doi:10.1093/iwcomp/x

Older Adults' Concurrent and Retrospective Think-Aloud Verbalizations for Identifying User Experience Problems of VR Games

Mingming Fan[1,2], Vinita Tibdewal[3], Qiwen Zhao[3], Lizhou Cao[3], Chao Peng[3], Runxuan Shu[3] and Yujia Shan[3]

[1] *The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China*
[2] *The Hong Kong University of Science and Technology, Hong Kong SAR, China*
[3] *Rochester Institute of Technology, Rochester, NY, USA*

While virtual reality (VR) games are beneficial for older adults to improve their physical functions and cognitive abilities, VR research often does not include older adults. Our review of the proceedings of major HCI conferences (i.e., ASSETS, CHI, CHI PLAY, CSCW, and DIS) between 2016 and 2020 shows that only three out of 352 VR-related papers involved older adults. Consequently, older adults tend to encounter user experience (UX) problems with VR. One common way to identify UX problems is to conduct usability testing with think-aloud (TA) protocols. As VR games tend to be perceptually and physically demanding, older adults might need to allocate more resources to VR content and interaction and thus have fewer resources for thinking aloud. This raises the question of whether TA protocols are still a viable approach to detecting UX problems of VR games for older adult participants. To answer this question, we conducted usability testing with older adults who played two common types of VR games (i.e., the exergame and experience game) using concurrent and retrospective TA protocols (i.e., CTA and RTA), which are widely used in industry. We analyzed participants' TA verbalizations and uncovered how different categories of verbalizations indicate UX problems. We further show how older adults perceived the effects of thinking aloud on their game experiences in two TA protocols and offer design implications.

Categories and subject descriptors: human-centered computing; human computer interaction (HCI); empirical studies in HCI

Keywords: Older adults, think-aloud protocols, virtual reality, VR games, verbalization, UX problems, user experience

RESEARCH HIGHLIGHTS

- We investigated whether concurrent and retrospective think-aloud protocols (CTA and RTA) are suitable for evaluating the user experience of immersive VR games with older adults by studying their verbalizations and subjective experiences.
- We showed that think-aloud verbalization categories and their proportions were similar for both CTA and RTA though CTA had a higher percentage of "Action Description" and RTA had higher percentages of "User Experience" and "Explanation."
- We showed that some verbalization categories were more indicative of UX problems than others. For example, "Problem Formulation" and "User

Experience" were most indicative of UX problems among all verbalization categories. Such trends were similar in both CTA and RTA.
- Older adults felt thinking aloud had minimal effects on their VR game experiences in both CTA and RTA.
- Based on our results, we offer two implications for using the TA method: 1) both CTA and RTA are viable approaches to identifying UX problems of immersive VR games with older adults; 2) As some verbalization categories are more indicative of UX problems than others, UX practitioners could prioritize their attention towards certain categories when identifying UX problems if pressed for time.

1. INTRODUCTION

Virtual reality (VR) has been shown to be able to help older adults improve physical functions, such as reducing the incident rate of falling (Mirelman et al., 2016), improving functional balance (Rendon et al., 2012), and enhancing reaction speed (Wüest et al., 2014). Moreover, VR can also help older adults improve cognitive abilities, such as improving executive function (Gamito et al., 2019) and memory (Optale et al., 2010).

Despite of the potential physical and cognitive benefits of VR for older adults, VR applications (apps) are often designed and evaluated without input from them. To understand to what extent the HCI research community studied VR for older adults, we reviewed recent (2016-2020) five years' proceedings of five major HCI conferences (i.e., ASSETS, CHI, CHI PLAY, CSCW, and DIS), where older adults and VR related papers tend to be published, by searching for the following keywords in the title and abstract: VR, virtual reality, and mixed reality. Among the 352 retrieved papers (CHI: 257, CHI PLAY: 48, ASSETS: 6, CSCW: 1, and DIS: 40), only two CHI papers and one DIS paper focused on older adults. This suggests a lack of VR research focusing on older adults.

Meanwhile, VR has been attracting increasingly more attention from the HCI community. The number of VR-related papers in the proceedings of these five conferences increased exponentially from six in 2016 to 67 in 2020. As VR continues to garner more attention from both industry and academia and becomes increasingly integrated into people's daily lives, it is imperative to ensure that VR is accessible to older adults. One important step to ensure VR's accessibility for older adults is to effectively identify user experience (UX) problems that older adults encounter when they use VR apps so that these problems could then be fixed.

The arguably most widely used method to identify UX problems is to conduct usability testing with think-aloud (TA) protocols (Fan et al., 2020b; Mcdonald et al., 2012). It has been used to identify UX problems that older adults encounter with 2D apps (e.g., websites and games) (Fan et al., 2021; Luger et al., 2014; Huang et al., 2012; Chung et al., 2015; Brewer et al., 2016; Lin et al., 2019). Compared to 2D apps, immersive VR apps rendered in a head-mounted display provide fully immersive environments and are more likely to cause visual fatigue such as eye strain, dizziness, and overall discomfort Cobb et al. (1999); Iskander et al. (2018); Park and Lee (2020). Furthermore, while mouse clicks and touch are typically used to interact with 2D apps, hand gestures and even whole-body motions are employed to interact with immersive VR apps. Thus, immersive VR apps tend to be more physically demanding than 2D apps. Consequently, users may attend more to immersive visuals and engage more in physical interactions and have relatively fewer resources to allocate to think aloud when using immersive VR apps. This raises the question of whether usability testing with TA protocols is still effective to identify UX problems of immersive VR apps. Furthermore, due to age-related declines in perception, motor skills, and cognitive ability, older adults might have fewer resources than their younger counterparts to allocate to think aloud when using immersive VR apps. As a result, it is important to examine whether usability testing with think-aloud protocols is still an effective method to identify UX problems of immersive VR apps for older adult users.

We take an initial exploration to answer this question by assessing the effectiveness of usability testing with two common types of TA protocols in identifying UX problems of immersive VR games for older adults from two perspectives. The first perspective is to analyze older adults' think-aloud verbalizations (i.e., utterances) and uncover *how different types of verbalizations are indicative of UX problems differently*. Such indications could inform UX designers to better focus their attention on verbalizations that are more indicative of UX problems so that they could identify UX problems more effectively. The second perspective is to understand *whether and to what extent older adults might perceive thinking aloud affects their VR game experiences*.

Our exploration investigates these two perspectives on two common types of TA protocols—concurrent think-aloud (CTA), in which users work on a task while at the same time verbalizing their thought processes, and retrospective think-aloud (RTA), in which users first work on the task and then verbalize their thought processes by watching the recording (e.g., video). The specific research questions (RQs) are as follows:

- RQ1: What do older adults verbalize with two TA protocols (i.e., CTA and RTA) while playing VR games?

- RQ2: How might older adults' verbalizations in CTA and RTA indicate UX problems they encounter?
- RQ3: How do older adults perceive the effect of thinking aloud on their VR game experiences in CTA and RTA?

To answer RQs, we recruited older adults who aged 60 and older, based on the World Health Organization (WHO)'s definition of aging population [1], to participate in usability testing. During the usability testing, participants played two VR games while thinking aloud using two TA protocols (i.e., CTA and RTA) respectively. The two VR games were chosen to cover two common types of VR games that older adults often play. One is the VR exercise game (i.e., *the exergame*), which aims to promote physical exercises for older adults in a slow and controlled manner. The other is the VR experience game (i.e., *the experience game*), which aims to provide scenic environments for older adults to explore and experience. After usability testing with two VR games using CTA and RTA, the participants were asked to fill in questionnaires and were interviewed to understand their perceived effects of thinking aloud on their gaming experiences.

In sum, our results show that not all verbalization categories were equally indicative of UX problems; it was similar for both CTA and RTA in terms of how the verbalization categories were indicative of UX problems; both CTA and RTA had little perceived effects on participants' VR game experiences, and were thus both viable approaches to use with older adults for VR games; and VR game types affected participants' preferences for the TA protocols. Overall, participants preferred CTA for the VR experience game and RTA for the VR exergame game. Finally, we present the implications of our findings.

2. BACKGROUND AND RELATED WORK

2.1. VR for Older Adults

Virtual reality (VR) has been shown to have positive effects on older adults' physical functions. For example, when training on a treadmill with VR for six weeks, the incident rate of falling was significantly lower than training without VR (Mirelman et al., 2016). A randomized controlled six-week balance intervention delivered through VR shows that dynamic balance in older adults was significantly improved (Rendon et al., 2012). Wuest et al. found that exercise-based VR games helped to improve older adults' gait- and balance-related physical performance (Wüest et al., 2014). Bisson et al. also found improvement in functional balance and mobility and decrease in reaction time after a 10-week training with a VR game (Bisson et al., 2007).

Moreover, VR has also been shown to have positive effects on older adults' cognitive abilities. For example, Gamito et al. compared VR based cognitive stimulation (VR-CS) with traditional paper-and-pencil cognitive stimulation (PP-CS) with older adults and found that VR-CS had higher improvement in cognition and executive functions than PP-CS (Gamito et al., 2019). Optale et al. showed that older adults had significant improvement in memory tests after six months of VR memory training (Optale et al., 2010). In addition, older adults were found to become more positive toward VR (Roberts et al., 2019; Syed-Abdul et al., 2019) and only experience minimal motion sickness when using VR (Huygelier et al., 2019). This suggests that there is an opportunity for VR to improve older adults' life quality.

However, VR technologies are often designed and evaluated with young populations and older adults are usually not included. As indicated in the Introduction, our literature review of the recent five years' (2016-2020) proceedings of five HCI and accessibility conferences revealed that only three out of 352 VR-related papers focused on older adults as the target population. With fast-paced development in VR technologies, they might become increasingly less accessible to older adults. Thus, it is imperative to understand older adults' experiences in VR to better pinpoint problems they encounter. In this work, we took a step to understand older adults' experiences in VR.

Two common types of VR games have often been used for older adults to promote their health. One type of VR games is **exergames**, which are designed to promote physical exercises by asking older adults to move their limbs and body to interact with game elements (Finkelstein et al., 2011; Bolton et al., 2014; Eisapour et al., 2018). The other type of VR games is **experience games**, which are designed to provide scenarios (e.g., virtual shopping (Laver et al., 2012)). Exposing to natural environments in particular can have a positive impact on health and well-being (Ulrich, 1981; Li, 2010; Park and Mattson, 2009; Ulrich et al., 1991). In this research, we used one exergame and one experience game to study older adults' VR experiences.

2.2. Think-Aloud Protocols and Older Adults

While qualitative methods, such as interviews, surveys, and focus groups, may allow us to understand older adults' attitudes towards VR applications based on *retrospective* self-reports (Roberts et al., 2019), think-aloud (TA) protocols would allow UX evaluators to gain access to older adults' inner thought processes that reflect their real-time interaction flows and experiences with VR

[1] https://www.who.int/health-topics/ageing

applications, which are unavailable in their retrospective self-reports. Although think-aloud protocols have been used to study older adults' thought processes during web information seeking and software use (Huang et al., 2012; Luger et al., 2014; Chung et al., 2015; Lin et al., 2019), little research is known whether TA protocols are viable approaches to study older adults' user experience with VR applications as they might need to allocate more perceptual and cognitive resources toward immersive contents in VR and consequently have relatively fewer resources for verbalizing their thought processes (i.e., thinking aloud).

Concurrent think-aloud (CTA) and retrospective think-aloud (RTA) are two types of TA protocols widely used in industry (Fan et al., 2020b; Mcdonald et al., 2012). It has been a debatable topic as to which of the two protocols is better to use. In terms of participants' task performance, Van et al. found that while both CTA and RTA revealed a similar number and type of usability problems, the task success rate was higher in RTA than CTA (Van Den Haak and De Jong, 2003). In contrast, Alshammar et al. found that there was no difference in task completion rates between CTA and RTA, but CTA was more effective than RTA for finding usability problems (Alshammari et al., 2015). While CTA has been shown to be more popular among UX practitioners (Fan et al., 2020b; Mcdonald et al., 2012), RTA may be less susceptible to the influence of the task difficulty (Van Den Haak et al., 2003). Moreover, Chatrangsan and Petrie investigated older adults use of both CTA and RTA and found they preferred CTA over RTA (Chatrangsan and Petrie, 2017).

In terms of participants' verbalizations, previous research suggested that CTA might lead to a higher percentage of present tense and affective utterances, while RTA might lead to higher percentage of past tense, insight, and cognitive utterances (Olmsted-Hawala and Bergstrom, 2012). In contrast, McDonald et al. found that while verbalizations in CTA often pointed to UX problems, verbalizations in RTA were useful in understanding those UX problems (Mcdonald et al., 2013). In sum, prior studies suggested discrepancies in the similarities and differences of CTA and RTA in terms of success rates, task completion time, and participants' preferences. This has motivated us to examine: how older adults' verbalizations in CTA and RTA would reveal problems with VR games and their preferences.

2.3. Think-aloud Verbalization Categorization

While it has become increasingly easier to conduct think-aloud usability test sessions with more participants, such as via remote usability testing, analyzing test sessions is often arduous (Fan et al., 2020b; Mcdonald et al., 2012). This process often entails analyzing participants' verbalizations and reviewing test session recordings to pinpoint problems. To facilitate analyzing participants' verbalizations, researchers categorized them into common *Verbalization Categories* (Cooke, 2010; Elling et al., 2012; Fan et al., 2019; Zhao and McDonald, 2010; Hertzum et al., 2015). Cooke categorized participants' verbalizations on a website with CTA into five categories (i.e., Procedure, Reading, Observation, Explanation, and Others) and found over half of the verbalizations were Reading (Cooke, 2010). Elling et al. replicated Cooke' study with three different websites and confirmed these five verbalization categories (Elling et al., 2012). These five verbalization categories were also used by a recent study (Fan et al., 2019). In addition to the five-category scheme (Cooke, 2010; Elling et al., 2012), Zhao et al. identified ten verbalization categories (i.e., Reading, Action Description, Action Evaluation, Result Evaluation, Problem Formulation, User Experience, Casual Explanation, Impact, Recommendation and Task Confusion) from CTA sessions (Zhao and McDonald, 2010). These ten verbalization categories were later confirmed by McDonald et al. in their analysis of verbalizations from both CTA and RTA sessions (Mcdonald et al., 2013). Hertzum et al. reviewed verbalization categories from previous work (Bower, 1990; Cooke, 2010; Van Den Haak, 2006; Mcdonald et al., 2013; Zhao and McDonald, 2010; Zhao et al., 2014) and categorized the verbalizations of their participants into six more succinct categories (i.e., Action Description, Explanation, System Observation, Redesign Proposal, Domain Knowledge and User Experience) (Hertzum et al., 2015) . Inspired by previous research, when we categorized participants' verbalizations in this research, we referenced these verbalization categories and adapted their definitions to better interpret our data. In cases where these categories did not apply, our research team discussed and created our own categories.

3. METHOD

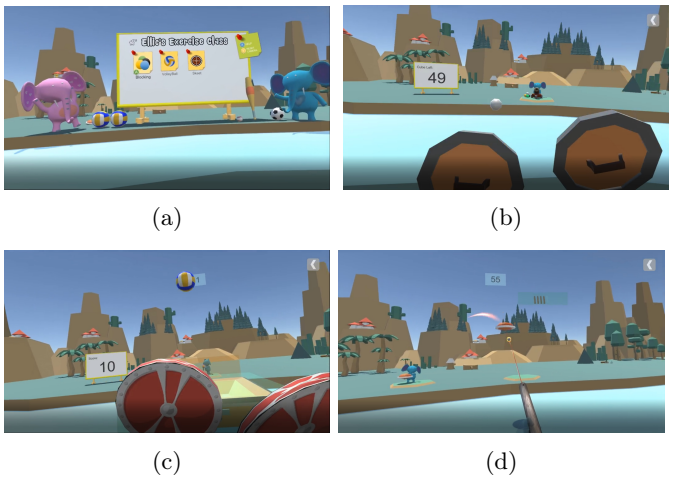
3.1. Participants

We recruited participants from local senior centers via paper flyers and personal visits to the senior centers. We also encouraged participants to contact their friends or colleagues who might also be interested in the study. In the end, eight older adults participated in the study (6 females and 2 males; median age: 77 and age range: 67-85). Table 1 shows their demographic information and prior VR experience. To measure participants' prior VR experience, we used the US National Institutes of Health

(NIH)'s proficiency scale, which describes an individual's level of proficiency in a particular competency [2].

Table 1. Participants' demographic info and the study design (i.e., VR games and TA protocols were counter-balanced).

ID	Age	Prior VR Experience	Design (A: the exergame, B: the experience game)
P1	77	None	A (CTA), B (RTA)
P2	77	None	B (CTA), A (RTA)
P3	77	Novice (limited experience)	A (RTA), B (CTA)
P4	81	None	B (RTA), A (CTA)
P5	73	Fundamental awareness (Heard of it, no experience)	A (CTA), B (RTA)
P6	70	Fundamental awareness (Heard of it, no experience)	B (CTA), A (RTA)
P7	67	Novice (limited experience)	A (RTA), B (CTA)
P8	78	Novice (limited experience)	B (RTA), A (CTA)

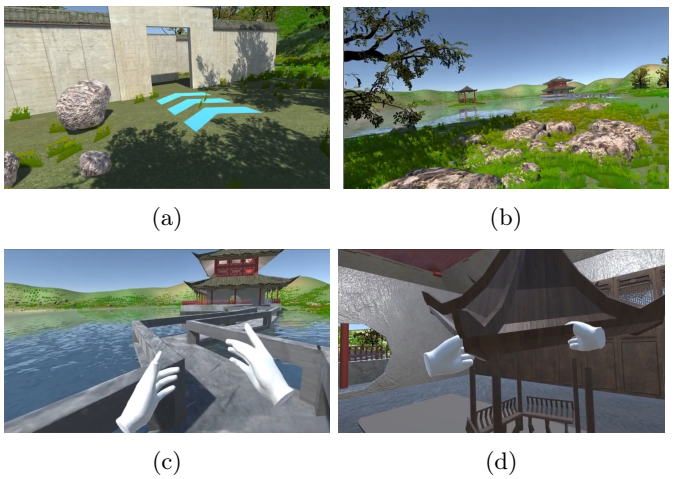


(a) (b) (c) (d)

Figure 1: The screenshots of the exergame: (a) the game selection menu; (b) the ball blocking sub-game; (c) the volleyball sub-game; (d) the skeet shooting sub-game.

3.2. VR Games

Game Selection: Popular VR games, such as Beat Saber [3], require intensive movements and balance skills and are not designed for older adults. Although VR games, such as Zen Zone [4] and Alcove [5], are designed for older adults with computer-rendered scenery, such games mostly provide passive VR experience, such as playing pre-animated scenes or 360-degree videos or have limited opportunity for older adults to interact with the content beyond simple button-pressing. In contrast, we wanted to immerse older adults into virtual environments where they could actively interact with game elements by moving their hands and body with an appropriate level of intensity. Implementing our own games allowed us to customize game elements and interactions for older adults. As a result, we designed and implemented our own games for the study. We conducted five informal pilot study sessions with older adults by asking them to interact with the early versions of the game and provide feedback (e.g., what they liked and disliked and whether the interactions required were difficult for them to complete). We then incorporated the feedback to ensure our VR games were interactive and safe for older adults. Next, we introduce the two types of VR games we designed for the user study: A VR *Exergame* and a VR *Experience Game*.



(a) (b) (c) (d)

Figure 2: The screenshots of the VR experience game: (a) the entrance of the game; (b) a view in the garden; (c) a view on the bridge to the a palace; (d) Lego-like puzzle pieces for building a pavilion.

Exergame: The exergame has three different sub-games to engage players in different physical exercises. Figure 1 shows the sub-game selection scene and the three sub-game scenes. One sub-game is a ball blocking game, the goal of which is to protect a matrix of translucent cubes behind the player from being hit by the balls using virtual shields in their hands. During the game, the player needs to move hands and arms to block the balls with the shields. The second sub-game was a volleyball game where the player needs to bounce a ball into the basket placed in front of them. The goal of the gameplay is to let the player progressively gain a sense of arm and body position adjustments so that they can improve their body

[2] https://hr.nih.gov/about/faq/working-nih/competencies/what-nih-proficiency-scale
[3] https://beatsaber.com/
[4] https://thezen.zone/
[5] https://alcovevr.com/

control. The third sub-game was a skeet shooting game. The player needs to shoot the skeet with a virtual rifle before it falls to the ground. The rifle can load a maximum of five bullets. The player has to reload the rifle after all bullets are fired. The goal of the gameplay is to train the player's responsiveness and hand-eye coordination.

Experience Game: The VR experience game renders natural landscapes such as waterfalls, lakes, bridges, rocks, mountains, and an ancient palace building in a traditional Chinese garden. Figure 2 shows some screenshots of the game, which provides a natural scenery to older adults to enjoy and interact with. The player starts at the gate of the garden (Figure 2 a) and moves the handheld controllers up and down to mimic walking in the garden. The player could move their head to view different garden scenery and use handheld controllers to walk in the garden and interact with items in it. The goal for the player is to reach the palace building as shown in the far end of Figures 2 b and c and use the handheld controller to pick up puzzle pieces from the ground to build a Lego-like pavilion there (Figure 2 d).

3.3. Apparatus

We developed the VR games using Unity 3D game engine and deployed it on a Windows 10 computer. We used Oculus Quest, which includes a head-mounted display (HMD) and two handheld controllers, to render the two VR games and provide immersive and interactive VR experiences. We chose Oculus Quest because it was it was recently released and popular HMD VR equipment in late 2019 when we started the project. In addition, it can be connected to an external monitor, where we could observe and record what participants see for later analysis.

3.4. Study Design

The study was approved by the ethics review committee in our institution. During the usability testing, we asked the participants to play two VR games, one exergame and one experience game, and verbalize their thought processes using CTA and RTA respectively. We counter-balanced the order of the games and TA protocols to alleviate potential order effect. The last column of Table 1 shows the counter-balance design.

3.5. Study Procedure

The study took an average of 75 minutes. Participants first signed the consent form and then went through the following phases: Pre-test questionnaire, CTA practice session, VR preparation, VR practice session, RTA practice session, two VR game sessions, one each with CTA and RTA, and post-test questionnaire and interview.

At the beginning of the study, the moderator asked the participants to inform her during the VR experience if they experienced vertigo or needed any medical assistance. The moderator also closely monitored the situation in case any emergency assistance was needed. The moderator also informed participants that they could stop their participation at any time without giving any reason. In the end, all participants completed the study and no one experienced any vertigo or other medical conditions.

Pre-test Questionnaire: Participants were asked to provide demographic information, such as age, gender, and mention if they have any medical condition such as balance issues or motor impairments that might affect the study results. They were also asked to rate their previous experience with VR, if any. Table 1 shows the information.

CTA Practice: We followed Ericsson and Simon's guidelines (Ericsson and Simon, 1984) to conduct CTA. We first introduced CTA to participants and played a short online video tutorial (Nielsen, 2014) about how to perform CTA. We then asked participants to practice CTA by setting an alarm on a physical alarm clock.

VR Preparation: Participants were asked if they would prefer to sit down or stand while playing the games. We then helped them wear "Sea-Bands" [6] on their wrists as a precaution for motion sickness if they agreed to do so. Then the moderator helped the participants put on Oculus Quest headset.

VR Practice: As participants had no or limited experience with VR, we asked them to play a practice VR game "Oculus First Steps" [7] to get familiar with the VR headset and handheld controllers. The moderator provided instructions and answered their questions when needed. While participants were playing the game, we recorded the game using a screen recorder and also the whole setup in which they played the game.

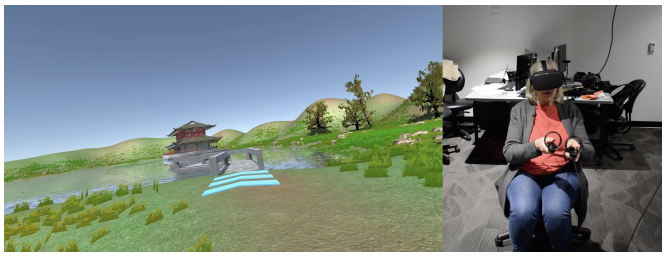
RTA Practice: After participants finished playing the practice VR game, they were shown the video recording of their recorded session and were asked to think aloud.

Two Game Sessions with CTA and RTA: The moderator informed participants what to expect in the games and how to use the handheld controllers. Then, the participants played two VR games using CTA and RTA. Participants were asked to think aloud using CTA and RTA according to the counter-balance table 1.

Figure 3 shows the games and the study setup. Participants played the three sub-games of the exergame in one session and could take a break if needed. We did not ask participants to verbalize during sub-game switching. Thus, sub-game switching did not increase any verbalizations.

[6] https://www.sea-band.com/product/adult-pack/
[7] https://www.oculus.com/experiences/quest/1863547050392688

(a)



(b)

Figure 3: VR games and the study set-up: (a) a participant was standing and playing the exergame; (b) another participant was sitting and playing the experience game.

Posture while Playing the Games: We informed participants that they could choose to play VR games either standing up or sitting in a chair, and all participants chose to stand while playing the two VR games except one who sat during the experience game. Figure 3 shows two participants playing two VR games.

To ensure participants' safety, the participants were asked to tell the moderator if they felt dizzy or sick during the games. Moreover, the moderator closely monitored the participants to ensure participants' safety and also reminded them to keep talking if they fell into silence for more than 15 seconds.

Game Exposure Time: We set the maximum time for each game as 10 mins to avoid fatigue, but allowed participants to wrap up when the time was up. In practice, the average game time of the exergame and experience game was 11 (SD=3) and 8 (SD=2) mins respectively and the difference was not significant. Thus, the exposure time was comparative.

Post-test Questionnaire and Interview: After each VR game session, we asked participants to answer seven 5-point Likert (1: strongly disagree, 5: strongly agree) questions, which were framed in both positive and negative tones, to understand their thinking aloud experiences: *Q1: It was easy to perform and concentrate on the tasks*; *Q2: It was easy to verbalize my thoughts*; *Q3: I felt distracted by the evaluator*; *Q4: I felt verbalizing my thoughts unnatural*; *Q5: I felt verbalizing my thoughts unpleasant*; *Q6: I felt verbalizing my thoughts tiring*; *Q7: I felt verbalizing my thoughts time-consuming*. The lower the rating is, the less negative effect thinking aloud had on their game experiences. We also interviewed participants to understand their ratings of the questionnaires with two questions: "How was your overall experience of TA protocols/VR games?" and "What is your preference for CTA and RTA? Why?"

4. DATA ANALYSES

Participants' TA verbalizations during the usability testing of the two VR games were the primary data to answer our RQs. We transcribed their think-aloud verbalizations (i.e., utterances), categorized them using a widely-used categorization strategy in the literature (Cooke, 2010; Elling et al., 2012; Fan et al., 2019), identified UX problems that participants encountered, quantified the correlations between the verbalizations and UX problems, and compared the effects of TA protocols and VR games on these correlations. Moreover, we further analyzed their questionnaire ratings and interview feedback to better understand their experiences of thinking aloud when playing VR games. Next, we present the details of these analyses.

4.1. Categorizing Verbalizations

Following prior work that analyzed think-aloud verbalizations (e.g., (Cooke, 2010; Elling et al., 2012; Hertzum et al., 2015; Zhao et al., 2014; Fan et al., 2019; Zhao and McDonald, 2010)), we segmented participants' VR game sessions into small segments based on pauses between verbalizations (i.e., utterances) and the semantics of verbalizations (Cooke, 2010). One segment typically contained words, phrases or one or a few sentences that were semantically related and separated by pauses. We manually transcribed participants' verbalizations in each segment to ensure accuracy.

Two researchers independently reviewed verbalizations of each segment and assigned a category label to it by referring to five categorization schemes used in the literature (Hertzum et al., 2015; Zhao and McDonald, 2010; Cooke, 2010; Elling et al., 2012; Fan et al., 2019) while also keeping it open if a new category label was needed. Two researchers discussed any conflicts to try to resolve them. In cases for which the two researchers did not reach an agreement, the third researcher joined the discussion to consolidate the category labels. Table 2 shows the nine categories, definitions, and examples. Seven were adopted and updated from the literature and two (i.e., Supplement and Moderator Intervention) were new categories. The inter-rater reliability for this analysis using Cohen's Kappa was 0.95.

Table 2. Verbalization categories and their descriptions with examples from the study

Verbalization Categories	Descriptions	Examples
Action Description (Zhao and McDonald, 2010; Hertzum et al., 2015)	Participants describing WHAT they did, are currently doing or will/should do.	*"Getting ready to do the paddle and "tak"." - P6 (Exergame)*
Explanation (Hertzum et al., 2015; Cooke, 2010; Elling et al., 2012; Fan et al., 2019)	Participants describing WHY they performed, are performing or will perform a particular task.	*"Aha I am supposed to ... because I just got 2 points. Now I am really beginning to understand." - P8 (Exergame)*
Redesign Proposal (Hertzum et al., 2015)	Participants providing suggestions and recommendations to improve the game or their experience of playing the game.	*"But I would have been excited to stand on the rock. Gets you a little bit of perspective you know?" - P6 (Exergame)*
System Observation (Hertzum et al., 2015)	Participants describing the features and visual layout of the VR game.	*"Well there at first I read the sign and the directions. Then I saw the blue lines pointing. And I wanted to know what they were pointing at." - P5 (Experience Game)*
User Experience (Zhao and McDonald, 2010; Hertzum et al., 2015)	Participants expressing positive or negative feelings and experiences while playing the VR game.	*"Okay..we have a beautiful scene here. Very nice! Looks like it might be a pagoda type temple." - P7 (Experience Game)*
Problem Formulation (Zhao and McDonald, 2010)	Participant verbalization that indicates difficulty or uncertainty while playing the games which is expressed as a negative feeling, generally caused by a system based issue(s).	*"The balls are coming at me and I'm trying to figure out what to do with them and I'm..I think I don't know if I am supposed to be hitting them into this thing in front of me or not." - P8 (Exergame)*
Others (Cooke, 2010; Elling et al., 2012)	Participant verbalization that are not relevant to the game.	*"But I suppose, nowhere to take your shoes off, so that would've been disrespectful." - P5 (Experience Game)*
Moderator Intervention	While playing the game, if the participants were quiet for more than 15 seconds, the moderator was reminding them to keep talking.	*"So what are you thinking?" - P2 (Experience Game)*
Supplement	Participants mentioning something that they realized while watching the recorded video during the RTA.	*"What's the shadow? I didn't see the shadow before." - P1 (Experience Game)*

4.2. Identifying UX Problems

For each segment, two researchers independently assessed whether the participant encountered a problem by referencing classic usability heuristics and design principles (Nielsen, 1994a; Norman, 2013). They then discussed any conflicts in their problem categorization. In cases for which the two researchers did not reach an agreement, the third researcher joined the discussion to try to resolve the conflicts. The example UX problems and corresponding think-aloud verbalizations are shown in Table 3.

4.3. Computing How Verbalization Categories Indicate UX Problems

4.3.1. Verbalization Category Proportions

We first calculated the percentage of verbalization segments in each category for CTA and RTA respectively to understand how verbalizations were distributed among different categories and how TA protocols might affect the distribution. We then calculated the percentage of verbalization segments in each category for two types of games respectively to understand how games might affect the distribution.

4.3.2. Precision, Recall, and F-measure of Verbalization Categories for identifying UX Problems

We calculated precision, recall, and F-measure of each verbalization category to understand how well each category is indicative of UX problems. These measures are commonly used to calculate the performance of a machine learning classifier and have recently been used to understand the relationship between verbalization categories and UX problems (Fan et al., 2019, 2020a, 2021). If i denotes a verbalization category, which can be any one of the nine categories in Table 2, then Precision and Recall of the category i are as follows:

Precision(i) = Number of segments with a problems in category i/Total number of segments in category i;

Table 3. The UX problems with usability heuristics violated and the corresponding problem descriptions and examples.

UX Problems (*Usability heuristics violated* Nielsen (1994a))	Example problems and think-aloud verbalizations
Users could not receive appropriate feedback to know the system status in time (*Visibility of system status*)	**Exergame**: Participants were confused about how the scoring system works in the Volleyball game. For example, P3 verbalized, *"Uhm..I couldn't figure out how the score was going and but you know if I started to look at the score, I found that I couldn't keep track of what I was supposed to be doing by hitting the ball over. So that was the problem."*;
The design did not speak users' language or failed to match users' mental model (*Match between system and the real world*)	**Experience game**: The garden design in the VR experience game did not match the participants' mental model of a Chinese garden. P5 said, *"So I was thinking, this is the strangest Chinese garden that I've ever seen. I didn't see a lot of urns. I didn't see any goldfish."*
Users could not back out of a process or undo an action easily (*User control and freedom*)	**Exergame**: Participants accidentally selected the same game that they played earlier. However, they did not know how to go back to the navigation menu and were forced to play it again. P1 verbalized, *"Ahh.. Skeet. No we did Skeet. I don't want Skeet. I want to do HELP. Shoot the target. No I don't want to shoot the target. I have no idea what to do here."*; **Experience game**: Participants were doing the walking action but they were unable to control the walking speed of their avatar in the game which made them dizzy. P2 verbalized, *"Are we starting? Am I not moving? Oh there we go..ohh ohh..now it's going really fast. It's kind of dizzying"*
The design failed to prevent problems from happening in the first place (*Error prevention*)	**Experience game**: Participants were moving their hands in the wrong direction and that's why they were unable to move. P6 was annoyed and verbalized, *"How come I'm not moving? I'm frustrated. I'm not going anywhere."*
The design required users to memorize certain information (*Recognition rather than recall*)	**Experience game**: Participants were provided with all the instructions to navigate in the garden at the beginning of the game, and they were expected to remember them. In the middle of the game, P3 got confused about the functionality of the buttons on the remote control and verbalized, *"That's the one under my index finger, right? The grip trigger. Which one is it?"*
The design did not provide users with different ways to accomplish a task (*Flexibility and efficiency of use*)	**Experience game**: The participants can use walking functionality instead of teleportation to move in the garden. However, this was not suggested to them when teleportation was not working for them. P3 verbalized, *"I think I'll go over there. I think it won't let me. Over to the bridge. Nope. Not gonna let me go there. Ohh looks like some butterflies over there. Maybe I'll go up this hill. Nope. It won't let me. Alright."*
The visual design failed to support users' primary goal effectively (*Aesthetic and minimalist design*)	**Exergame**: There was a visual indicator in the Shooting exergame that showed the number of bullets loaded in the gun. However, participants either did not notice it or they did not understand what it meant. They were left wondering what happened when the gun stopped working as it went out of bullets. P6 saw the reload bar and said, *"So I'm seeing this little orange bar flashing there lightly. I'm not sure what that means."*
The design failed to provide users with effective error messages that could indicate problems and suggest solutions (*Help users recognize, diagnose, and recover from errors*)	**Exergame**: The participant wanted to select a new game and pressed the wrong button and did not understand why they are not able to select a new game. P5 verbalized, *"Now I gotta choose again? I have no idea. Oh. I don't see anything happening after skeet."*
Users could not get additional support to complete tasks from system documentation (*Help and documentation*)	**Exergame**: The instructions in the game did not help the participants in understanding how to play the Blocking game. P4 was confused with the instructions displayed in the game and verbalized: *"Cubes left. Were they supposed to hit the thing? Woah. I don't understand the instructions as you can see."*

Recall(i) = Number of segments with a problems in category i/Total number of segments with a problem;

Intuitively, if a category i has a higher Precision, UX evaluators have a higher chance to find a problem by examining a segment of category i than a segment of other categories. In contrast, if a category i has a higher Recall, UX evaluators have a higher chance to catch more problems by examining the segments of category i than the segments of other categories.

In practice, if a category has a higher Precision, it usually has a relatively lower Recall. Thus, it is also necessary to have one single measure to combine the effects of Precision and Recall. One common such measure is F-measure, which is calculated using the following equation.

$F-measure(i) = \frac{2*Precision(i)*Recall(i)}{Precision(i)+Recall(i)}$

Intuitively, if a category i has a higher F-measure, UX evaluators have an overall higher chance of finding a problem to examine the segments of category i than the segments of other categories.

All our analyses were based on the percentages of different verbalization categories instead of the absolute numbers. Thus, our findings reflect the relative importance of each category in revealing UX issues.

4.4. Analyzing Questionnaire and Interview Data

We collected questionnaire ratings and brief interview feedback to understand participants' experiences of thinking aloud while playing VR games. Specifically, we performed Shapiro-Wilk normality tests on participants' ratings of seven questions related to their experiences with CTA and RTA (See the last paragraph of Sec 3.5). For the ones that met the normal distribution requirement, We performed the paired t-test if the normal distribution requirement was met or Wilcoxon Signed Rank test if the requirement was not met for the ratings of each question to see if there is any significant difference.

Furthermore, two researchers independently coded interviews and discussed the codes afterward to resolve any potential conflicts. In the end, we identified four main themes: experiences and preferences of CTA and RTA, VR game experiences, and difficulties encountered.

5. RESULTS

5.1. Verbalization Category Proportions (RQ1)

5.1.1. Verbalization Category Proportions by TA Protocols

Table 4 shows the percentage and number of segments in each verbalization category in CTA and RTA. We observed four similarities. First, "Problem Formulation" appeared most frequently for both CTA and RTA. Second, "Action Description", "User Experience", and "System Observation" were among the next group of most frequently appeared categories with a similar percentage range (between 14.3% and 24.5%). Third, "Explanation", "Moderator Intervention", and "Others" were among the next group of categories with a similar percentage range (between 10.8% and 3.3%). Lastly, "Redesign proposal" was the least frequently appeared common category between CTA and RTA.

Table 4. The mean (standard deviation) of percentages of segments in each verbalization category for each TA protocol.

Category	CTA	RTA
Problem Formulation	28.7% (16.7%)	23.5% (8.9%)
Action Description	24.5% (11.2%)	19.1% (10.9%)
User Experience	18.9% (9.4%)	19.4% (11.8%)
System Observation	14.3% (8.9%)	15% (5.5%)
Moderator Intervention	10.8% (6.0%)	7.0% (5.0%)
Explanation	4.9% (2.1%)	8.6% (4.6%)
Others	4.4% (3.2%)	4.8% (3.2%)
Redesign Proposal	3.3% (0.4%)	5.0% (2.9%)
Supplement	0	4.9% (2.6%)

There were also differences between CTA and RTA. First, Supplement, by definition, was unique to RTA. Second, CTA had a higher percentage of "Action Description" (24.5%) than RTA (19.1%). Second, RTA had a higher percentage of Explanation (8.6%) than CTA (4.9%). Lastly, CTA had higher percentage of "Moderator Intervention" (10.8%) than RTA (7.0%).

5.1.2. Verbalization Category Proportions by Games

Table 5. The mean (standard deviation) of percentages of segments in each verbalization category for each VR game.

Category	Exergame	Experience Game
Problem Formulation	28.4% (15.4%)	25.0% (13.2%)
User Experience	20.8% (10.8%)	17.8% (10.0%)
Action Description	16.3% (6.3%)	26.5% (12.1%)
System Observation	12.5% (7.9%)	16.2% (7.1%)
Moderator Intervention	9.9% (6.9%)	4.2% (5.2%)
Explanation	4.6% (2.9%)	5.4% (6.0%)
Others	3.8% (3.1%)	1.7% (3.2%)
Supplement	6.0% (3.9%)	1.6% (1.9%)
Redesign Proposal	1.7% (1.9%)	1.4% (3.0%)

Table 5 shows the percentage and number of segments in each verbalization category in the exergame and the experience game. It shows three patterns: 1) the proportions of verbalization categories were overall similar for both the Exergame and the Experience game; 2) the "Problem Formulation" category was the most popular category for both the Exergame (28.4%) and the Experience game (25%), followed by "User Experience" and "Action Description"; 3) while the percentages of most categories were similar for the two types of games, there was one difference that the Experience game had a higher percentage of "Action Description" than the Exergame. One potential explanation could be that when playing the Exergame, participants tended to focus more of their physical and cognitive resources on blocking a ball, bouncing a volleyball, or shooting a skeet. Thus, they had relatively less cognitive resources to verbalize their actions than when they played the Experience game, which was less cognitively demanding.

5.2. How Each Verbalization Category Indicates UX Problems (RQ2)

5.2.1. UX Problems

The exact UX problems were not the focus of this research. Instead, we were interested in understanding how participants' verbalizations indicate UX problems (RQ2). The implication was that verbalization patterns indicating UX problems could be extracted with computational approaches and used for inferring UX problems automatically. Nonetheless, we describe example UX problems, the UX heuristics that were violated (Nielsen, 1994a), and examples in the Table 3.

In this section, we present example think-aloud verbalizations of each verbalization category to illustrate how they indicated UX problems in Table 6. In the next subsection, we quantify the correlations between verbalization categories and problems.

As segments of "Moderator Intervention" and "Others" categories did not relate to UX problems, we did not include them in the following analysis.

5.2.2. Quantifying How Verbalization Categories Indicate UX Problems.

We first calculated Precision, Recall and F-measure of each verbalization category for identifying UX Problems. Table 7 shows the results. First, "Problem Formulation" had the highest scores for all three measures, which suggested that it was the most indicative of UX problems among all categories. Second, all categories had relatively low recall except "Problem Formulation." This suggested that problems were widely distributed among different categories and thus UX evaluators would only be able to locate a small percentage of the problems if they only checked one of these categories. Third, "Explanation", "Redesign Proposal", and "Supplement" had Precision of around or above 0.50. This suggested that UX evaluators would be able to identify a problem with around or above 50% chance if they reviewed a segment from any of these categories.

5.3. Perceived Effects of Thinking Aloud on VR Game Experiences (RQ3)

To understand whether usability testing with thinking aloud is a viable approach to detect UX problems of VR games for older adults, we asked participants how much they felt thinking aloud had affected their VR game experiences with 5-point Likert scale questions (see the last paragraph in Section 3.5). The first two columns of Table 8 shows participants' perceptions of thinking aloud grouped by TA protocols. Results suggest that overall participants felt it was easy to verbalize thoughts with both CTA and RTA. Moreover, they did not feel thinking aloud was unnatural, unpleasant or distracting. Wilcoxon-Signed Ranks test results show that there was no significant difference in their ratings for CTA and RTA.

To understand whether VR games affect participants' perceptions of thinking aloud, we computed the mean and standard deviation of their ratings of the questions grouped by the games. The last two columns of Table 8 shows the results. Results of paired t-tests or Wilcoxon-Signed Rank tests found no significant difference for all ratings except the rating of the naturalness of verbalization between two types of games (i.e., *I felt verbalizing thoughts unnatural*) ($t(7) = 2.65, p = .03$). Specifically, participants felt that verbalizing their thoughts was relatively more unnatural during the exergame ($M = 2.38, SD = .86$) than during the experience game ($M = 1.38, SD = .48$). Participants' feedback during interviews echoed the effect of the games on their experience of thinking aloud. For example, some participants preferred RTA when playing the exergame because they did not have to verbalize their thoughts while playing the exergame and thereby could better focus on the game while playing: "*I had to remember on the active think-aloud to actually do my think-aloud while playing the game. While playing the game, I'm trying to concentrate and figure out what it is and I keep forgetting the talk but once I remembered, talking was a piece of cake. - P8.* On the other hand, some participants preferred CTA when playing the experience game because they did not have to recall their thoughts later like in RTA.

We noticed that participants tended to pause more often and thereby need to be reminded more often to think-aloud when playing the exergame than the experience game in CTA than in RTA. Indeed, the moderator intervened more often in CTA sessions ($M =$

Table 6. The example think-aloud verbalizations of each verbalization category that indicate problems.

Verbalization category	Example think-aloud verbalizations indicating problems
Problem Formulation	*"I never figured out how to get in closer to the table because there's shadows of the hands, right? Every time I tried to move closer to the table, it wasn't working." - P1 (Experience Game)*; *"Now reload. Shoot. Joystick? I don't know. Is it...They didn't give me any instructions on how to reload! Okay. So where am I supposed to be shooting?" - P5 (Exergame)*
Action Description	*"But I was trying to get into the doorway. But I wouldn't go in." - P4 (Experience Game)* *"And then here I was able to quickly go for the skeet and again I needed a rescue with the gun." - P6 (Exergame)*
Explanation	*"Here we go again. I still couldn't figure out how to move, because I tried the same way you walk." - P1 (Experience Game)* *"I think I mentioned how many, how many left and I said cubes left 46. I thought they meant balls but then I realized no it said cubes." - P6 (Exergame)*
Redesign Proposal	*"You know you should have lots of pink flowers and all kinds of pretty. things you know, gold fish ponds and I could find and get interested. Like I said I kept trying and I either get blocked by the water or boulders. So that's what I was thinking." - P5 (Experience Game)* *"There is only one problem with this game. I don't get to shoot any back at him [a character in the game]! So that's what's missing. The ability to see if you knock him. I wanna get a chance at him." - P1 (Exergame)*
System Observation	*"So there's a rock garden there. I'm looking at the rock garden. I'm turning and I can see grass and another rock garden" - P7 (Experience Game)* *"And then what was that gun over there on the side? I guess my opponent." - P3 (Exergame)*
User Experience	*"Alright. So I was thinking, this is the strangest Chinese garden that I've ever seen. I didn't see a lot of urns. I didn't see any goldfish." - P5 (Experience Game)* *"I'm thinking this is difficult and I don't like guns. Not even doing this." - P4 (Exergame)*
Supplement	*"I bet you that would've been easier going around in here instead of that but I forgot about doing the walking" - P8 (Experience Game)* *"So I'm seeing this little orange bar flashing there lightly. I'm not sure what that means. I didn't see it while I was playing the game." - P6 (Exergame)*

Table 7. The Precision, Recall and F-measure of each verbalization category for locating UX problems.

Category	Precision	Recall	F-measure
Problem Formulation	1.00	0.50	0.67
User Experience	0.49	0.18	0.26
Action Description	0.32	0.13	0.18
Explanation	0.70	0.07	0.13
System Observation	0.26	0.07	0.11
Redesign Proposal	0.67	0.02	0.04
Supplement	0.55	0.02	0.04

$2.5, SD = 2.27$) than in RTA sessions ($M = 1.83, SD = 1.94$). This was likely because participants tended to be more occupied cognitively and physically by the exergame content than the experience game and therefore tended to forget thinking aloud more often. This was evident when the moderator prompted them to think aloud, most of them mentioned that they were not thinking anything but focusing on figuring out the game.

6. DISCUSSION

We first present the key findings of the study and then discuss how these findings answer our RQs.

Table 8. The ratings of the questions regarding participants' experiences of two thinking aloud protocols.

Questions	TA Protocols		Games	
	CTA	RTA	Exergame	Experience
It was easy to perform and concentrate on the tasks	4.4(0.5)	3.6(1.3)	3.8(1.2)	4.3(0.9)
I felt verbalizing thoughts was easy	4.5(0.5)	4.1(1.0)	4.1(1.0)	4.5(0.5)
I felt distracted by the evaluator	1.3(0.5)	1.4(0.5)	1.3(0.5)	1.4(0.5)
I felt verbalizing thoughts was difficult	1.5(0.8)	1.9(0.6)	1.9(0.9)	1.5(0.5)
I felt verbalizing thoughts was unnatural	1.9(1.1)	1.9(0.6)	2.4(0.9)	1.4(0.5)
I felt verbalizing thoughts was unpleasant	1.6(0.7)	1.5(0.5)	1.8(0.7)	1.4(0.5)
I felt verbalizing thoughts was tiring	1.9(0.8)	1.5(0.5)	1.8(0.7)	1.6(0.7)
I felt verbalizing thoughts was time-consuming	1.8(0.9)	1.5(0.5)	1.6(0.7)	1.6(0.7)

6.1. Key Findings

We investigated whether two common types of think-aloud protocols (i.e., CTA and RTA) are suitable methods for identifying the UX problems of immersive VR games with older adults by analyzing their verbalizations and subjective experiences of the protocols. Our findings show that: 1) participants' verbalization categories and the corresponding proportions are similar for both CTA and RTA; 2) Some verbalization categories are more indicative of UX problems than other in both CTA and RTA. For example, "problem formulation" and "user experience" are more indicative of UX problems among all verbalization categories; 3) Participants felt that thinking aloud had minimal effects on their immersive VR game experiences in both CTA and RTA.

6.2. Verbalization Category Proportions (RQ1)

6.2.1. CTA vs. RTA

As Table 4 shows, both CTA and RTA had a similar set of categories and similar proportions for most categories. However, CTA had a higher percentage of "Action Description" than RTA. One potential reason could be that it was easier to describe *what they were doing* (i.e., "Action Description") when participants were still doing it in CTA than to recall that action later in RTA. In contrast, RTA had a higher percentage of "User Experience" and "Explanation" than CTA. This might be because in RTA participants tended to verbalize explanations (i.e., "Explanation") or experiences (i.e., "User Experience") of what they were doing instead of directly describing *what they were doing* (i.e., "Action Description"). For example, in the exergame session, P5 verbalized in CTA, *"I'm trying..I'm trying to get this ball across the net."* ("Action Description"); in contrast, P6 verbalized about similar experience in RTA, *"Getting it into the box. It did matter how hard you pushed and how the direction you pushed the ball, so there was some degree of accuracy."* ("Explanation"). In the experience game session, P6 verbalized in CTA, *"I am...trying to move [but did move anywhere]...Oooh [lost balance while walking]...Okay."* ("Action Description"); in contrast, regarding similar experience, P4 verbalized in RTA, *"It was very dizzying. I guess I was feeling, you know, the vertigo."* ("User Experience"). Another difference was that CTA had a higher percentage of "Moderator Intervention", which suggested that participants forgot to think aloud more often in CTA than in RTA. This might be because participants had to prioritize their cognitive resources toward playing VR games in CTA and thus tended to forget to verbalize their thoughts. In contrast, participants watched recorded sessions while thinking aloud in RTA. Thus, they did not have to allocate cognitive resources toward playing the games anymore and thus tended to be able to verbalize their thoughts more often.

6.2.2. the Exergame vs. the Experience Game

As Table 5 shows, the proportion of eight out of the nine categories were similar for two types of games. This suggested that participants verbalized similar proportions of content in eight out of the nine categories regardless of the types of games. However, one difference was that participants verbalized more content of "Action Description" in the experience game than the exergame. While the experience game was relatively relaxing and participants mostly wandered around in the VR space, the exergame was relatively intense and participants had to observe the game elements (e.g., moving balls or skeet) and move their arms to block or shoot these elements. As a result, participants tended to have less cognitive resource to verbalize their actions ("Action Description") in the exergame than in the experience game. The implication is that while UX evaluators could expect similar proportions of most categories of verbalizations when older adults are playing a VR game, they should also expect fewer "Action Description" verbalizations if the VR game consumes a significant amount of cognitive resource as exergames.

6.3. How Verbalization Categories Indicate UX Problems (RQ2)

"Problem Formulation" was, by definition, always related to problems. Thus, it had a Precision of 1. Its overall Recall was 0.5, which meant that 50% of the problems were in "Problem Formulation" segments. Although the segments of the rest categories shared the other 50% of the problems, UX evaluators would be able to catch 68% of the problems by just checking segments of one additional category "User Experience". This was promising given that these two categories were the most indicative of UX problems for both CTA and RTA. Moreover, this number would raise to 81% if adding another category "Action Description". This suggested that it was possible for UX evaluators to prioritize their attentions toward certain categories of segments if they were pressed for time.

While the trends between categories and UX problems were similar for CTA and RTA, one difference was that "Explanation" had higher precision and F-measure in RTA than CTA. This suggested that "Explanation" was more indicative of problems in RTA than CTA. This was partially because RTA had a higher percentage of "Explanation" than CTA. One implication is that when pressed for time, UX evaluators would be better off check segments of "Problem formulation", "User Experience", and "Explanation" in RTA.

Furthermore, "Redesign Proposal" was rare in both CTA and RTA. Thus, UX evaluators should not expect to hear design recommendations from participants often. While the literature suggested UX evaluators not actively ask participants for design recommendations as doing so would alter their task behavior (Fan et al., 2020b; Mcdonald and Petrie, 2013), our results showed that UX evaluators should pay attention to such redesign proposals if participants voluntarily proposed. This is because the Precision of "Redesign Proposal" was relatively high for both CTA and RTA.

6.4. Perceived Effects of Thinking Aloud on Participants' VR game experiences (RQ3)

Results in Section 5.3 showed that thinking aloud had minimal effects on older adults' VR game experiences. Furthermore, TA protocols did not have any significant effect on this perception. This suggested that both CTA and RTA are viable usability testing methods to uncover UX problems of VR games for older adults without significantly affecting their game experiences.

Our results also showed that the game type affected their perceptions of CTA and RTA. Participants felt that verbalizing thoughts while playing the exergame at the same time (CTA) was relatively more unnatural than verbalizing their thoughts when watching the recording of their play session (RTA). The potential reason was that the exergame required them to devote more cognitive and motor resources toward the game and thus had relatively fewer resources allocated to verbalize thoughts at the same time. The implication is that if the testing VR game is cognitively or physically demanding for older adults, they would likely verbalize fewer thoughts of Action Description category in CTA than RTA. If the testing goal includes understanding older adults' interaction issues through their actions, then RTA would be a better choice than CTA though both CTA and RTA would likely produce similar proportions of other categories of verbalizations.

6.5. Implications for Choosing CTA or RTA for Evaluating VR Apps for Older Adults

Prior work suggested that older adults might prefer CTA over RTA as CTA could be more natural to them as they often "talk themselves through" when using a new technology (Chatrangsan and Petrie, 2017). Our research extends on this line of work and provides two design implications. First, our results show that both CTA and RTA have little negative effect on older adults' VR game experiences. This suggests that both CTA and RTA are viable usability testing methods to use for understanding older adults' experiences with VR games. Moreover, our study also highlights some trade-offs between CTA and RTA. On the one hand, if a VR game requires users to interact with game elements frequently or intensively, RTA might be a better choice than CTA because RTA frees users from the need to allocate some cognitive resources to thinking aloud while interacting with game elements frequently or intensively. On the other hand, CTA might be a better choice if the VR game stimulates rich visceral experiences, which might be hard for older adults to recall after completing the game as in RTA.

Second, our results provide insights into how older adults' verbalizations in CTA and RTA indicate the UX problems that they experience in VR games. The design implication is that computational methods could be developed to process participants' verbalizations (i.e., utterances) and draw the UX evaluator's attention towards the verbalization categories that are more indicative of UX problems. In doing so, the UX evaluator might be able to identify UX problems more efficiently or not miss any UX problems. Furthermore, such verbalization patterns indicating UX problems could also be extracted and used to train machine learning (ML) models to automatically detect UX problems (e.g., (Fan et al., 2020a; Harms, 2019)). It remains open questions how to achieve such goals and what roles ML and UX

evaluators should play to better identify and resolve the UX problems of VR games.

7. LIMITATIONS AND FUTURE WORK

We took a first step to understand older adults' experiences with two common types of VR games through their verbalizations in two types of think-aloud protocols (CTA and RTA). While shedding light on the effects of two TA protocols on older adults' VR game experiences, our work has some limitations and also opens up new opportunities for future work.

Small Samples. The outbreak of COVID-19 prevented us from conducting more in-person studies. Consequently, our findings were based on a small number of participants' data. Given the limited samples, we did not perform statistical analysis on the numbers related to the verbalizations. Thus, it remains unknown whether the differences in verbalization categories, think-aloud protocols, VR games are statistically significant. More studies with a larger sample size are warranted to answer this question. Nevertheless, as an initial exploration, our work has uncovered the categories of verbalizations, how each category indicates UX problems, and older adults' perceived effects of thinking aloud while playing VR games. Ultimately, these results suggest that usability testing with these two types of think-aloud protocols is a viable approach to identifying UX problems of VR games for older adults.

Sensor Data. In this research, we focused on understanding whether think-aloud is effective to identify UX problems of immersive VR games for older adults through their verbalizations and their perceived effects of thinking aloud on their game experiences. Toward this goal, we did not analyze heart rate data in this work. However, as prior work suggested that there is a different physiological response in the body when when playing against a computer versus playing against a human player (Mandryk et al., 2006), one interesting future work is to explore whether heart rate data and perhaps other types of sensor data (e.g., facial expression, body gestures) could be leveraged to infer UX problems of VR games.

Interactions between the Moderator and Participants. We followed Ericcson and Simon's guidelines (Ericsson and Simon, 1984) to minimize the interactions between the moderator and participants. The moderator only reminded participants to "keep talking" when they fell into silence for a while. With this minimal interaction, our study results suggest that participants did not feel much negative effect of thinking aloud on their VR game experiences. However, it is not uncommon that moderators would actively probe participants by asking questions, which is known as relaxed think-aloud protocols. It remains unknown whether "nagging" from the moderator in such related protocols would have effects (e.g., intrusiveness, ethics) on older adults' VR game experiences.

In addition, we had one moderator to conduct all the studies to keep the interactions between the moderator and participants consistent. In practice, it is possible that multiple moderators conduct usability testing with think-aloud protocols. The moderator's prior experience and way of conducting the session might affect participants' think-aloud experiences. Thus, it is also important to further understand how moderators' ways of conducting test sessions (e.g., consistency) might affect participants' think-aloud and VR game experiences.

Prior VR Experiences. All of our participates had no or limited prior experiences with VR. Older adults with rich VR experience might have more cognitive resources they can allocate toward thinking aloud and thus may have different preferences for the two TA protocols. More research is warrant to understand how older adults' VR experiences may affect the findings.

Types of Immersive VR Applications. We used two common types of VR games (i.e., exergames and experience games) for this initial exploration to understand whether CTA and RTA are feasible methods to identify UX problems among older adult users. These two VR games were single-user VR games. Researchers have recently begun to study social VR for older adults (Baker et al., 2019). The multi-user scenarios require collaboration and communication among older adults, which is different from single-user game scenarios. It is interesting to explore how multi-user VR games might affect older adults' experiences with CTA and RTA protocols and how their verbalizations might suggest UX problems differently. Second, we used immersive VR games for this research. The UX requirements of games may not be the same as general applications. For example, having appropriate levels of challenges might be beneficial to keep gamers engaged; however, challenges might not be appreciated by users when everyday applications. It is imperative to investigate whether and how older adults' experiences with CTA and RTA might change when they use non-game VR applications.

Severity of UX Problems. We annotated whether a segment had a problem or not, but did not annotate the severity of UX problems. The severity of a UX problem was determined by many factors, such as frequency with which the problem occurs and their potential market impact (Nielsen, 1994b). Future work should explore whether older adults' verbalizations (i.e., utterances) also suggest the severity of UX problems.

Human-AI Collaboration for Uncovering UX Problems. Analyzing usability test sessions is time-consuming as it often entails reviewing session recordings and scrutinizing users' actions and verbalizations to

pinpoint UX problems (Fan et al., 2020b; Mcdonald et al., 2012). However, in practice UX evaluators often have short time budget to complete their analysis. Thus, there is a need for fast-paced UX analysis methods. One implication of our findings is that UX evaluators could catch 68% of the problems by only examining two categories of segments (i.e., the Recall value of the sum of the first two rows in Table 7) or 81% if adding an additional category. This suggests that not all segments of a usability test session are equally indicative of UX problems. Thus, it is possible to capture a majority of the problems by only focusing on segments that are more indicative of UX problems, which might be preferable when the UX team is pressed for time. To make this happen, future work could leverage artificial intelligence (AI) to automatically categorize the verbalizations into categories and design human-AI collaboration tools (e.g.,Fan et al. (2020c, 2022); Soure et al. (2021)) to help UX evaluators better allocate their attention toward the segments that are more indicative of UX problems.

8. CONCLUSION

We have studied older adults' think-aloud (TA) verbalizations in two TA protocols (CTA and RTA) when they played two common types of VR games and uncovered how different verbalization categories indicated UX problems that they experienced. Specifically, we have identified nine TA verbalization categories and found that the proportions of the verbalization categories were overall similar for both CTA and RTA. While CTA had higher proportions of "Action Description" and "Moderator Intervention" categories, RTA had higher proportions of "User Experience" and "Explanation."

Moreover, the proportions of verbalization categories were roughly the same for the two types of VR games (i.e., the exergame and the experience game). Furthermore, we have proposed three measures (i.e., precision, recall, F-measure) to quantify how verbalization categories indicate UX problems. Our findings show that the how verbalization categories indicate UX problems were overall similar for both CTA and RTA. These results suggest that older adults' TA verbalizations in both CTA and RTA are effective in uncovering the problems they encounter when playing VR games.

Additionally, we have studied older adults' subject experiences of thinking aloud with CTA and RTA when playing VR games and found that older adults felt thinking aloud with CTA and RTA had little effect on their VR game experiences. This suggests that older adults are receptive of both CTA and RTA as part of usability testing for uncovering UX issues of VR games.

That said, there are still some differences between CTA and RTA. Older adult participants felt that thinking aloud with RTA was relatively easier than CTA if they had to interact frequently with the game elements such as the VR exergame in our study. They also felt that thinking aloud with CTA was relatively more straightforward than RTA when interacting with VR games that invited visceral experiences, such as the VR experience game in our study; in contrast, verbalizing visceral experiences in RTA required them to recall such experiences after completing the game, which was more challenging than verbalizing such experiences right when they occurred as in CTA.

In sum, our research provides qualitative and quantitative evidence that despite thinking aloud (TA) requires extra efforts from older adults when playing immersive VR games, older adults do not feel verbalizing their thoughts in both CTA and RTA affect their VR game experiences and different types of verbalizations can be utilized to better pinpoint UX problems they encountered. Furthermore, our research also provides trade-offs between two TA protocols for different types of VR games. Last but not least, our study was conducted with a small number of older adults. Given the small sample size, we did not perform statistical tests. Future work should validate our findings with a larger sample size and statistical inferences.

ACKNOWLEDGEMENTS

We would like thank our reviewers for their constructive feedback. We would also like to thank our participants for their participation.

REFERENCES

Alshammari, T., Alhadreti, O. & Mayhew, P. (2015) When to ask participants to think aloud: A comparative study of concurrent and retrospective think-aloud methods. *International Journal Of Human Computer Interaction*. **6**, 48-64

Baker, S., Waycott, J., Carrasco, R., Hoang, T. & Vetere, F. (2019) Exploring the Design of Social VR Experiences with Older Adults. *Proceedings Of The 2019 On Designing Interactive Systems Conference*. pp. 303-315

Bisson, E., Contant, B., Sveistrup, H. & Lajoie, Y. (2007) Functional balance and dual-task reaction times in older adults are improved by virtual reality and biofeedback training. *Cyberpsychology Behavior*. **10**, 16-23

Bolton, J., Lambert, M., Lirette, D. & Unsworth, B. (2014) PaperDude: a virtual reality cycling exergame. *CHI'14 Extended Abstracts On Human Factors In Computing Systems*. pp. 475-478

Bowers, V. (1990) Concurrent versus retrospective verbal protocol for comparing window usability. *Doctoral Dissertations*.

Brewer, R., Morris, M. & Piper, A. (2016) " Why would anybody do this?" Understanding Older Adults' Motivations and Challenges in Crowd Work. *Proceedings Of The 2016 CHI Conference On Human Factors In Computing Systems*. pp. 2246-2257 , https://doi.org/10.1145/2858036.2858198

Chatrangsan, M. & Petrie, H. (2017) The usability and acceptability of tablet computers for older people in Thailand and the United Kingdom. *International Conference On Universal Access In Human-Computer Interaction*. pp. 156-170

Chung, J., Chaudhuri, S., Le, T., Chi, N., Thompson, H. & Demiris, G. (2015) The use of think-aloud to evaluate a navigation structure for a multimedia health and wellness application for older adults and their caregivers. *Educational Gerontology*. **41**, 916-929

Cobb, S., Nichols, S., Ramsey, A. & Wilson, J. (1999) Virtual reality-induced symptoms and effects (VRISE). *Presence: Teleoperators Virtual Environments*. **8**, 169-186

Cooke, L. (2010) Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions On Professional Communication*. **53**, 202-215

Eisapour, M., Cao, S. & Boger, J. (2018) Game Design for Users with Constraint: Exergame for Older Adults with Cognitive Impairment. *The 31st Annual ACM Symposium On User Interface Software And Technology Adjunct Proceedings*. pp. 128-130

Elling, S., Lentz, L. & De Jong, M. (2012) Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions On Professional Communication*. **55**, 206-220

Ericsson, K. & Simon, H. (1984) Protocol analysis: Verbal reports as data. The MIT Press.

Fan, M., Li, Y. & Truong, K. (2020) Automatic Detection of Usability Problem Encounters in Think-aloud Sessions. *ACM Transactions On Interactive Intelligent Systems (TiiS)*. **10**, 1-24, https://doi.org/10.1145/3385732

Fan, M., Lin, J., Chung, C. & Truong, K. (2019) Concurrent think-aloud verbalizations and usability problems. *ACM Transactions On Computer-Human Interaction (TOCHI)*. **26**, 1-35, https://doi.acm.org/10.1145/3325281

Fan, M., Shi, S. & Truong, K. (2020) Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey.. *Journal Of Usability Studies*. **15**, 85-102

Fan, M., Wu, K., Zhao, J., Li, Y., Wei, W. & Truong, K. (2020) VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions On Visualization And Computer Graphics (TVCG)*. **26**, 343-352 , http://dx.doi.org/10.1109/TVCG.2019.2934797

Fan, M., Yang, X., Yu, T., Liao, V. & Zhao, J. (2022) Human-AI Collaboration for UX Evaluation: Effects of Explanations and Synchronization. *Proceedings of the ACM on Human-Computer Interaction*. **6(CSCW1)**, 1-32. https://doi.org/10.1145/3512943

Fan, M., Zhao, Q. & Tibdewal, V. (2021) Older Adults' Think-Aloud Verbalizations and Speech Features for Identifying User Experience Problems. *Proceedings Of The 2021 CHI Conference On Human Factors In Computing Systems*. https://doi.org/10.1145/3411764.3445680

Finkelstein, S., Nickel, A., Lipps, Z., Barnes, T., Wartell, Z. & Suma, E. (2011) Astrojumper: Motivating exercise with an immersive virtual reality exergame. *Presence: Teleoperators And Virtual Environments*. **20**, 78-92

Gamito, P., Oliveira, J., Morais, D., Coelho, C., Santos, N., Alves, C., Galamba, A., Soeiro, M., Yerra, M., French, H. & Others (2019) Cognitive stimulation of elderly individuals with instrumental virtual reality-based activities of daily life: pre-post treatment study. *Cyberpsychology, Behavior, And Social Networking*. **22**, 69-75

Harms, P. (2019) Automated usability evaluation of virtual reality applications. *ACM Transactions On Computer-Human Interaction (TOCHI)*. **26**, 1-36

Hertzum, M., Borlund, P. & Kristoffersen, K. (2015) What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal Of Human-Computer Interaction*. **31**, 557-570

Huang, M., Hansen, D. & Xie, B. (2012) Older adults' online health information seeking behavior. *Proceedings Of The 2012 IConference*. pp. 338-345

Huygelier, H., Schraepen, B., Ee, R., Abeele, V. & Gillebert, C. (2019) Acceptance of immersive head-mounted virtual reality in older adults. *Scientific Reports*. **9**, 1-12

Iskander, J., Hossny, M. & Nahavandi, S. (2018) A review on ocular biomechanic models for assessing visual fatigue in virtual reality. *IEEE Access*. **6** pp. 19345-19361

Laver, K., Lim, F., Reynolds, K., George, S., Ratcliffe, J., Sim, S. & Crotty, M. (2012) Virtual reality grocery shopping simulator: Development and usability in neurological rehabilitation. *Presence*. **21**, 183-191

Li, Q. (2010) Effect of forest bathing trips on human immune function. *Environmental Health And Preventive Medicine*. **15**, 9-17

Lin, C., Neafsey, P. & Strickler, Z. (2009) Usability testing by older adults of a computer-mediated health communication program. *Journal Of Health Communication*. **14**, 102-118

Luger, T., Houston, T. & Suls, J. (2014) Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *Journal Of Medical Internet Research*. **16**, e16

Mandryk, R., Inkpen, K. & Calvert, T. (2006) Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour Information Technology*. **25**, 141-158

McDonald, S., Edwards, H. & Zhao, T. (2012) Exploring think-alouds in usability testing: An international survey. *IEEE Transactions On Professional Communication*. **55**, 2-19

McDonald, S. & Petrie, H. (2013) The effect of global instructions on think-aloud testing. *Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*. pp. 2941-2944

McDonald, S., Zhao, T. & Edwards, H. (2013) Dual verbal elicitation: the complementary use of concurrent and retrospective reporting within a usability test. *International Journal Of Human-Computer Interaction*. **29**, 647-660

Mirelman, A., Rochester, L., Maidan, I., Del Din, S., Alcock, L., Nieuwhof, F., Rikkert, M., Bloem, B., Pelosin, E., Avanzino, L. & Others (2016) Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial. *The Lancet*. **388**, 1170-1182

Nielsen, J. Usability engineering. (Elsevier,1994)

Nielsen, J. Severity Ratings for Usability Problems. Nielsen Norman Group: Evidence-Based User Experience Research, Training, And Consulting. (1994), https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

Nielsen, J. (2014) Demonstrate Thinking Aloud by Showing Users a Video. *nielsen Norman Group: Evidence-Based User Experience Research, Training, And Consulting*

Norman, D. (2013) The design of everyday things: Revised and expanded edition. Basic books.

Olmsted-Hawala, E. & Bergstrom, J. (2012) Think-aloud protocols: Does age make a difference. *Proceedings Of Society For Technical Communication (STC) Summit, Chicago, IL*.

Optale, G., Urgesi, C., Busato, V., Marin, S., Piron, L., Priftis, K., Gamberini, L., Capodieci, S. & Bordin, A. (2010) Controlling memory impairment in elderly adults using virtual reality memory training: a randomized controlled pilot study. *Neurorehabilitation And Neural Repair*. **24**, 348-357

Park, S. & Mattson, R. (2009) Ornamental indoor plants in hospital rooms enhanced health outcomes of patients recovering from surgery. *The Journal Of Alternative And Complementary Medicine*. **15**, 975-980

Park, S. & Lee, G. (2020) Full-immersion virtual reality: Adverse effects related to static balance. *Neuroscience Letters*. **733** pp. 134974

Rendon, A., Lohman, E., Thorpe, D., Johnson, E., Medina, E. & Bradley, B. (2012) The effect of virtual reality gaming on dynamic balance in older adults. *Age And Ageing*. **41**, 549-552

Roberts, A., De Schutter, B., Franks, K. & Radina, M. (2019) Older adults' experiences with audiovisual virtual reality: perceived usefulness and other factors influencing technology acceptance. *Clinical Gerontologist*. **42**, 27-33

Soure, E., Kuang, E., Fan, M. & Zhao, J. (2021) CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Transactions On Visualization And Computer Graphics*. **28**, 643-653

Syed-Abdul, S., Malwade, S., Nursetyo, A., Sood, M., Bhatia, M., Barsasella, D., Liu, M., Chang, C., Srinivasan, K., Raja, M. & Others. (2019) Virtual reality among the elderly: a usefulness and acceptance study from Taiwan. *BMC Geriatrics*. **19**, 223

Ulrich, R. (1981) Natural versus urban scenes: Some psychophysiological effects. *Environment And Behavior*. **13**, 523-556

Ulrich, R., Simons, R., Losito, B., Fiorito, E., Miles, M. & Zelson, M. (1991) Stress recovery during exposure to natural and urban environments. *Journal Of Environmental Psychology*. **11**, 201-230

Van Den Haak, M. & De Jong, M. (2003) Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. *IEEE International Professional Communication Conference, 2003. IPCC 2003. Proceedings.*, 285-287

Van Den Haak, M., De Jong, M. & Jan Schellens, P. (2003) Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour Information Technology*. **22**, 339-351

Van den Haak, M., De Jong, M. & Schellens, P. (2006) Constructive Interaction: An Analysis of Verbal Interaction in a Usability Setting. *IEEE Transactions On Professional Communication*. **49**, 311-324

Wüest, S., Borghese, N., Pirovano, M., Mainetti, R., Langenberg, R. & Bruin, E. (2014) Usability and effects of an exergame-based balance training program. *Games For Health: Research, Development, And Clinical Applications*. **3**, 106-114

Zhao, T. & McDonald, S. (2010) Keep talking: an analysis of participant utterances gathered using two concurrent think-aloud methods. *Proceedings Of The 6th Nordic Conference On Human-Computer Interaction: Extending Boundaries*. pp. 581-590

Zhao, T., McDonald, S. & Edwards, H. (2014) The impact of two different think-aloud instructions in a usability test: a case of just following orders?. *Behaviour Information Technology*. **33**, 163-183