill, and T. Erickson. Human-agent collaboration: Can an agent be a partner? In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1289–1294, 2017. 2

[5] D. A. Boehm-Davis and R. Remington. Reducing the disruptive effects of interruption: A cognitive framework for analysing the costs and benefits of intervention strategies. *Accident Analysis & Prevention*, 41(5):1124–1129, 2009. 8

[6] J. M. Bradshaw, P. J. Feltovich, H. Jung, S. Kulkarni, W. Taysom, and A. Uszok. Dimensions of adjustable autonomy and mixed-initiative interaction. In *International workshop on computational autonomy*, pp. 17–39. Springer, 2003. 2

[7] C. Brown, J. Hicks, C. H. Rinaudo, and R. Burch. The use of augmented reality and virtual reality in ergonomic applications for education, aviation, and maintenance. *Ergonomics in Design*, 31(4):23–31, 2023. 1

[8] A. C. Brualdi. Classroom questions. *ERIC - Education Resources Information Center*, 1998. 5, 9

[9] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 520–527, 1999. 2

[10] N. Cha, A. Kim, C. Y. Park, S. Kang, M. Park, J.-G. Lee, S. Lee, and U. Lee. Hello there! is now a good time to talk? opportune moments for proactive interactions with smart speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), sep 2020. doi: 10.1145/3411810 1

[11] M. Choi, D. Cui, M. Volonte, A. Koilias, D. Kao, and C. Mousas. Toward understanding the effects of intelligence of a virtual character during an immersive jigsaw puzzle co-solving task. *ACM Transactions on Applied Perception*, 22(2):1–28, 2025. 2

[12] L. Christenbury and P. P. Kelly. Questioning: A path to critical thinking. *ERIC - Education Resources Information Center*, 1983. 5, 9

[13] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran. Interactive team cognition. *Cognitive science*, 37(2):255–285, 2013. 2

[14] S. Coxe, S. G. West, and L. S. Aiken. The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment*, 91(2):121–136, 2009. 7

[15] C. de Armas, R. Tori, and A. V. Netto. Use of virtual reality simulators for training programs in the areas of security and defense: a systematic review. *Multimedia Tools and Applications*, 79(5):3495–3515, 2020. 1

[16] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5503–5512, 2017. 3

[17] B. J. Dixon, M. J. Daly, H. Chan, A. Vescan, I. J. Witterick, and J. C. Irish. Augmented image guidance improves skull base navigation and reduces task workload in trainees: A preclinical trial. *The Laryngoscope*, 121(10):2060–2064, 2011. doi: 10.1002/lary.22153 2

[18] S. Ergün, A. M. Karadeniz, S. Tanrıseven, and I. Y. Simsek. Ar-supported induction cooker ar-si: One step before the food robot. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–5, 2020. doi: 10.1109/ICHMS49158.2020.9209362 2

[19] S. Ergün, A. M. Karadeniz, S. Tanrıseven, and I. Y. Simsek. Ar-supported induction cooker ar-si: One step before the food robot. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–5, 2020. doi: 10.1109/ICHMS49158.2020.9209362 2

[20] D. Escobar-Castillejos, J. Noguez, F. Bello, L. Neri, A. J. Magana, and B. Benes. A review of training and guidance systems in medical surgery. *Applied Sciences*, 10(17), 2020. doi: 10.3390/app10175752 2

[21] S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004. 7

[22] S. Garofalo, S. Giovagnoli, M. Orsoni, F. Starita, and M. Benassi. Interaction effect: Are you doing the right thing? *PLoS One*, 17(7):e0271668, 2022. 7, 8

[23] C. George, M. Spitzer, and H. Hussmann. Training in ivr: investigating the effect of instructor design on social presence and performance of the vr user. In *Proceedings of the 24th ACM symposium on virtual reality software and technology*, pp. 1–5, 2018. 2

[24] M. Giuliani and A. Knoll. Evaluating supportive and instructive robot roles in human-robot interaction. In *Social Robotics: Third International Conference, ICSR 2011, Amsterdam, The Netherlands, November 24-25, 2011. Proceedings 3*, pp. 193–203. Springer, 2011. 2

[25] M. Häring, J. Eichberg, and E. André. Studies on grounding with gaze and pointing gestures in human-robot-interaction. In *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4*, pp. 378–387. Springer, 2012. 2

[26] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 5, 6

[27] M. Hertzum and K. Hornbæk. Frustration: Still a common user experience. *ACM Trans. Comput.-Hum. Interact.*, 30(3), jun 2023. doi: 10.1145/3582432 3

[28] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999. 1, 2

[29] T. N. Höffler and D. Leutner. Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, 17(6):722–738, 2007. doi: 10.1016/j.learninstruc.2007.09.013 1

[30] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69, 2014. 2

[31] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette. Modeling trust in human-robot interaction: A survey. In *International conference on social robotics*, pp. 529–541. Springer, 2020. 2

[32] A. Kim, J.-M. Park, and U. Lee. Interruptibility for in-vehicle multitasking: Influence of voice task demands and adaptive behaviors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(1), mar 2020. doi: 10.1145/3381009 1

[33] A. N. Kluger and A. DeNisi. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254, 1996. 1

[34] B. C. Lee and V. G. Duffy. The effects of task interruption on human performance: A study of the systematic classification of human behavior and interruption frequency. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 25(2):137–152, 2015. 7

[35] H. Liu, M. Choi, L. Yu, A. Koilias, L.-F. Yu, and C. Mousas. Synthesizing shared space virtual reality fire evacuation training drills. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 459–464. IEEE, 2022. 2

[36] K.-Y. Liu, S.-K. Wong, M. Volonte, E. Ebrahimi, and S. V. Babu. Investigating the effects of leading and following behaviors of virtual humans in collaborative fine motor tasks in virtual reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 330–339. IEEE, 2022. 2

[37] R. Q. Mao, L. Lan, J. Kay, R. Lohre, O. R. Ayeni, D. P. Goel, et al. Immersive virtual reality for surgical training: a systematic review. *Journal of Surgical Research*, 268:40–58, 2021. 1

[38] I. Misra, R. Girshick, R. Fergus, M. Hebert, A. Gupta, and L. Van Der Maaten. Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2018. 3

[39] S. Mohammed, L. Ferzandi, and K. Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, 36(4):876–910, 2010. 2

[40] T. O. Nelson, R. J. Leonesio, R. S. Landwehr, and L. Narens. A comparison of three predictors of an individual's memory performance:

the individual's feeling of knowing versus the normative feeling of knowing versus base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(2):279, 1986. 5

[41] T. O. Nelson and L. Narens. Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of verbal learning and verbal behavior*, 19(3):338–368, 1980. 5

[42] D. G. Novick and S. Sutton. What is mixed-initiative interaction. In *Proceedings of the AAAI spring symposium on computational models for mixed initiative interaction*, vol. 2, p. 12, 1997. 2

[43] J. J. Ockerman and A. R. Pritchett. Preliminary investigation of wearable computers for task guidance in aircraft inspection. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*, pp. 33–40. IEEE, 1998. 2

[44] J. Pirker. The potential of virtual reality for aerospace applications. In *2022 IEEE Aerospace Conference (AERO)*, pp. 1–8. IEEE, 2022. 1

[45] A. Pradhan, A. Lazar, and L. Findlater. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(4):1–27, 2020. 2

[46] J. Qin, Z. Chen, W. Zhang, D. Guan, Z. Wu, and M. Zhao. Research on active interaction design for smart speakers agent of home service robot. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*, pp. 253–263. Springer, 2019. 2

[47] E. Ramil Brick, V. Caballero Alonso, C. O'Brien, S. Tong, E. Tavernier, A. Parekh, A. Addlesee, and O. Lemon. Am i allergic to this? assisting sight impaired people in the kitchen. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, p. 92–102. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3462244.3481000 2

[48] C. Rich and C. L. Sidner. Diamondhelp: A generic collaborative task guidance system. *AI Magazine*, 28(2):33, Jun. 2007. doi: 10.1609/aimag.v28i2.2038 2

[49] J. Rickel and W. L. Johnson. Task-oriented collaboration with embodied agents in virtual worlds. *Embodied conversational agents*, pp. 95–122, 2000. 2

[50] P. Rudman and M. Zajicek. Autonomous agent as helper - helpful or annoying? In *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, IAT '06, p. 170–176. IEEE Computer Society, USA, 2006. doi: 10.1109/IAT.2006.41 3

[51] R. Rzayev, G. Karaman, K. Wolf, N. Henze, and V. Schwind. The effect of presence and appearance of guides in virtual reality exhibitions. In *Proceedings of mensch und computer 2019*, pp. 11–20. Association for Computing Machinery, 2019. 2

[52] M. Salem, K. Rohlfing, S. Kopp, and F. Joublin. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 ro-man*, pp. 247–252. IEEE, 2011. 2

[53] S. Schmidt, O. Ariza, and F. Steinicke. Intelligent blended agents: Reality–virtuality interaction with artificially intelligent embodied virtual humans. *Multimodal Technologies and Interaction*, 4(4):85, 2020. 3

[54] V. Schwind, P. Knierim, C. Tasci, P. Franczak, N. Haas, and N. Henze. " these are not my hands!" effect of gender on the perception of avatar hands in virtual reality. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 1577–1582, 2017. 3

[55] R. Semmens, N. Martelaro, P. Kaveti, S. Stent, and W. Ju. Is now a good time? an empirical study of vehicle-driver communication timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300867 1

[56] Z. Shi, Y. Feng, and A. Lipani. Learning to execute actions or ask clarification questions. In *Findings of the association for computational linguistics: NAACL 2022*, pp. 2060–2070, 2022. 1, 2

[57] V. L. Smith and H. H. Clark. On the course of answering questions. *Journal of memory and language*, 32(1):25–38, 1993. 5, 7

[58] R. J. Stone, P. B. Panfilov, and V. E. Shukshunov. Evolution of aerospace simulation: From immersive virtual reality to serious games. In *Proceedings of 5th International Conference on Recent Advances in Space Technologies-RAST2011*, pp. 655–662. IEEE, 2011. 1

[59] W. R. Swartout, J. Gratch, R. W. Hill Jr, E. Hovy, S. Marsella, J. Rickel, and D. Traum. Toward virtual humans. *AI Magazine*, 27(2):96–96, 2006. 2

[60] J. Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312, 1994. 4

[61] B. Tang, H. A. Frye, A. E. Gelfand, and J. A. Silander. Zero-inflated beta distribution regression modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 28(1):117–137, 2023. 7

[62] T. Tofade, J. Elsner, and S. T. Haines. Best practice strategies for effective use of questions as a teaching tool. *American journal of pharmaceutical education*, 77(7):155, 2013. 5, 9

[63] Vociebot.ai.2019. Lifepod is taking reservations for 'proactive' voice assistant and smart speaker for the elderly, 2019. https://voicebot.ai/2019/06/27/lifepod-is-taking-reservations-for-proactive-voice-assistant-and-smart-speaker-for-the-elderly/. 2

[64] J. Vora, S. Nair, A. K. Gramopadhye, A. T. Duchowski, B. J. Melloy, and B. Kanki. Using virtual reality technology for aircraft visual inspection training: presence and comparison studies. *Applied ergonomics*, 33(6):559–570, 2002. 1

[65] M. Walker, T. Phung, T. Chakraborti, T. Williams, and D. Szafir. Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy. *ACM Transactions on Human-Robot Interaction*, 12(4):1–39, 2023. 9

[66] I. Wang, L. Buchweitz, J. Smith, L.-S. Bornholdt, J. Grund, J. Ruiz, and O. Korn. Wow, you are terrible at this! an intercultural study on virtual agents giving mixed feedback. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8, 2020. 5, 6

[67] J. I. Westbrook. Interruptions to clinical work: How frequent is too frequent? *Journal of graduate medical education*, 5(2):337–339, 2013. 8

[68] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011. 5

[69] M. K. Wozniak, R. Stower, P. Jensfelt, and A. Pereira. What you see is (not) what you get: A vr framework for correcting robot errors. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 243–247, 2023. 1

[70] J. Zou, M. Aliannejadi, E. Kanoulas, M. S. Pera, and Y. Liu. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM Transactions on Information Systems*, 41(1):1–25, 2023. 3

[71] J. Zou, E. Kanoulas, and Y. Liu. An empirical study on clarifying question-based systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 2361–2364, 2020. 3

[72] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, G. M. Smith, et al. *Mixed effects models and extensions in ecology with R*, vol. 574. Springer, 2009. 5