Engaging Participants during Selection Studies in Virtual Reality

Difeng Yu* Qiushi Zhou† Benjamin Tag‡ Tilman Dingler§ Eduardo Velloso¶ Jorge Goncalves‖

The University of Melbourne, Melbourne, Australia

ABSTRACT

Selection studies are prevalent and indispensable for VR research. However, due to the tedious and repetitive nature of many such experiments, participants can become disengaged during the study, which is likely to impact the results and conclusions. In this work, we investigate participant disengagement in VR selection experiments and how this issue affects the outcomes. Moreover, we evaluate the usefulness of four engagement strategies to keep participants engaged during VR selection studies and investigate how they impact user performance when compared to a baseline condition with no engagement strategy. Based on our findings, we distill several design recommendations that can be useful for future VR selection studies or user tests in other domains that employ similar repetitive features.

Index Terms: Human-centered computing—HCI design and evaluation methods—User studies; Human-centered computing—Virtual reality; Human-centered computing—Pointing;

1 INTRODUCTION

As selecting objects is one of the fundamental tasks in virtual reality (VR), selection studies are prevalent in VR research (see reviews [2, 42]). Researchers conduct these studies to evaluate and compare different interaction techniques [4, 75], gather empirical evidence for pointing models [80, 84] and attempt to understand user selection behavior [6]. Such studies usually require participants to complete hundreds or even thousands of repetitive trials to measure user performance more accurately (e.g., [4, 41, 59, 74]). In a typical selection trial, participants are asked to select a single color target (normally a sphere), possibly among a set of distractors (often spheres in a different color), in a monochrome VR environment [4, 69, 78]. Given the repetitive and tedious nature of many of these experiments, participants are likely to become disinterested with the task at hand [46, 71, 80]. More physically demanding tasks, such as the ones that require mid-air pointing [32, 34, 83] and body/head-based selection [49, 85] can also trigger fatigue more quickly. As a result of these factors, participants may disengage from the selection task and perform poorly, which can influence the study results [12, 14, 51, 74, 80].

The challenge of participant engagement in VR selection studies was first, to the best of our knowledge, mentioned by Wingrave and Bowman in their early work [80]. In their work, they found that initially the participants were motivated to perform well in order not to fail the researcher. However, after a period of time, when participants became disengaged, their main concern was to complete the experiment so that they could leave. The researchers further argued, "this is problematic for comparing early and later trials". Surprisingly, this issue was seldom mentioned thereafter in VR selection literature.

We summarize three possible reasons for the lack of research regarding this problem: 1) researchers rely on counterbalanced study designs, for example, half of the participants complete condition A before B, with the other half finishing B before A. In this case, they assume the influence of disengagement will cancel out between different testing conditions; 2) studies commonly allow participants to take rests at different points in the experiment to help them recover; and 3) monetary incentives (e.g., vouchers) are sometimes used to motivate participants to complete the task. These methods can certainly work to some extent. However, even with a counterbalanced design, study results might still get affected because the engagement level of different individuals can decrease at different speeds [12]. Thus, the ones who disengage faster can "fail" more conditions than the others which can be harmful for small sample studies. Furthermore, research suggested that counterbalancing does not eliminate the risk of order effects [13, 43]. Having a short rest can surely recover muscle fatigue, but not necessarily the tedium. According to Flow Theory [14, 51], participants can experience boredom once they are over-skilled (because of the numerous repetitions) in the task. Finally, payments do not always ensure high engagement with the study and can sometimes undermine participants' internal motivation [15, 50]. Accordingly, a more ideal solution is that participants remain engaged throughout the experiment.

While there exists extensive research on participant motivation and engagement in other types of experiments (e.g., [3, 10, 22, 25, 39, 76]), it remains unknown if these approaches are applicable to VR selection experiments. Furthermore, it is unclear how these approaches would compare to traditional task settings, and if and how they would affect user performance. Thus, in our work, we raise the following research questions: ***RQ1.*** *Is there any evidence of participant disengagement in VR selection studies?* ***RQ2.*** *Will disengagement influence study results?* ***RQ3.*** *How can we keep participants engaged during VR selection studies?* ***RQ4.*** *How will different engagement strategies affect user performance?*

To answer these questions, we conducted a user study comparing four motivational strategies (mini-story, companion and encouragement, texture and animation, and ambient music) and a baseline (traditional selection task settings). In this paper, we first introduce works that are related to our research. We then present the study framework and the user study. After that, we discuss our findings and conclude the paper. The main contributions of this research include:

- An exploratory experiment of the participant disengagement problem during VR selection studies.
- The use and evaluation of different motivational strategies in VR selection studies and their impact on participant engagement.
- A set of design recommendations that can be useful for future VR selection studies or user experiments that employ similar repetitive features.

2 RELATED WORK

In this section, we first review earlier attempts to engage participants in selection studies. We then introduce the motivation theory, with

*e-mail: difengy@student.unimelb.edu.au
†e-mail: qiushi.zhou@unimelb.edu.au
‡e-mail: benjamin.tag@unimelb.edu.au
§e-mail: tilman.dingler@unimelb.edu.au
¶e-mail: eduardo.velloso@unimelb.edu.au
‖e-mail: jorge.goncalves@unimelb.edu.au

2642-5254/20/$31.00 ©2020 IEEE
DOI 10.1109/VR46266.2020.00-33

an emphasis on intrinsic motivation and its related essential factors. Finally, we go through the motivational strategies and intrinsic motivators that are more related to our research.

2.1 Selection Studies and Participant Engagement

While extensive research has been conducted regarding object selection in VR, little work has considered the participant engagement problem in selection studies. Wingrave and Bowman, after recognizing that participant engagement could invalidate their results, attempted to engage participants in their selection study by displaying participants' trial time in VR to foster their competitiveness and thus their engagement [80]. Cassidy *et al.* [11] modified the traditional selection task and developed *FittsFarm*, which aimed at involving children in a selection study with stylus-tablet input. The children (participants) were asked to drag and drop the apple (the first target) to feed the lion (the second target). Henze and Boll [29] deployed a gamified software to the wild for a selection study and collected a substantial amount of input data. They found a weak correlation when applying Fitts's law model to the collected data. Further, Knoche *et al.* [40] have conducted gamified selection studies in the wild and the lab. However, none of these works confirmed that there was an effect of disengagement through empirical data. Furthermore, the usefulness and potential effects of the motivational strategies remain unknown, as there was no comparison to a control group.

2.2 Motivation Theory

An extensive body of research has investigated how to make people feel motivated and energized toward certain ends (e.g., [39, 54, 55]). Among them, two types of motivations have been identified: *extrinsic motivation* and *intrinsic motivation* [62]. Extrinsic motivation refers to doing something because it leads to a separable outcome (such as a monetary reward), whereas intrinsic motivation is defined as doing something since it is inherently engaging (doing a sport because it is internally rewarding rather than to win any prizes) [7, 62, 79]. While increasing extrinsic and intrinsic motivation can both induce a higher level of engagement [62], our work focuses on intrinsic motivation, where engagement derives from the completion of the task itself.

One recent work has reviewed literature in games and psychology, and summarized six key factors that are related to intrinsic motivation [60]. The first three elements, namely *competence*, *autonomy*, and *relatedness*, come from the Self-Determination Theory (SDT) [63]. They specify people's innate needs for engaging with optimal challenges and experiencing mastery, self-organizing their own behaviors, and connecting with others [15]. In addition, the *curiosity* component is also frequently mentioned in the literature, which is the degree to which people can continue to arouse and satisfy their inquisitiveness [45]. The *immersion* element normally refers to the "suspension of disbelief" and can be composed by, for example, compelling storytelling and creating a lively game-world [27, 61]. The final factor is *domination*, which can be interpreted as people's needs to exert influence on others [60]. The six factors mentioned above align well with a more comprehensive meta-synthesis of different player types [27].

2.3 Motivational Strategies and Intrinsic Motivators

Previous works have reviewed motivational strategies and intrinsic motivators in great detail [1, 54, 55, 60]. Here, we only focus on the ones that are more relevant to our study.

Gamification [16, 67], which is "the use of game elements in non-game contexts", has been extensively applied in research to, for example, improve user performance [5, 24, 76] or increase people's engagement and motivation in doing certain activities [20, 52]. It normally refers to including design elements such as points, leaderboards, and levels which can potentially influence the need for competence and social relatedness (mainly leaderboards) and enhance participants' performance [35, 47, 56]. Other components like meaningful stories and avatars can bring a sense of autonomy, immersion, and curiosity, but do not relate to performance directly [26, 35].

A similar but different concept, which was named "juiciness", explicitly considers the condition when a player's action can trigger multiple visual and audio reactions [31, 36]. It can be the use of animation, particle, and sound effects to create positive and engaging user experience [30]. Research indicated that juiciness elements could facilitate intrinsic motivation, but not necessarily improve user performance [31].

Companion and encouragement (praise) are essential means to motivate or persuade people to reinforce their behavior [55]. Such facilitation can be carried out by the presence of virtual characters [23, 53], and they have been shown to raise people's motivation, both intrinsically and extrinsically [19]. They may also increase social relatedness, especially when a user is wearing a VR headset, which "isolates" them from the real world. Previous work suggests that participants performed a simple task better, but a challenging one worse, in the presence of a virtual character in AR [48].

Music is an effective remedy for anxiety which is especially harmful to study engagement according to Flow Theory [14, 51, 68]. Music can also result in higher enjoyment [37, 70], reduce motion sickness [38], and it has been actively used in games and movies to immerse players and audiences in the experience [61]. Researchers have identified that there is a relationship between music tempo, which is the speed of the underlying beat of the music, and user performance [18, 65]. For example, participants were found to be much quicker in performing a target selection task on smartphones in music conditions (both fast and slow tempo music) comparing to a silent condition [65].

3 Study Framework

In this section, we present the study framework, including the multi-directional tapping selection task, the four study scenarios that are incorporated with different intrinsic motivators and the baseline condition. We furthermore present the eight measurements that were used for the scenario comparisons.

3.1 Experimental Task

We can classify the existing VR selection tasks into two main types: one generates targets in pre-defined and predictable locations [4, 49], the other randomizes the target position under certain constraints (e.g., a fixed distance to a home button where each selection starts from) [78]. In this research, we used the multi-directional tapping task [69] under the first type of task design as it has been standardized for selection studies [33]. In addition, the randomized design creates a higher workload for participants to collect the same amount of trials, as the cursor has to return to the home button after each selection.

Our selection task follows the standard ISO9241-9 design [33, 69] and presents 21 spherical targets organized in a circle (see Figure 1). The participants are required to select the targets in a clockwise manner, following the path shown in Figure 1E. For each trial, a participant needs to control the ray, which emanates from the hand position, to point at the highlighted goal target and presses the trigger button to confirm the selection. After the selection confirmation, a short sound will be provided to indicate the correctness of the selection. The next trial will start with a new goal target highlighted. The participant then moves the cursor to the new target. After completing one circle of 21 targets, another round of selection starts. Note that the first selection is discarded for each circle following previous research [82, 84], leaving 20 timed trials.

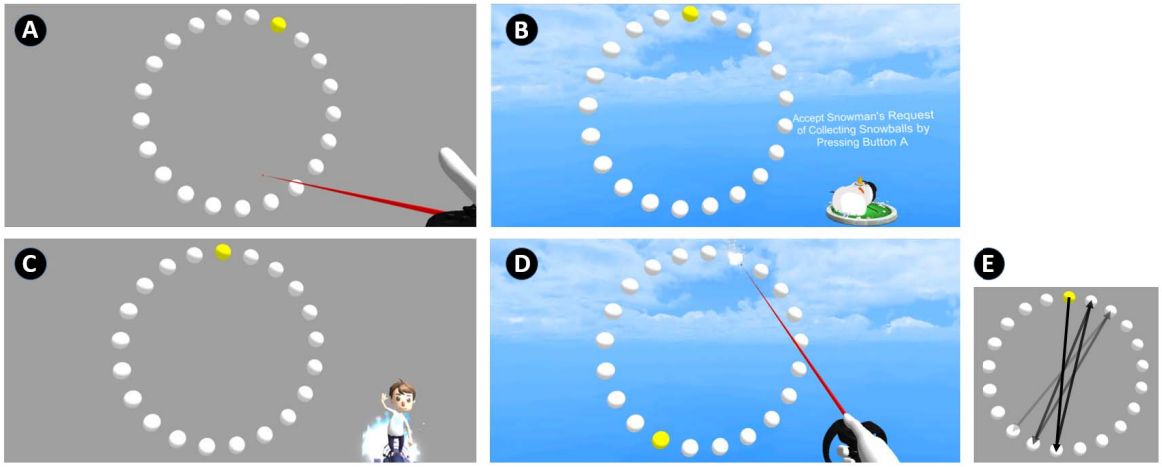

Figure 1: A demonstration of the scenarios used in this work (A-D). (A) *Baseline* (or *Music*, as the only difference between the two is the audio); (B) *MiniStory*: the blue sky and the snowman were used to immerse participants in the storyline; (C) *Coach*: a virtual coach will accompany and encourage the participants during the rest period of the experiment; and (D) *Shooting*: texture (blue sky) and animation (the explosion of the target) were provided. (E) The participants were required to follow the path indicated by the arrows to select the targets in order.

3.2 Study Scenarios

The design rationale of the scenarios was that the motivational strategies should not shift the speed-accuracy trade-off severely but positively engage participants in the study, with potentially improved overall performance in the long run. For example, approaches like a leaderboard are not ideal as they can make participants more "competitive" in the experiment, increasing selection speed but lower the accuracy, to win a better position on the board [64]. Embedding such elements are likely to strongly influence the study results, and the test results no longer represent natural selection conditions. On the other hand, using methods that can increase perceived user engagement and persist high performance in the longer term can be beneficial for such studies. Therefore, based on the rationale, we browsed through the potential strategies in the literature (e.g., [54, 55, 60]) and carefully chose the following four strategies: story (*MiniStory*), companion and encouragement (*Coach*), texture and animation (*Shooting*), and ambient music (*Music*). Across all the conditions described below, the objects (white color), target (yellow color), virtual hands, and selection ray (red color) remained consistent.

- **Baseline**. The *Baseline* scenario mimicked other VR selection study settings (e.g., [4, 77]) with a grey background (see Figure 1A). A correct or wrong sound was given according to the correctness of each selection.
- **MiniStory**. The *MiniStory* scenario contained a fully-animated virtual character (a snowman) and blue sky to immerse users in the story (see Figure 1B) [61]. Before the task started, participants were asked by the snowman to collect snowballs as it was melting because of the strong sunlight. After accepting the mission, participants then proceeded to the selection task. The snowman disappeared when the users were performing the task, in order not to distract them thus influence the performance. The snowman, which became larger because of the collected snowballs, showed up between the resting period and requested users to get more snowballs. After finishing the task, the snowman thanked the participants for their help. All scripts are documented in the supplementary material and the sound feedback was the same as *Baseline*. By bring the immersion and curiosity with the story [26, 35], we expected users to be more engaged in the study.
- **Coach**. The *Coach* scenario had another fully-animated virtual character (a coach) on a grey background (see Figure 1C) [53]. Before the task, the coach welcomed participants to the selection study, and then moved to the right side of the users, saying "I will be accompanying you in this scenario". Participants were not able to see the character when performing the task, if not rotating their head to the right. This ensured that the character did not influence user performance directly. After finishing certain blocks, the coach would say some encouraging words to inspire the participants [55]. The scripts are documented in the supplementary material, and the sound feedback was the same as *Baseline*. As suggested by previous work (e.g., [18]), the encouragement and companionship provided by the virtual character can possibly raise participants' motivation and increase their engagement level.
- **Shooting**. The *Shooting* scenario used a blue sky texture to immerse users in this game-like scenario. Animation and particle effects (the explosion of the target) would be triggered when users selected the right object (see Figure 1D). In the meantime, a smashing sound would be played to simulate the explosion of the target. The "juiciness" elements [31, 36] happened once the selection was made; thus, we hypothesized that they would not affect user performance directly. We expected that these juiciness elements could create positive and engaging user experience [30].
- **Music**. Previous work suggests that music tempo has a substantial impact on selection performance [65]. Therefore, to leverage its benefits, we used ambient music [1], which does not have a detectable tempo. The music was played throughout the whole scenario. The visual settings and sound feedback were the same as the baseline condition. Since music can provide various benefits such as heal anxiety [68] and lead to higher level of enjoyment [70], participants could possibly be more engaged in the study.

3.3 Measurements

- **Selection Time**: The elapsed time between when the target appearing and the selection being made (by pressing the trigger).
- **Error Rate**: The percentage of error trials for each condition.
- **Throughput (TP)**: The unified term which combines both speed and accuracy. We calculated TP via Equation 1, where ID_e is the effective index of difficulty, MT is the selection time, A_e is the average actual movement distance, and $SD_{x,y}$ is the bivariate

[1] From `https://youtu.be/RQcLIm-s75U`.

endpoint deviation (more details in previous work [69, 72, 73, 82, 84]). Throughput is independent from the speed-accuracy trade-off, which allows us to compare "fast but reckless" and "slow but careful" selections [44, 82].

$$TP = \frac{ID_e}{MT} = \frac{\log_2\left(\frac{A_e}{4.133 \times SD_{x,y}} + 1\right)}{MT} \quad (1)$$

- **Self-Assessment Manikin (SAM) Questionnaire** [9]: A pictorial assessment technique that measures a person's emotion response (pleasure, arousal, and dominance) towards stimulus. We used the 5-point SAM scales in this study.
- **User Engagement Scale Short Form (UES-SF)** [57]: A questionnaire that quantifies user engagement in interaction tasks. Since not all subscales are relevant to our study, we only used focused attention (FA, the feeling of absorption and losing track of time), aesthetic appeal (AE, the attractiveness and virtual appeal of the interface), and reward (RW, the worthiness and interestingness of the experience) factors, with three questions for each factor. We used 5-point scales for each question.
- **Raw NASA-TLX** [28]: An assessment tool that rates the perceived workload. The total workload is composed of six subscales rated on 5-point scales, including mental demand, physical demand, temporal demand, performance, effort, and frustration.
- **Semi-Structured Interviews**. We conducted a semi-structured interview with each participant at the end of the experiment. We were interested in determining what motivated participants to complete the experiment. We also discussed the study disengagement problem and potential ways to make selection studies more engaging with participants.
- **Observations**: We recorded observation notes regarding participants' task performance. The observations focused on participants' posture, facial expression, and their interactions in the virtual environment (from a computer screen that duplicated what participants saw in VR).

4 User Study

The purpose of the study is to explore and answer the research questions (***RQ1-4***) raised earlier in the paper. To achieve that, we compare four scenarios based on different motivational strategies and the baseline scenario, which mimics traditional selection studies in VR.

4.1 Experimental Design and Procedure

The study employed a 5 × 6 within-subjects design with two factors: Scenario and Block. The Scenario factor is composed of *Baseline*, *MiniStory*, *Coach*, *Shooting*, and *Music*. The Block factor has 6 subsequent blocks, each contains 60 selection trials. The order of the Scenario was counterbalanced using a Latin Square design.

The whole experiment lasted about 40-50 minutes for each participant, which is comparable to other selection studies (e.g., [41, 77]). Participants were first introduced to the experiment, which was said to be "a selection study in virtual reality", and signed a consent form. They then filled in a pre-experiment questionnaire for their demographic information. After that, they were invited to wear the VR headset and were instructed how to perform the selection tasks. We informed the participant to complete the tasks as quickly and as accurately as possible, while no specific accuracy requirement was placed. Next, they proceeded to the warm-up phase, with 20 practice trials (in the *Baseline* setting). They were allowed to raise questions at this point. The formal experiment was divided into five sections corresponding to the evaluation of five scenarios. In each section, they first practiced the selection for 20 trials, and then proceeded to the timed trials. They were asked (by our program) to take a rest every 120 trials and were requested to complete the questionnaires after finishing each scenario, namely SAM, UES-SF, and Raw NASA-TLX, as mentioned in Study Framework. The questionnaires were presented in VR as previous work has shown that this can reduce study duration and user disorientation [66]. Participants were compensated with a $10 voucher after completing the experiment.

4.2 Participants, Apparatus, and Materials

We recruited 21 participants (10F/11M), aged between 19-39 (mean = 24.5 ± 4.6), from diverse educational backgrounds from a local university campus. One participant was removed from the final analysis due to severe disengagement (discussed later). The remaining participants had normal or corrected-to-normal vision, and rated their familiarity with VR as moderate (average 3.5 ± 0.9 on a 5-point scale).

Participants wore an Oculus Rift headset and interacted with our application using an Oculus Touch wireless controller. The experiment was conducted on an Intel Core i9 processor PC with a dedicated NVIDIA GTX 1080 graphics card. The software was developed using C#.NET and ran on the Unity3D platform.

4.3 Empirical Results

In total, we collected 36,000 data points (20 participants × 5 scenarios × 6 blocks × 60 repetitions) from the experiment. We removed the outliers that deviated by more than three standard deviations from the averaged selection time in each condition, a common practice in these type of experiments (e.g., [84]). As a result, 479 data points (∼1.3%) were discarded, leaving 35,521 data points for analysis.

4.3.1 General Trend

To avoid ambiguity with 'Block', we define a new term called *SequenceBlock*, which refers to the block number ordered in time sequence over the whole experiment. For example, SequenceBlock 10 means the fourth block of trials during the second scenario in the entire experiment. Considering we have 5 scenarios and 6 blocks, there are 30 SequenceBlocks in total.

From Figure 2, we were able to identify a clear trend that, as the experiment went on, the averaged selection time of all participants kept decreasing. The error rate was relatively stable in the first half of the experiment (until around SequenceBlock 15) but started increasing in the latter half. The throughput steadily increased up to SequenceBlock 20, and then stabilised for the remaining blocks.

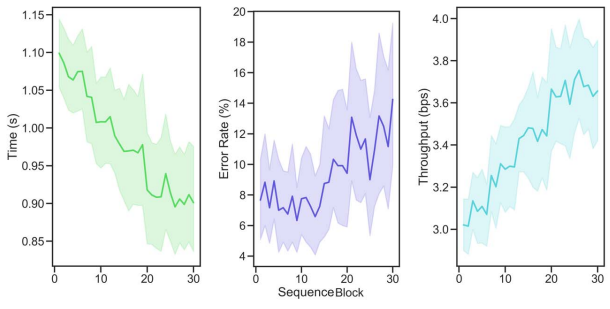

Figure 2: The general trend of selection time (left), error rate (middle), and throughput (right) over the whole experiment (5 Scenarios × 6 Blocks). The error bands indicate ± 95% confidence intervals.

4.3.2 Repeated Measures Analysis

Selection Time. Shapiro-Wilk tests indicate that the selection time data of one condition (*Shooting*, Block 4) was not normally distributed ($p = .009$). To conduct repeated measures ANOVA (RM-ANOVA), a Box-Cox transformation with $\lambda = 0$ (log-transformation) was applied to correct non-normal residuals [4, 8]. A

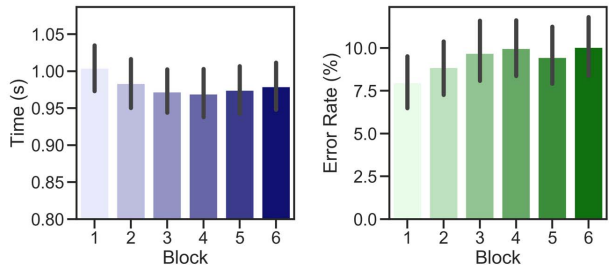

Figure 3: The averaged selection time (left) and error rate (right) of six blocks. The error bars indicate ± 95% confidence interval.

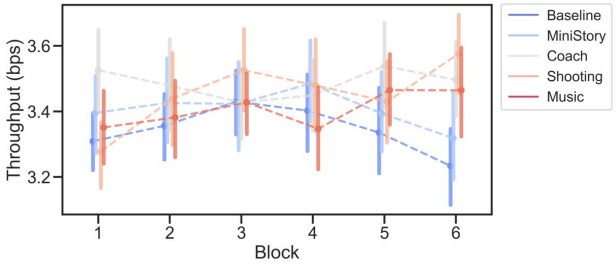

Figure 4: The averaged throughput of different scenarios over six blocks. The error bands indicate ± 1 standard error. The throughput of the *Baseline* scenario kept dropping in the last four blocks.

RM-ANOVA[2] indicated that BLOCK ($F_{5,95} = 6.14, p < .001, \eta_G^2 < .01$) had a significant main effect on selection time, but not SCENARIO ($F_{2.48,47.06} = 0.80, p = .478, \eta_G^2 = .01$). Pairwise comparisons found significant differences between Block 1 and Block 2-5 (all $p < .050$), and a marginally significant difference with Block 6 ($p = .077$). An interaction between SCENARIO and BLOCK was identified ($F_{20,380} = 1.85, p = .015, \eta_G^2 < .01$); however, no clear results could be drawn from the interaction effect.

Error Rate. The error rate data underwent pre-processing through Aligned Rank Transform (ART) [81] to take into account the non-normal distribution. RM-ANOVA tests reveal that both SCENARIO ($F_{4,551} = 5.06, p < .001$) and BLOCK ($F_{5,551} = 2.67, p = .021$) had significant main effects on error rate. The averaged error rates across six blocks are presented in Figure 3, right. Post-hoc pairwise comparisons showed that *Baseline* and *Shooting* ($p = .003$), *MiniStory* and *Shooting* ($p = .023$), *Music* and *Shooting* ($p = .002$) were significantly different from each other. In addition, differences in error rate between Block 1 and Block 4 ($p = .026$), Block 1 and Block 6 ($p = .029$) were statistically significant. No interaction effects were found in the error rate data ($F_{20,551} = 0.58, p = .926$).

Throughput. Shapiro-Wilk tests show that the throughput data was normally distributed (all $p > .1$). RM-ANOVA tests indicate that both SCENARIO ($F_{2.53,48.03} = 0.53, p = .636, \eta_G^2 < .01$) and BLOCK ($F_{5,95} = 1.48, p = .203, \eta_G^2 < .01$) did not have a significant effect on throughput. A marginal interaction effect was found between SCENARIO and BLOCK ($F_{9.26,175.94} = 1.84, p = .063, \eta_G^2 = .01$).

4.3.3 Analysis within Blocks

As shown Figure 4, the throughput of *Baseline* dropped continuously from the third block onward. This is also indicated by a RM-ANOVA test on *Baseline* with later four blocks ($F_{3,57} = 3.48, p = .021, \eta_G^2 = .02$) [3]. Pairwise comparisons show that Block 6 had a significantly lower throughput than Block 3 ($p = .037$). Therefore, we conducted paired-sample t-tests on the last block (where the scenarios also diverged the most) to assess how the throughput of *Baseline* differed from other scenarios (*MiniStory*, *Coach*, *Shooting*, and *Music*). We summarize the detailed results, including t-tests, Cohen's *d* (effect size), mean value, and 95% confidence interval for each scenario in Table 1. The results suggest that *Coach*, *Shooting*, and *Music* scenarios had significantly higher throughput than *Baseline* in the last block (Block 6), with medium effect sizes.

Table 1: The table summarizes the results from paired-sample t-tests (upper-tailed, significant level $\alpha = 0.05$, comparing *Baseline* to the other four scenarios), effect size based on Cohen's *d*, mean value (μ), and 95% confidence interval (CI) for each scenario in the last block.

Scenario	t	p	Sig?	d	μ	CI
Baseline	-	-	-	-	3.23	[3.01, 3.46]
MiniStory	0.80	.217	no	0.15	3.32	[3.05, 3.58]
Coach	2.08	.026	**yes**	0.50	3.50	[3.27, 3.72]
Shooting	2.60	.009	**yes**	0.65	3.58	[3.34, 3.81]
Music	1.86	.039	**yes**	0.40	3.46	[3.19, 3.74]

4.3.4 Analysis based on Individuals

Four representative individual results are shown in Figure 5. The data trends of P2 and P4 imply disengagement, while P11 had relatively stable performance. We also included P20, whose data were eliminated from the analysis, because of the severe disengagement with a high error rate throughout the whole study (except for the first block).

4.4 Questionnaire Results

We collected a total of 1,800 questionnaire answers (20 participants × 5 scenarios × 18 questions) from the experiment. During the data analysis procedure, when using RM-ANOVA tests, the subjective data were pre-processed through ART [81] to take any non-normal distributions into account.

4.4.1 Validity Checking

RM-ANOVA tests were conducted on all scales (three emotion responses, three engagement subscales, and six workload ratings) across the scenario appearance order (from the first scenario experienced to the last scenario). No significant difference was found (all $p > .05$). This suggests that participants kept a relatively consistent rating from the start of the experiment to the end; rather than, for example, giving later scenarios a lower score because they were affected by previous scenarios.

4.4.2 Repeated Measures Analysis

Emotion Response. A RM-ANOVA test shows that SCENARIO did not have a significant effect on pleasure ($F_{4,76} = 1.47, p = .220$), arousal ($F_{4,76} = 0.34, p = .846$), or dominance ($F_{4,76} = 1.30, p = .280$).

Engagement. RM-ANOVA tests indicate that SCENARIO had significant main effects on focused attention ($F_{4,76} = 11.78, p < .001$), aesthetic appeal ($F_{4,76} = 3.78, p = .007$), and reward ($F_{4,76} = 5.42, p < .001$). Pairwise comparisons revealed that *MiniStory*, *Shooting*, and *Music* led to higher focused attention values than *Baseline* (all $p < .001$). In addition, the focused attention of *Shooting* was significantly higher than *Coach* ($p = .001$). The post-hoc tests also show that *Shooting* ($p = .011$) and *Music* ($p = .034$) had higher aesthetic appeal rating than *Baseline*. Moreover, *MiniStory*

[2]For all RM-ANOVA tests, Greenhouse-Geisser adjustments were applied when the sphericity assumption was violated and Bonferroni corrections were used in pairwise comparisons.

[3]The RM-ANOVA result on *Baseline* with all blocks was $F_{5,95} = 2.21, p = .059, \eta_G^2 = .02$ for reference.

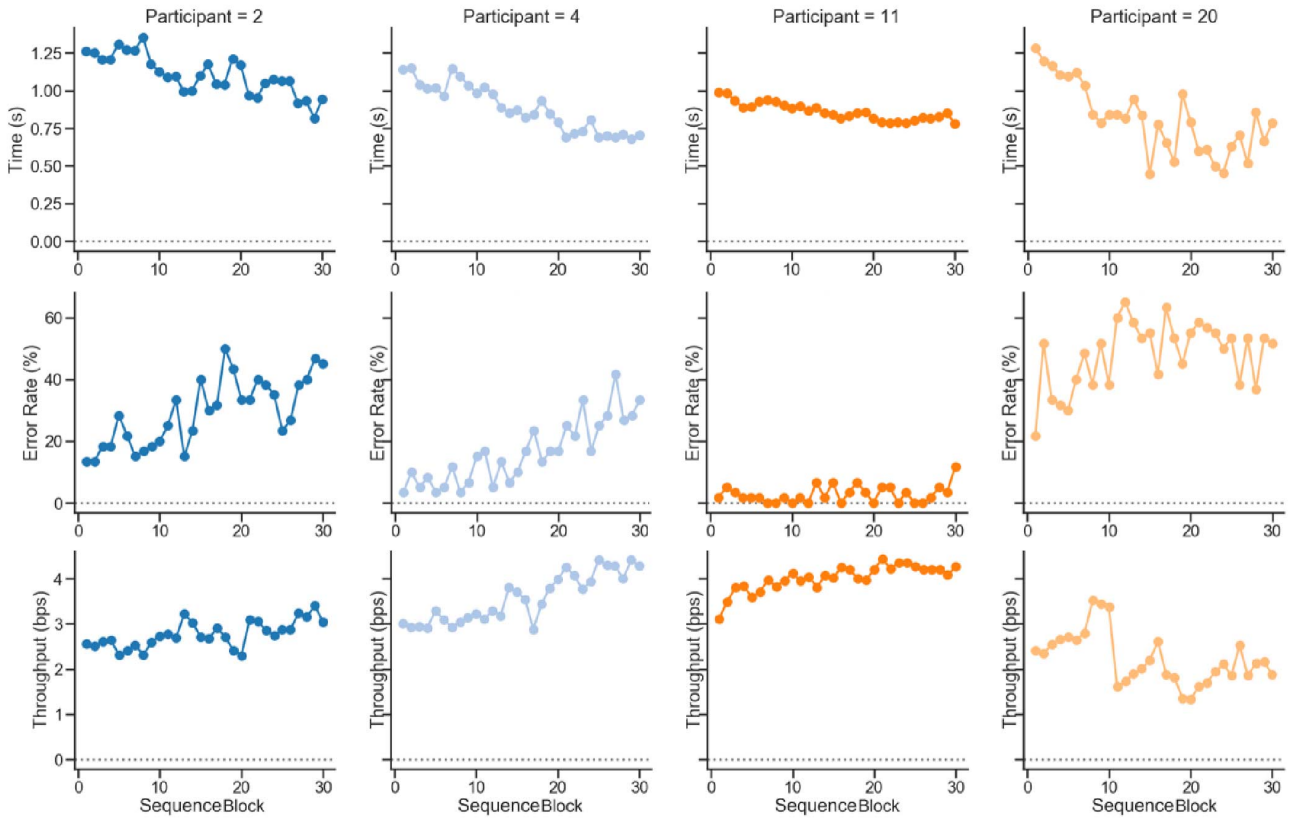

Figure 5: The selection time, error rate, and throughput of four different individuals over the whole experiment.

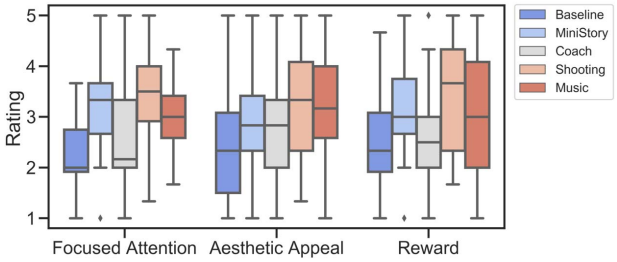

Figure 6: The box plots from the results of user engagement scales.

($p = .031$) and *Shooting* ($p = .001$) had higher reward rating than *Baseline*. The reward value of *Shooting* was also higher than *Coach* ($p = .020$). The results of the user engagement scales are summarized in Figure 6.

Workload. RM-ANOVA tests did not show significant effects of SCENARIO on perceived workload: Mental Demand ($F_{4,76} = 0.46, p = .765$), Physical Demand ($F_{4,76} = 1.20, p = .318$), Temporal Demand ($F_{4,76} = 0.70, p = .595$), Performance ($F_{4,76} = 1.17, p = .330$), Effort ($F_{4,76} = 0.15, p = .963$), and Frustration ($F_{4,76} = 0.58, p = .679$).

4.5 Interviews and Observations

When asked about what motivated them to finish the study, 10 participants explicitly mentioned that it was because they wanted to leave the experiment. One stated, for example, "*I want to finish the five scenarios as soon as possible*". Some others were keen to receive their vouchers, saying like "*I want to have the voucher and leave*". Among the ten participants, two of them brought up the commitment they made before the study. "*I have to finish because I've committed to doing so*", as one said. Interestingly, two participants reported that they wanted to 'test' themselves during the experiment. One (P3) noted that "*I want to see how focused I can be during the study*". The other (P15) answered that "*I want to examine myself... I feel competitive during the task, and I want to be the best in the study*", although the participants knew that researchers might not be interested in finding the best one. Some participants said that they were curious about the purpose of the study (N=2), or their intention was to experience VR studies (N=2).

During the interviews, several participants explicitly mentioned that the snowman (N=9), the shooting game (N=7), the music (N=9), and the encouragement from the VR character (N=1) were helpful, while the baseline condition was tedious and boring (N=2). "*It (the experiment) was actually not that boring, and the shooting element was the best*", "*The animation was quite interesting*", "*The music one (scenario) was very nice*", uttered by different participants. However, negative comments (N=2) also existed towards the engagement methods. For example, P15 stated that "*I tried not to listen to what was said by the snowman and the boy [coach] because they felt artificial*".

Participants maintained different sitting postures during the experiment according to our observations. Participants first started in a comfortable sitting position. As the experiment proceeded, some changed to other postures, including crossing the legs, leg-shaking, tilting the head, etc. Some of them yawned at certain points during the study and kept adjusting the headset. Noticeably, two participants held the shooting posture in the study, and one of the two

was actively interacting with the virtual character before the formal experiment started (e.g., waving the hands). One participant, whose data were discarded as mentioned before, started selecting objects randomly after two scenarios.

5 Discussion

In this section, we first discuss the answers to the research questions (***RQ1-4***), which were raised at the beginning of the paper. We then present other findings that are related to study engagement. Finally, we offer recommendations for future selection studies.

5.1 Research Questions

RQ1. *Is there any evidence of participant disengagement in VR selection studies?* Yes, three factors from our study suggest that participants are likely to disengage from the experiment in the *Baseline* scenario, which was designed to be similar to current practices in selection studies. First, the throughput of the baseline condition dropped significantly in the last four blocks, as shown in Figure 4, while this was not the case for the scenarios *Music*, *Shooting*, and *Coach*. This indicates that, overall, participants were not able to maintain their initial performance in the *Baseline* scenario. Second, all subscales from the user engagement questionnaire were rated lower than borderline (3) for *Baseline*. Participants found it was not able to capture their attention (focused attention scale) and was neither appealing nor rewarding. Third, qualitative data shows that participants found the *Baseline* scenario tedious and boring.

RQ2. *Will disengagement influence study results?* The short answer is that it will influence the results, but the effect size depends on the purpose and sample size of the study. In *Baseline*, user performance dropped in the last few blocks—the disengagement does matter when the absolute performance of a condition is measured. However, when the goal of the study is to compare different variables (e.g., techniques, scenarios, etc.), and when these factors are counterbalanced, the answer is not as straightforward. As we see from the individual plots, participants disengaged from the study at different points, which means that the participant who disengaged earlier would "fail" more conditions than the others. In some cases, we are probably still able to trust the results, given numbers of participants repeating the condition multiple times in a balanced order. The effect from the disengagement might be averaged out (but it certainly increases variances). Nevertheless, the disengagement problem will be much more harmful to small sample studies, as a single disengaged participant can significantly impact the final results.

RQ3. *How can we keep participants engaged during VR selection studies?* Our study results suggest that introducing a mini-story (*MiniStory*), companion and encouragement (*Coach*), texture and animation (*Shooting*), and ambient music (*Music*) all had positive effect on study engagement. The empirical results indicate that participants were able to sustain their performance throughout six blocks in the *Coach*, *Shooting*, and *Music* scenarios. In addition, *MiniStory*, *Shooting*, and *Music* generally led to participants feeling more engaged and led to higher intensity of flow experience [14] than *Baseline*. *MiniStory* and *Music* were both considered appealing to participants, while *MiniStory* and *Shooting* were seen as more rewarding than *Baseline*. The qualitative data provides further evidence that adding motivational elements can make the study experience more engaging.

RQ4. *How will different engagement strategies affect user performance?* On the one hand, engagement strategies including *Coach*, *Shooting*, and *Music* can help participants maintain their study performance (throughput) longer than *Baseline*. On the other hand, some approaches might shift the speed-accuracy trade-off [84]. While the throughput was relatively stable across the five scenarios (RM-ANOVA was not able to find statistical significant differences between SCENARIO), *Shooting* led to higher error rates than *Baseline*. That is, participants tended to be quicker but more cursory in terms of selection.

5.2 Other Findings

Apart from answering the research questions, we also identified other findings regarding user performance and subjective feedback.

Our results show that the overall trend of selection time kept decreasing, and the throughput kept increasing. We identified two possible reasons for this. First, learning effects, conceivably the acquaintance to the controller helped improve the performance (throughput). As participants got more used to the controller, they were able to complete the task faster and more accurately. Ultimately, what limits our performance with input devices is information processing [21]. The more experience the participants have with an input device, the less the device itself matters and the closer they get to the information processing limit. However, learning effects alone do not explain the raise of error rates, especially the increase around SequenceBlock 15. The second reason might be that the participants were somewhat disengaged from the study at about SequenceBlock 15, therefore, starting to shift the speed-accuracy trade-off and ignore some of the errors. That is, they became "fast and inaccurate" users from "slow and precise" users. Based on this rationale, we can reasonably infer that the overall error rate throughout the whole study might be a useful identifier for participant disengagement. Furthermore, it is also indicated that although engagement strategies can engage users relatively better than the *Baseline* setting, the effects might not be able to last for the whole experiment. As time passes, engaging strategies will become less attractive as participants become accustomed to them.

Similar patterns can be found for each scenario (within the six blocks). The statistical analysis indicates that BLOCK had significant main effects on both selection time and error rate, but not throughput. Our post-hoc reasoning revealed that the selection strategy of participants was transformed from a slower but more accurate way to a faster but more cursory approach.

In terms of subjective feedback, the engagement strategies used in this study led to higher user engagement, but not necessarily higher pleasure, arousal, dominance, or perceived workload. Different participants had different opinions toward the engagement strategies; they generally appreciated it, while a few thought some scenarios were somewhat distracting.

According to our observations, postures [17] and facial expressions [58] (although only half of the face can be seen when the user is wearing a head-mounted display) are helpful indicators of participant disengagement. For example, participants tended to cross their legs and sighed when disengaging from the experiment. In contrast, they were more likely to be fully engaged in the study if they were actively interacting with the virtual avatars (e.g., waving to the virtual character) and sometimes smiled because of the scenario contents.

5.3 Design Recommendations

Based on the study results and the discussion above, we distilled several design recommendations for future selection studies in VR. These takeaway messages might also be useful for studies that employ similar repetitive features and are outside of the field of VR.

R1. Caution about the disengagement problem during selection studies in VR is always required. Some participants may self-engage themselves throughout the whole study, while others may not. The ones who quickly disengage from the experiment can "fail" more conditions than the others, leading to unfair comparisons. The problem can be mitigated by increasing sample size and balancing the study design. However, awareness of the issue of disengagement can be particularly harmful to the results of small sample studies.

R2a. Identification of any evidence of participant disengagement through the following three ways: **(1)** During the experiment, observation if there is any boredom behaviour or expression from participants. **(2)** Ahead of the formal analysis, averaging the results across the whole study to see the overall trend of different dependent variables—ideally, the performance should be increasing (because of the learning effect) or relatively constant. **(3)** Diagnosing the overall trend to see if there is any abnormal increment of error rates or decrease of performance (like selection time and throughput) as the study goes on. If there is, it is likely to suggest that participants were, to some extent, disengaged from the experiment at that point.

R2b. If participants were determined to be disengaged from the study, removing the data of those participants to preserve the reliability of the collected data has to be considered.

R3. If possible, run pilot studies before the formal one and determine the length of the experiment through the overall performance trend. Desirably, the trials should end before the point where the global trend of error rate sharply increases, or the performance clearly decreases. It is not ideal to run overly lengthy and repetitive studies, where participants are likely to be disengaged in the latter part of the study.

R4. Consider using motivation strategies when the experiment is long and repetitive. Our study results suggest introducing ministory, companion and encouragement, texture and animation, and ambient music can all increase the engagement level. However, be cautious about their potential impact on user performance. For example, when using the texture and animation elements (*Shooting* in our case), remember that they may shift the speed-accuracy trade-off of participants. Furthermore, do not apply complex visual/audio effects, as they can be distracting during the studies.

5.4 Limitations

We identify some limitations of our study. First, we were not able to completely cancel out the fatigue effect in the study, although enough breaks were given throughout the whole experiment, and the chosen input technique was not very demanding. Physical fatigue might interplay with the disengagement factor, and either can have an impact on user performance. Second, varying target width and movement amplitude could better simulate real selection tasks. However, as we wanted to assess user performance across the whole study, fixing the factors allowed us to compare performance between different blocks. Finally, more complex physiological sensors, like EEG and eye-tracking, could be useful to detect disengaged participants. However, no standardized methods exist for identifying and analyzing disengagement with such technologies, and as such future work on this topic is needed.

6 Toolbox for Selection Study in VR

We open-source a toolbox to facilitate the design of more engaging object selection tasks and to speed up VR selection research. Although some evaluation gadgets have been developed in previous works [82], we found no such tools existed in VR. Our tool is based on the standard ISO9241-9 task [33, 69], and runs on the Unity3D platform with C# scripts (no need to install dependencies if not accessing the scripts). The target widths and movement amplitudes can be easily adjusted through text input. Our tools can record data, including selection time, errors, selection endpoints, with an option of recording the selection trace. All selection trials are logged in .txt files. Engagement strategies (including *MiniStory*, *Coach*, *Shooting*, and *Music*) can be enabled if required. The tool and its source code can be downloaded from `https://github.com/Davin-Yu/EngageWithVRSelection`.

7 Conclusion and Future Work

Through this research, we elicited evidence that participant disengagement problems existed in traditional VR selection study settings. Participants can feel bored during repetitive selection tasks and their performance decreases as the study proceeds. Moreover, different participants disengaged from the experiment at different speeds, which can cause variances in study results. To deal with the problem, four engagement strategies, including mini-story, companion and encouragement, texture and animation, and music, were empirically evaluated. Our results show that all approaches improve the engagement level of the participants, but depending on the methods used, they might shift the speed-accuracy trade-off. We further distilled a set of design recommendations that can be useful for future VR selection studies, as well as experiments that employ similar repetitive features. We encourage researchers to be aware of the disengagement problem that can appear in user studies. We also expect future research on this topic using different motivational strategies that can lead to further research outcomes.

Acknowledgments

We thank our participants for their time, and the reviewers for their professionalism and dedication that helped improve our paper.

References

[1] M. L. Ambrose and C. T. Kulik. Old friends, new faces: motivation research in the 1990s. *Journal of Management*, 25(3):231 – 292, 1999. doi: 10.1016/S0149-2063(99)00003-3

[2] F. Argelaguet and C. Andujar. A survey of 3d object selection techniques for virtual environments. *Computers & Graphics*, 37(3):121 – 136, 2013. doi: 10.1016/j.cag.2012.12.003

[3] T. August and K. Reinecke. Pay attention, please: Formal language improves attention in volunteer and paid online experiments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 248:1–248:11. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300478

[4] M. Baloup, T. Pietrzak, and G. Casiez. Raycursor: A 3d pointing facilitation technique based on raycasting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 101:1–101:12. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300331

[5] S. C. Barathi, D. J. Finnegan, M. Farrow, A. Whaley, P. Heath, J. Buckley, P. W. Dowrick, B. C. Wuensche, J. L. J. Bilzon, E. O'Neill, and C. Lutteroth. Interactive feedforward for improving performance and maintaining intrinsic motivation in vr exergaming. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 408:1–408:14. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173982

[6] M. D. Barrera Machuca and W. Stuerzlinger. The effect of stereo display deficiencies on virtual hand pointing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 207:1–207:14. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300437

[7] M. V. Birk, R. L. Mandryk, and C. Atkins. The motivational push of games: The interplay of intrinsic motivation and external rewards in games for training. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, pp. 291–303. ACM, New York, NY, USA, 2016. doi: 10.1145/2967934.2968091

[8] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964. doi: 10.1111/j.2517-6161.1964.tb00553.x

[9] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49 – 59, 1994. doi: 10.1016/0005-7916(94)90063-9

[10] A. Caraban, E. Karapanos, D. Gonçalves, and P. Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human*

Factors in Computing Systems, CHI '19, pp. 503:1–503:15. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300733

[11] B. Cassidy, J. C. Read, and I. S. MacKenzie. Fittsfarm: Comparing children's drag-and-drop performance using finger and stylus input on tablets. In *IFIP Conference on Human-Computer Interaction*, pp. 656–668. Springer, 2019.

[12] M. S. Christian, A. S. Garza, and J. E. Slaughter. Work engagement: A quantitative review and test of its relations with task and contextual performance. *Personnel psychology*, 64(1):89–136, 2011.

[13] A. Cockburn and C. Gutwin. Anchoring effects and troublesome asymmetric transfer in subjective ratings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300592

[14] M. Czikszentmihalyi. *Flow: The psychology of optimal experience*. New York: Harper & Row, 1990.

[15] E. L. Deci and R. M. Ryan. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4):227–268, 2000.

[16] S. Deterding, R. Khaled, L. E. Nacke, and D. Dixon. Gamification: Toward a definition. In *CHI 2011 gamification workshop proceedings*, vol. 12. Vancouver BC, Canada, 2011.

[17] S. S. D'Mello, P. Chipman, and A. Graesser. Posture as a predictor of learner's affective engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, 2007.

[18] J. Edworthy and H. Waring. The effects of music tempo and loudness level on treadmill exercise. *Ergonomics*, 49(15):1597–1610, 2006. doi: 10.1080/00140130600899104

[19] A. Eyck, K. Geerlings, D. Karimova, B. Meerbeek, L. Wang, W. IJsselsteijn, Y. De Kort, M. Roersma, and J. Westerink. Effect of a virtual coach on athletes' motivation. In *International Conference on Persuasive Technology*, pp. 158–161. Springer, 2006.

[20] S. S. Feger, S. Dallmeier-Tiessen, P. W. Woźniak, and A. Schmidt. Gamification in science: A study of requirements in the context of reproducible research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 460:1–460:14. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300690

[21] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[22] D. R. Flatla, C. Gutwin, L. E. Nacke, S. Bateman, and R. L. Mandryk. Calibration games: Making calibration tasks enjoyable by adding motivating game elements. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pp. 403–412. ACM, New York, NY, USA, 2011. doi: 10.1145/2047196.2047248

[23] B. Fogg. Computers as persuasive social actors. 2003.

[24] J. Goncalves, S. Hosio, D. Ferreira, and V. Kostakos. Game of words: Tagging places through crowdsourcing on public displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, pp. 705–714. ACM, New York, NY, USA, 2014. doi: 10.1145/2598510.2598514

[25] J. Goncalves, V. Kostakos, E. Karapanos, M. Barreto, T. Camacho, A. Tomasic, and J. Zimmerman. Citizen Motivation on the Go: The Role of Psychological Empowerment. *Interacting with Computers*, 26(3):196–207, 07 2013. doi: 10.1093/iwc/iwt035

[26] J. Hamari, J. Koivisto, H. Sarsa, et al. Does gamification work?-a literature review of empirical studies on gamification. In *HICSS*, vol. 14, pp. 3025–3034, 2014.

[27] J. Hamari and J. Tuunanen. Player types: A meta-synthesis. 2014.

[28] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909

[29] N. Henze and S. Boll. It does not fitts my data! analysing large amounts of mobile touch data. In *IFIP Conference on Human-Computer Interaction*, pp. 564–567. Springer, 2011.

[30] K. Hicks, P. Dickinson, J. Holopainen, K. Gerling, et al. Good game feel: An empirically grounded framework for juicy design. 2018.

[31] K. Hicks, K. Gerling, G. Richardson, T. Pike, O. Burman, and P. Dickinson. Understanding the effects of gamification and juiciness on players. In *2019 IEEE Conference on Games (CoG)*, pp. 1–8, Aug 2019. doi: 10.1109/CIG.2019.8848105

[32] K. Hinckley, R. Pausch, J. C. Goble, and N. F. Kassell. A survey of design issues in spatial input. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, UIST '94, pp. 213–222. ACM, New York, NY, USA, 1994. doi: 10.1145/192426.192501

[33] ISO. Iso 9241-9:2000. *Ergonomic requirements for office work with visual display terminals (VDTs) - Part 9 - Requirements for non-keyboard input devices*, 2000.

[34] S. Jang, W. Stuerzlinger, S. Ambike, and K. Ramani. Modeling cumulative arm fatigue in mid-air interaction based on perceived exertion and kinetics of arm motion. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 3328–3339. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025523

[35] Y. Jia, B. Xu, Y. Karanam, and S. Voida. Personality-targeted gamification: A survey study on personality traits and motivational affordances. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 2001–2013. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858515

[36] J. Juul and J. S. Begy. Good feedback for bad players? a preliminary study of 'juicy'interface feedback. In *Proceedings of first joint FDG/DiGRA Conference, Dundee*, 2016.

[37] A. Keesing, M. Ooi, O. Wu, X. Ye, L. Shaw, and B. C. Wünsche. Hiit with hits: Using music and gameplay to induce hiit in exergames. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW 2019, pp. 36:1–36:10. ACM, New York, NY, USA, 2019. doi: 10.1145/3290688.3290740

[38] B. Keshavarz and H. Hecht. Pleasant music as a countermeasure against visually induced motion sickness. *Applied Ergonomics*, 45(3):521 – 527, 2014. doi: 10.1016/j.apergo.2013.07.009

[39] K. Knaving, P. Woźniak, M. Fjeld, and S. Björk. Flow is not enough: Understanding the needs of advanced amateur runners to design motivation technology. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 2013–2022. ACM, New York, NY, USA, 2015. doi: 10.1145/2702123.2702542

[40] H. Knoche, A. Christensen, and S. A. Pedersen. A comparison of gamified hci studies with lab and crowd participants. *EAI Endorsed Trans. Creative Technologies*, 4(11):e1, 2017.

[41] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise head- and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 81:1–81:14. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173655

[42] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.

[43] I. S. MacKenzie. *Human-computer interaction: An empirical research perspective*. Newnes, 2012.

[44] I. S. MacKenzie and P. Isokoski. Fitts' throughput and the speed-accuracy tradeoff. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 1633–1636. ACM, New York, NY, USA, 2008. doi: 10.1145/1357054.1357308

[45] T. W. Malone. Toward a theory of intrinsically motivating instruction. *Cognitive science*, 5(4):333–369, 1981.

[46] M. Martin, G. Sadlo, and G. Stew. The phenomenon of boredom. *Qualitative Research in Psychology*, 3(3):193–211, 2006. doi: 10.1191/1478088706qrp066oa

[47] E. D. Mekler, F. Brühlmann, K. Opwis, and A. N. Tuch. Do points, levels and leaderboards harm intrinsic motivation?: An empirical analysis of common gamification elements. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, Gamification '13, pp. 66–73. ACM, New York, NY, USA, 2013. doi: 10.1145/2583008.2583017

[48] M. R. Miller, H. Jun, F. Herrera, J. Y. Villa, G. Welch, and J. N. Bailenson. Social interaction in augmented reality. *PloS one*, 14(5):e0216290, 2019.

[49] K. Minakata, J. P. Hansen, I. S. MacKenzie, P. Bækgaard, and V. Rajanna. Pointing by gaze, head, and foot in a head-mounted display. In

Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA '19, pp. 69:1–69:9. ACM, New York, NY, USA, 2019. doi: 10.1145/3317956.3318150

[50] M. Musthag, A. Raij, D. Ganesan, S. Kumar, and S. Shiffman. Exploring micro-incentive strategies for participant compensation in high-burden studies. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pp. 435–444. ACM, New York, NY, USA, 2011. doi: 10.1145/2030112.2030170

[51] J. Nakamura and M. Csikszentmihalyi. The concept of flow. In *Flow and the foundations of positive psychology*, pp. 239–263. Springer, 2014.

[52] S. Oberdörfer, D. Heidrich, and M. E. Latoschik. Usability of gamified knowledge learning in vr and desktop-3d. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 175:1–175:13. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300405

[53] C. S. Oh, J. N. Bailenson, and G. F. Welch. A systematic review of social presence: definition, antecedents, and implications. *Front. Robot. AI 5: 114. doi: 10.3389/frobt*, 2018.

[54] H. Oinas-Kukkonen and M. Harjumaa. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 24(1):28, 2009.

[55] R. Orji and K. Moffatt. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health Informatics Journal*, 24(1):66–91, 2018. doi: 10.1177/1460458216650979

[56] R. Orji, G. F. Tondello, and L. E. Nacke. Personalizing persuasive strategies in gameful systems to gamification user types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 435:1–435:14. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3174009

[57] H. L. O'Brien, P. Cairns, and M. Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28 – 39, 2018. doi: 10.1016/j.ijhcs.2018.01.004

[58] R. W. Picard. *Affective computing*. MIT press, 2000.

[59] A. A. Ramcharitar and R. J. Teather. Ezcursorvr: 2d selection with virtual reality head-mounted displays, Jan. 2018. doi: 10.20380/GI2018.15

[60] S. Roohi, J. Takatalo, C. Guckelsberger, and P. Hämäläinen. Review of intrinsic motivation in simulation-based game testing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 347:1–347:13. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173921

[61] R. Rouse III. *Game design: Theory and practice*. Jones & Bartlett Learning, 2010.

[62] R. M. Ryan and E. L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[63] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

[64] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69:371 – 380, 2017. doi: 10.1016/j.chb.2016.12.033

[65] Z. Sarsenbayeva, N. van Berkel, E. Velloso, V. Kostakos, and J. Goncalves. Effect of distinct ambient noise types on mobile interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2):82:1–82:23, July 2018. doi: 10.1145/3214285

[66] V. Schwind, P. Knierim, N. Haas, and N. Henze. Using presence questionnaires in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 360:1–360:12. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300590

[67] K. Seaborn and D. I. Fels. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74:14 – 31, 2015. doi: 10.1016/j.ijhcs.2014.09.006

[68] S. Seinfeld, I. Bergstrom, A. Pomes, J. Arroyo-Palacios, F. Vico, M. Slater, and M. V. Sanchez-Vives. Influence of music on anxiety induced by fear of heights in virtual reality. *Frontiers in psychology*, 6:1969, 2016.

[69] R. W. Soukoreff and I. S. MacKenzie. Towards a standard for pointing device evaluation, perspectives on 27 years of fitts' law research in hci. *International journal of human-computer studies*, 61(6):751–789, 2004.

[70] M. J. Stork, M. Y. Kwan, M. J. Gibala, and K. G. Martin. Music enhances performance and perceived enjoyment of sprint interval exercise. *Medicine and science in sports and exercise*, 47(5):1052–1060, 2015.

[71] L. Svendsen. *A philosophy of boredom*. Reaktion Books, 2005.

[72] R. J. Teather, A. Pavlovych, W. Stuerzlinger, and I. S. MacKenzie. Effects of tracking technology, latency, and spatial jitter on object movement. In *2009 IEEE Symposium on 3D User Interfaces*, pp. 43–50, March 2009. doi: 10.1109/3DUI.2009.4811204

[73] R. J. Teather and W. Stuerzlinger. Pointing at 3d targets in a stereo head-tracked virtual environment. In *2011 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 87–94, March 2011. doi: 10.1109/3DUI.2011.5759222

[74] R. J. Teather and W. Stuerzlinger. Visual aids in 3d point selection experiments. In *Proceedings of the 2Nd ACM Symposium on Spatial User Interaction*, SUI '14, pp. 127–136. ACM, New York, NY, USA, 2014. doi: 10.1145/2659766.2659770

[75] H. Tu, S. Huang, J. Yuan, X. Ren, and F. Tian. Crossing-based selection with virtual reality head-mounted displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 618:1–618:14. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300848

[76] N. van Berkel, J. Goncalves, S. Hosio, and V. Kostakos. Gamification of mobile experience sampling improves data quality and quantity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):107:1–107:21, Sept. 2017. doi: 10.1145/3130972

[77] L. Vanacken, T. Grossman, and K. Coninx. Exploring the effects of environment density and target visibility on object selection in 3d virtual environments. In *2007 IEEE Symposium on 3D User Interfaces*, March 2007. doi: 10.1109/3DUI.2007.340783

[78] L. Vanacken, T. Grossman, and K. Coninx. Multimodal selection techniques for dense and occluded 3d virtual environments. *International Journal of Human-Computer Studies*, 67(3):237 – 255, 2009. Current trends in 3D user interface research. doi: 10.1016/j.ijhcs.2008.09.001

[79] T. Verhagen, F. Feldberg, B. van den Hooff, S. Meents, and J. Merikivi. Understanding users' motivations to engage in virtual worlds: A multipurpose model and empirical testing. *Computers in Human Behavior*, 28(2):484 – 495, 2012. doi: 10.1016/j.chb.2011.10.020

[80] C. A. Wingrave and D. A. Bowman. Baseline factors for raycasting selection. In *Proceedings of Virtual Reality International*, 2005.

[81] 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*, CHI '11, pp. 143–146. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1978963

[82] J. O. Wobbrock, K. Shinohara, and A. Jansen. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 1639–1648. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1979181

[83] T. S. Young, R. J. Teather, and I. S. MacKenzie. An arm-mounted inertial controller for 6dof input: Design and evaluation. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 26–35, March 2017. doi: 10.1109/3DUI.2017.7893314

[84] D. Yu, H.-N. Liang, X. Lu, K. Fan, and B. Ens. Modeling endpoint distribution of pointing selection tasks in virtual reality environments. *ACM Transactions on Graphics*, 38(6), 2019. doi: 10.1145/3355089.3356544

[85] D. Yu, H.-N. Liang, X. Lu, T. Zhang, and W. Xu. Depthmove: Leveraging head motions in the depth dimension to interact with virtual reality head-worn displays. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 103–114, Oct 2019. doi: 10.1109/ISMAR.2019.00-20