Revenge of the Sick: A Meta-Analysis of Washout Periods in Cybersickness Research

Tongyu Nie*, Ville Cantory*, Courtney Hutton Pospick*, Danhua Zhang, Haoyu Tan, Sam Adeniyi, Fei Wu, Daniel Zielasko, Isayas Berhe Adhanom, and Evan Suma Rosenberg

***Abstract*—Cybersickness remains a major barrier to the widespread adoption of virtual reality (VR) technologies, motivating researchers to investigate its causes and mitigation strategies through comparative human-subjects studies. These experiments may employ either within- or between-subjects designs. Although within-subjects designs require fewer participants and potentially offer higher statistical power, researchers need to wash out carryover effects of cybersickness symptoms before each session to avoid confounding the results. Prior studies have employed washout periods of varying lengths; some required testing conditions on different days , while others allowed only short breaks of less than 15 minutes. Although shorter washout periods are more convenient for experimenters, their impact on study outcomes has not been systematically investigated. In this work, we conducted a meta-analysis to evaluate the effects of study design and washout period length on cybersickness self-reports measured with the Simulator Sickness Questionnaire (SSQ). We found that short washout periods reduce statistical power compared to long washout periods or between-subjects designs. Based on these findings, we provide guidelines to help researchers design more reliable cybersickness studies and improve the generalizability of their results.**

***Index Terms*—Virtual reality, cybersickness, research methods, meta-analysis**

I. Introduction

Comparative experiments are essential for evaluating new cybersickness mitigation techniques in Virtual Reality (VR), but they are more challenging to conduct than many other Human-Computer Interaction (HCI) studies because of carryover effects across conditions. Carryover effects are the lingering influence of a previous treatment, condition, or stimulus on a subject's response to a subsequent, different treatment [1]. When comparing different conditions, researchers can choose either a within-subjects or a between-subjects design. Between-subjects experiments avoid compounding negative symptoms between conditions, but require larger sample sizes to distinguish statistical effects. In contrast, within-subjects designs require fewer participants. Additionally, this approach can potentially offer superior statistical power, because participants experience all conditions, thereby avoiding the confounding influence of individual differences that might otherwise obscure the effects of interest.

However, carryover effects across conditions are a critical issue in within-subjects designs, particularly when studying cybersickness, because symptoms can potentially persist for hours [2]. Researchers can counterbalance conditions, where they alternate the order of conditions to mitigate the impact of carryover effects on experimental outcomes. This is a common method to control for potentially confounding factors, such as learning effects over time. Counterbalancing increases design complexity and may introduce residual variability due to sequence-related interactions that are not fully eliminated. To reduce carryover effects, researchers can incorporate washout periods. Washout periods are intervals between conditions during which participants are allowed to return to their baseline state after the first intervention before proceeding to the next. Some researchers have used a longer washout period (time outside of VR elapsed between conditions), and tested conditions on different days in within-subjects designs to avoid carryover effects (e.g., [3]–[7]), which requires participants to visit the lab multiple times. Others have opted for shorter washout periods (typically less than 20 minutes) to complete all conditions in a single visit (e.g., [8]–[13]).

Currently, published guidelines for cybersickness research are limited, and best practices are not well understood, resulting in wide variation in research practices. In particular, the impact of carryover effects in within-subjects designs has not yet been systematically studied. However, carryover effects can still influence experimental outcomes even after counterbalancing, for example, by reducing statistical power or introducing bias due to nonlinear effects. In this paper, we provide quantitative evidence, based on a meta-analysis of the Simulator Sickness Questionnaire (SSQ) scores reported in 55 cybersickness experiments, to reveal the impact of these design choices on study outcomes and enable researchers to make informed decisions that balance scientific rigor with resource constraints.

The Simulator Sickness Questionnaire [14] is a widely used tool to assess various simulator sickness symptoms. It includes 16 items spread across three subscales: Nausea (N), Oculomotor Discomfort (O), and Disorientation (D). The Total Severity (TS) score is calculated as a weighted sum of the subscales. Although single-value scales for sickness, such as the Fast Motion Sickness (FMS) score [15], offer

*Tongyu Nie, Ville Cantory, and Courtney Hutton Pospick contributed equally to this work. Tongyu Nie, Ville Cantory, Courtney Hutton Pospick, Danhua Zhang, Haoyu Tan, Sam Adeniyi, and Evan Suma Rosenberg are with the University of Minnesota. Emails: {nie00035, canto063, hutto070, zhan5954, tan00213, adeni026, suma}@umn.edu

Fei Wu was with the University of Minnesota. Email: wuxx1624@alumni.umn.edu

Daniel Zielasko is with the Technical University of Denmark. Email: danzi@dtu.dk

Isayas Berhe Adhanom is with the Texas State University. Email: isayas@txstate.edu

Manuscript received August 22, 2025. Revised May 30, 2026.

0000–0000/00$00.00 © 2021 IEEE

rapid feedback to gauge a participant's overall well-being, the SSQ is more comprehensive and is the most widely used tool in cybersickness research to assess participants' cybersickness. When we use the SSQ in our analysis, we restrict our evaluation to the Total Severity score rather than breaking it down by subscale, as was originally intended by the SSQ's authors [14].

A. Contribution

In this work, we focus on quantifying the impact of carryover effects on the statistical power of analyses based on SSQ scores, as it directly influences the likelihood of detecting a true effect. We first model how carryover effects influence the standard deviation and effect size. Then, through a meta-analysis, we estimate the effect sizes associated with different experimental designs: between-subjects design, within-subjects design with a long washout period (conditions tested on different days) , and within-subjects design with a short washout period (multiple conditions on the same day).

Specifically, we proposed the following hypotheses. **H1** is formulated based on the existing literature on adaptation to cybersickness [16]–[19], while **H2** is based on the carryover effect. Both effects could introduce heterogeneity to the results and decrease the overall effect size. Finally, we are interested in **H3** because it was previously reported by Peck et al. [20].

- **H1:** The effect size of the increase in cybersickness following VR exposure is larger in studies using between-subjects designs than in studies using within-subjects designs with a ***long*** washout period.
- **H2:** The effect size of the increase in cybersickness following VR exposure is larger in studies using between-subjects designs than in studies using within-subjects designs with a ***short*** washout period.
- **H3:** The effect size of the increase in cybersickness following VR exposure decreases as the proportion of female participants increases in a given study.

We found that short within-subjects designs exhibit a positive carryover effect, where prior exposure increases symptoms in subsequent conditions, while long within-subjects designs exhibit a negative carryover effect, where symptoms decrease. Based on the magnitude of carryover effects observed in the meta-analysis, we conducted a power analysis to estimate differences in statistical power across experimental designs. Our analysis shows that using a short within-subjects design results in lower statistical power compared to long within-subjects and between-subjects designs, assuming the same number of participants per condition. To our knowledge, this is the first quantitative evidence demonstrating that short washout periods compromise the power of within-subjects cybersickness studies. Our findings provide researchers with clearer expectations for study outcomes and help inform design choices in future cybersickness research.

II. Issues with Cybersickness Research

Zielasko et al. used hypothetical scenarios to discuss the potential impact of first- or higher-degree carryover effects on cybersickness studies in a position paper [21]. The authors argue that standard counterbalancing is often insufficient when dealing with the lingering and unpredictable carryover effects inherent to cybersickness research. Case studies demonstrate how these effects can lead to erroneous conclusions, particularly when symptoms like headaches are delayed or recovery patterns vary significantly between individuals. To overcome this issue, the paper recommends adopting more rigorous protocols, such as using between-subjects designs, increasing sample sizes, or allowing a 24–48 hour washout period between conditions. The potential outcomes depicted in this paper motivated this work. However, they neither provided data to support their scenarios nor offered quantitative guidelines.

A. Previous Reviews on Cybersickness

Over the past few years, several literature reviews have addressed cybersickness. Biswas et al. reviewed 223 papers on the causes, measurement, and mitigation of cybersickness and proposed a mitigation guideline [22]. A systematic review and meta-analysis of 96 papers examined cybersickness in walking-based locomotion and found overall low SSQ scores for isometric walking, but higher scores for redirected walking [23]. In 2023, a position paper analyzed 61 papers on VR locomotion research from IEEE VR, ISMAR, and VRST, identifying a potential replication crisis in the field [24]. Another review examined current research on cybersickness reduction based on 88 papers [9]. Tian et al. presented a review of contributing factors, measurement techniques, user profiles, and mitigation strategies [25]. In cybersickness research, Hirzle et al. identified the SSQ as the most widely used measure of general discomfort in VR and critically examined its role as the predominant tool for assessing cybersickness [26]. A meta-analysis of 22 human-subject studies published at IEEE VR found that the proportion of female participants significantly influences SSQ outcomes [20]. Rebenitsch and Owen reviewed research methods, theoretical frameworks, and established findings associated with cybersickness [27]. A survey by Duzmanska et al. investigated the temporal trajectory of symptom development, adaptation, and carryover, finding that symptoms can persist for up to four hours [2]. However, this work did not provide quantitative data on statistical power.

Several other meta-analyses have highlighted key factors contributing to cybersickness in VR. A review of 26 studies found that age, cognitive and physical abilities, and task complexity influenced symptom severity, with video content and controller-based locomotion exacerbating symptoms compared to exergames and gaze-directed methods [28]. Another review reported racial differences in cybersickness, with Black participants experiencing fewer symptoms than White participants; however, their study was limited by narrow demographics and participant exclusions [29]. Saredakis et al. found that gaming content and controller-based locomotion led to higher SSQ scores, while scenic content and older users showed reduced symptoms [30]. Finally, a meta-analysis identified motion sickness susceptibility as the strongest predictor of VR sickness, with moderate effects from sex and neurological

disorders, and weaker influences from real-world and technological experience [31].

Gaps in Literature: None of the previous surveys and meta-analyses assessed the effect of experiment design on the potential cybersickness outcome. In this paper, we focused on the impact of study design on the statistical power of cybersickness-related human-subject studies. We also expect the effect size of pre- and post-SSQ scores from papers included in our meta-analysis to be higher than Peck et al.'s [20] and van Gemert et al.'s [23] because we focused on studies that are designed to induce cybersickness with virtual locomotion.

B. Cybersickness Adaptation Effects

Adaptation is the most significant order effect observed in cybersickness studies, characterized by a reduction in cybersickness symptoms in trials that occur in later sessions compared to earlier sessions [32]. Almost all published studies have used counterbalancing to address this issue. However, even though this can minimize the effect on the mean difference, counterbalancing may increase the dispersion of cybersickness symptom distribution and reduce statistical power.

a) Studies Observing Adaptation: Multiple highly cited early works on cybersickness mitigation that used within-subjects designs had their results significantly affected by adaptation. Fernandes and Feiner's work on dynamic field-of-view restrictors is one of the most influential in cybersickness mitigation. However, they found that participants who used the FOV restrictor first were less affected by the no-restrictor condition [3]. Similarly, Cao et al. adopted a similar experimental design to evaluate the mitigation effect of rest frames [4]. They also found a significant interaction effect between order and condition, which the authors suggest occurred because users adapted to the VE after initially experiencing it with a rest frame. Liu et al. also attributed some of their non-significant results to order effects after analyzing their data at the order-group level [33]. All three papers mentioned above reported significant results using between-subjects data from different order groups. Another study on rotation gain in redirected walking observed more sickness in first-time users of CAVE systems [34].

b) Studies Investigating Adaptation: Other researchers have designed studies that specifically focused on the adaptation effect. Early studies found that the number of VR exposures is a more important factor than the time interval between them [35]. Adhanom et al. used repeated exposures with gradually increasing optical flow in an abstract virtual environment to allow participants to adapt, and found that the adaptation transferred when users were exposed to a realistic virtual environment [16]. Doty et al. exposed participants to FOV restriction conditions to induce adaptation and found that the effect successfully transferred to a no-FOV-restriction condition [17]. Another study also found the adaptation effect generalizes across different VR experiences [19]. Researchers have also investigated users' behavioral, cognitive, and physiological adaptation to cybersickness [36].

These findings suggest that adaptation is an important factor affecting the results of within-subjects cybersickness studies. We hypothesize that even without carryover effects, within-subject designs with long washout periods can still observe reduced effect sizes due to order effects.

C. Cybersickness Carryover Effects

Most researchers acknowledge the presence of carryover effects in experimental studies. However, their approaches to addressing this issue vary. To minimize these effects, many opt for between-subjects designs or schedule different conditions on separate days. In contrast, some researchers consider it acceptable to test all conditions within a single day, despite the potential for carryover.

The duration of recovery and the methods used to determine whether participants are fully recovered vary across studies. Ang and Quarles allowed participants to take a 1-minute break after each trial, which they extended by 1 minute at a time until participants reported a Fast Motion Sickness (FMS) scale of 1 (on a scale of 10 instead of 20 [15]) [10]. Other researchers chose to allow participants to rest for 5 minutes (e.g., [8], [37], [38]) or 10 minutes (e.g., [6], [11], [13]).

Some researchers set specific criteria that participants had to meet before proceeding to the next condition after resting. For example, some studies used SSQ score thresholds: one allowed participants to continue after their SSQ scores dropped below 10 [8], while another required a score of 0 before continuing [12]. Others allowed participants to proceed based on subjective reports of feeling fine [39] or based on heart rate data [40]. In some cases, breaks were optional, offered "in case of fatigue" [41]. To the best of our knowledge, only one paper conducted two similar studies using both short (Experiment 1) and long washout periods (Experiment 2) [6].

Short washout periods have been called into question by prior literature [21] and publicly discussed at forums such as the IEEE VR Workshop on Immersive Sickness Prevention [42]. However, to the best of our knowledge, there is currently no quantitative evidence demonstrating the impact of this design choice.

III. Mathematical Modeling and Estimation of Carryover Effects

In this section, we mathematically model how carryover effects affect cybersickness experiment results regarding the mean difference across conditions, the effect sizes at the condition level, and the methodology for using observed condition-level results to estimate the proportion of effects carried over from previous exposure. We are interested in estimating the proportion of the carryover effect for two reasons. First, we would like to show that, based on results from accumulated literature, there is evidence showing that within-subjects designs cannot fully eliminate carryover effects. Second, and more importantly, we aim to provide a foundation for subsequent sections where we conduct a power analysis to quantify how carryover effects compromise experiment outcomes. We consider a within-subjects study with two conditions in which we ensured homogeneity in all variables, including participants, except the stimulus between conditions. We also assume the experiment design counterbalanced the order of conditions,

where, in the first session, half of the participants receive condition 1 while the other half receive condition 2.

A. Modeling the Impact on Mean Difference

In this section, we review how a positive carryover effect decreases the mean difference while a negative one increases it. The results of this section are important for understanding why a positive carryover effect reduces statistical power, which is discussed later in Section VI-B3. Consider a basic within-subjects design with two conditions, control (C) and experimental (E). We denote the average true effect of the two conditions as M_C and M_E. The true mean difference is then $M_{C-E} = M_C - M_E$. We define k as the proportion of the first condition's effect that carries over into the second condition. Because the conditions are counterbalanced, the observed means of conditions C and E are shifted because half of the participants experience condition C before E, and vice versa. The observed means, M'_C and M'_E, are calculated as:

$$\begin{aligned} M'_C &= M_C + \frac{k}{2}M_E \\ M'_E &= M_E + \frac{k}{2}M_C \end{aligned} \tag{1}$$

As Equation 2 shows, when k is positive, the observed mean difference M'_{C-E} is smaller than the true mean difference M_{C-E}. Conversely, when $k < 0$, the observed mean difference is inflated.

$$\begin{aligned} M'_{C-E} &= M'_C - M'_E \\ &= (M_C + \frac{k}{2}M_E) - (M_E + \frac{k}{2}M_C) \\ &= (1 - \frac{k}{2})M_{C-E} \end{aligned} \tag{2}$$

B. Modeling the Impact on Standard Deviation

In this subsection, we model how carryover effects increase the standard deviation (SD) of the result distribution for one of the two conditions in the experiment.

By definition, with n data points, the standard deviation is:

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2} \tag{3}$$

We consider the result of a continuous measure of our condition following a normal distribution (M, SD). Due to counterbalancing, researchers must divide participants into two order groups: Group A experiences our condition first, while Group B experiences it second. Carryover effects add an increment proportional to the other condition's effect to our condition in Group B. To model the standard deviation of our condition's results under the influence of carryover, we follow these steps:

- Calculate the SD of Group A.
- Calculate the SD of Group B to account for carryover effects.
- Combine the SDs of Groups A and B.

1) Group A's SD: Since Group A contains half of the total participants and is not affected by carryover effects, its SD is:

$$\begin{aligned} SD^2_{half} &= \frac{1}{\frac{n}{2}-1}\sum_{i=1}^{n/2}(x_i - M) \\ &= \frac{2}{n-2}\sum_{i=1}^{n/2}(x_i - M) \\ &= \frac{2}{n-2}\frac{(n-1)SD^2}{2} \\ &= \frac{n-1}{n-2}SD^2 \end{aligned} \tag{4}$$

2) Group B's SD: Because Group B's SD is affected by effects carried over from the other condition (the mean of which we define as M_e), its SD is derived from the sum of two variables: the true value of our condition without carryover ($SD = SD_{half}$) and the carryover effect itself ($SD = k \cdot SD_{half}$). We use k to represent the proportion of the first condition's effect that carries over into the second condition. Due to the nature of carryover, especially when the washout period is short, the carryover effect is expected to be proportional to and strongly correlated with the effect of the previous condition.

Equation 5 shows the standard deviation of the sum of two variables (X and Y). The variable ρ represents the correlation coefficient between the first session's effect and the carryover effect. Because the carryover effect is assumed to be strongly correlated with the previous exposure, we set $\rho = 1$. Despite being ideal, this choice is conservative as it does not further introduce variations from the carryover effect.

$$SD_{X+Y} = \sqrt{SD_X^2 + SD_Y^2 + 2\rho SD_X SD_Y} \tag{5}$$

Since the observed results for Group B represent the sum of the true effect and the carryover effect, the SD for Group B is:

$$\begin{aligned} SD_{carry\_half} &= \sqrt{SD^2_{half} + k^2 SD^2_{half} + 2\rho k SD^2_{half}} \\ &= \sqrt{k^2 + 2\rho k + 1} \cdot SD_{half} \end{aligned} \tag{6}$$

3) Combining the SDs of Groups A and B: We can merge the mean and standard deviation from two groups with n_1 and n_2 data points together using equations 7 and 8 and get the combined descriptive statistics M_c and SD_c.

$$M_c = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2} \tag{7}$$

and

$$SD_c = \sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-1} + \frac{n_1 n_2 (M_1 - M_2)^2}{(n_1+n_2)(n_1+n_2-1)}} \tag{8}$$

When we merge the data from the first and second session together ($n_1 = n_2 = n/2$), the mean and SD of the combined final results are M_c and SD_c, which explains why within-subjects designs can increase the standard deviation of the experiment outcome.

$$M_c = M + \frac{1}{2}M_e \tag{9}$$

$$SD_c = \sqrt{\frac{(\frac{n}{2}-1)(SD^2_{half}+SD^2_{carry\_half})}{n-1} + \frac{\frac{n^2}{4}M_e^2}{n(n-1)}} = \sqrt{\frac{(n-2)(1+k^2+2\rho k+1)SD^2_{half}}{2(n-1)} + \frac{nM_e^2}{4(n-1)}} = \sqrt{\frac{(k^2+2\rho k+2)SD^2}{2} + \frac{nM_e^2}{4(n-1)}} \quad (10)$$

C. Estimating the Proportion of the Carryover Effect

Based on equation 10, we can further derive the estimation of k because we are interested in knowing how much cybersickness is carried-over to the second session. Note, when n is relatively large, $\sqrt{\frac{n-1}{n}} \approx 1$.

$$M_e = \pm\sqrt{\frac{4(n-1)}{n}\left[SD_c^2 - (1+\rho k+\frac{1}{2}k^2)SD^2\right]} \approx \pm 2\sqrt{SD_c^2 - \left(1+\rho k+\frac{1}{2}k^2\right)SD^2} \quad (11)$$

However, a paper only reports the combined result SD_c without the true SD that is unaffected by carryover, and the standard deviations reported by different studies are not directly comparable. As a result, we have to find another representation of k to estimate it.

On the other hand, effect sizes are comparable across studies, even with different sample sizes or designs, and are widely used in meta-analyses. Cohen's d is a type of effect size defined by equation 12, where M_{diff} and SD_{diff} can be considered as the delta-SSQ score distributions and r is a constant that we will talk more about later in Section IV-E. Because r will eventually cancel out in the calculation of power, we use $r = 0.5$ here to simplify calculation.

$$d = \frac{M_{diff}}{SD_{diff}}\sqrt{2(1-r)} \quad (12)$$

Given the definition of effect size in equation 12 with $r = 0.5$, we can have $SD = M/d$. Then the representation of M_e in equation 11 can be turned into:

$$M_e = \pm 2\sqrt{\frac{M_c^2}{d_c^2} - \left(1+\rho k+\frac{1}{2}k^2\right)\frac{M^2}{d^2}} = \pm 2\sqrt{\frac{\left(M+\frac{1}{2}M_e\right)^2}{d_c^2} - \left(1+\rho k+\frac{1}{2}k^2\right)\frac{M^2}{d^2}} \quad (13)$$

Divide both sides by M_e, replace M_e/M with k (see definition of k in Section III-A), and square both sides, we have Equation 14. Here, d_c is the estimated effect size with carryover effect (within-subjects studies) and d is the estimated effect size without carryover effects (between-subjects studies), which we will get from the meta-regression. We used $c_1 = 1/d_c^2$ and $c_2 = 1/d^2$ to simplify the notation.

$$1 = 4\left[\frac{\left(\frac{1}{k}+\frac{1}{2}\right)^2}{d_c^2} - \left(1+\rho k+\frac{1}{2}k^2\right)\frac{1}{k^2d^2}\right] = 4\left[c_1\left(\frac{1}{k^2}+\frac{1}{k}+\frac{1}{4}\right) - c_2\left(\frac{1}{k^2}+\frac{\rho}{k}+\frac{1}{2}\right)\right] = 4\left[(c_1-c_2)\frac{1}{k^2} + (c_1-\rho c_2)\frac{1}{k} + \left(\frac{c_1}{4}-\frac{c_2}{2}\right)\right] \quad (14)$$

We can see that equation 14 is a quadratic formula, which can be written as:

$$(c_1-c_2)\frac{1}{k^2} + (c_1-\rho c_2)\frac{1}{k} + \left(\frac{c_1}{4}-\frac{c_2}{2}-\frac{1}{4}\right) = 0 \quad (15)$$

Once we multiply both sides with $-k^2$, the equation becomes:

$$\frac{1}{4}(1-c_1+2c_2)k^2 + (\rho c_2 - c_1)k + (c_2-c_1) = 0 \quad (16)$$

Solving this quadratic equation yields two mathematical roots. As is standard when modeling physiological phenomena, we must identify the physically meaningful solution by evaluating these roots against established boundary conditions for cybersickness progression.

$$k = \frac{-(\rho c_2 - c_1) \pm \sqrt{(\rho c_2 - c_1)^2 - 4\cdot\frac{1}{4}(1-c_1+2c_2)(c_2-c_1)}}{2\cdot\frac{1}{4}(1-c_1+2c_2)} = \frac{(c_1-\rho c_2) \pm \sqrt{(\rho^2-2)c_2^2 + (3-2\rho)c_1c_2 + c_1 - c_2}}{\frac{1}{2}(1-c_1+2c_2)} \quad (17)$$

The solution in Equation 17 shows the proportion of carryover effect compared to the magnitude of M, the cybersickness level induced in the current condition. Theoretically, M_e should be correlated to the average induced cybersickness induced in the other condition, M_{other}; however, M and M_{other} tend to be approximately similar magnitude, and much higher than the carryover effect. Further, in the estimation of d_c, the difference between M and M_{other} is averaged out as we include the data from both conditions in this meta-analysis. Recall that $c_1 = 1/d_c^2$ and $c_2 = 1/d^2$, where d_c and d are the estimated effect sizes of within-subjects and between-subjects studies, respectively. Based on the literature search described in Section IV, we will estimate d_c (separately for long and short washout periods) and d in Section V to calculate the values of k for both washout durations.

IV. Method

A. Preliminaries

To investigate **H1**, **H2**, and **H3** and estimate the magnitude of the carryover effect, we conducted a meta-analysis and a meta-regression to evaluate the impact of washout periods on the effect size of SSQ outcomes. Additionally, we included the proportion of female participants as a moderator, as prior research suggests that a higher proportion of female participants is associated with smaller SSQ effect sizes [20]. We followed the PRISMA guidelines [43] when conducting the literature survey.

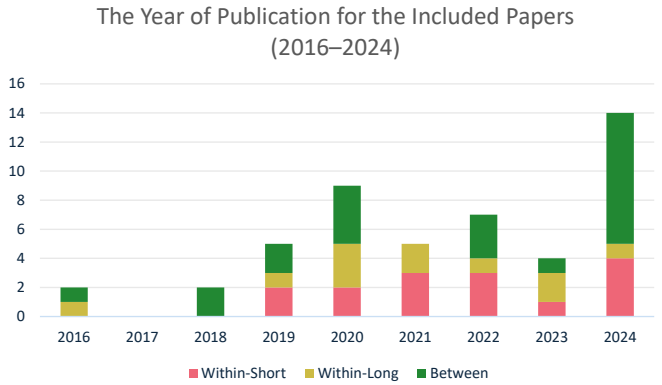

Fig. 1. The number of included papers published in each year, categorized into three categories: within-subjects with short washout periods, within-subjects with long washout periods, and between-subjects designs. Long washout periods were defined as testing different conditions on separate days, whereas short washout periods were defined as testing multiple conditions within the same day.

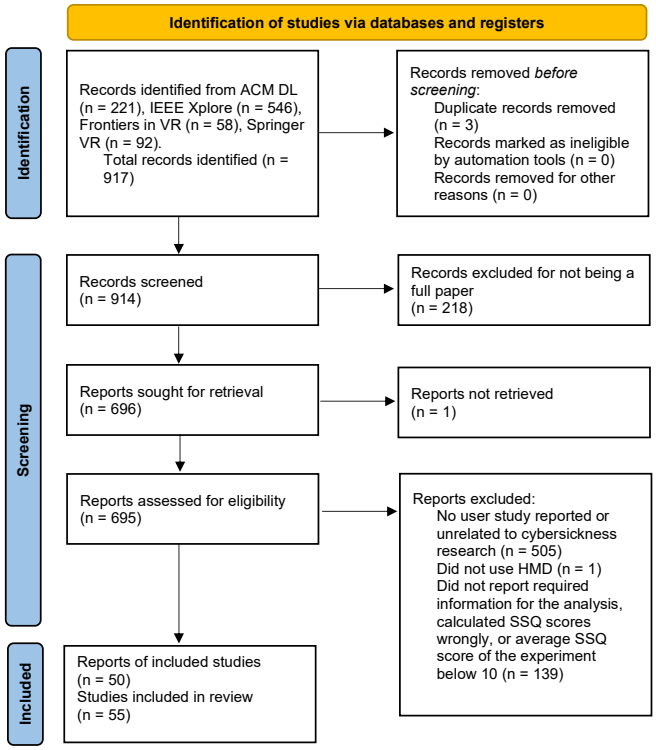

Fig. 2. The PRISMA flow diagram illustrates the process of identifying, screening, and selecting studies eligible for inclusion in our meta-analysis.

B. Literature Search

a) Year Range: The final literature search was conducted on December 4, 2024. We included studies published between 2016 and 2024, as 2016 marks the beginning of the consumer VR era. Notably, a paper by Fernandes and Feiner on cybersickness was published that year at IEEE 3DUI [3]. This work has been highly influential, widely cited not only for its findings on FOV restrictors and cybersickness mitigation, but also for its experimental design and the introduction of a novel symptom measurement method known as the relative discomfort score.

b) Keywords: We used the search terms: ("VR" OR "Virtual Reality") AND "*sickness," following the keyword strategy adopted by Biswas et al. in their recent review of cybersickness literature [22]. This approach reduces the risk of omitting relevant studies, as many papers refer to cybersickness as "visually induced motion sickness," "VR sickness," or "simulator sickness." However, this broad search also yields results that are not directly relevant. We manually screened all papers matching these keywords using the inclusion criteria listed in Section IV-C, specifically criterion 4, to ensure the exclusion of unrelated papers. We directly searched IEEE Xplore[1] and ACM Digital Library[2], as they publish many of the leading journals (e.g., IEEE TVCG, ACM TOG, ACM TAP) and conferences (e.g., IEEE VR, ISMAR; ACM CHI, VRST, SIGGRAPH, SAP) in the VR field. In line with the recommendations by Field and Gillett [44], we also manually searched additional journals outside IEEE and ACM. Specifically, we reviewed Frontiers in Virtual Reality[3] and Springer Virtual Reality[4], as both journals have a strong focus on VR and frequently publish studies on cybersickness.

C. Inclusion Criteria

We included reports from studies based on the following criteria:

1) Must be a full paper. No posters, extended abstracts, workshop papers, or short papers that are 3 pages or less.
2) Presented results of a human-subject study.
3) Used the head-mounted display (HMD) in their study.
4) Focus on the cause, mitigation, or measurement of cybersickness induced by visual stimulus.
5) Correctly calculated SSQ total scores (see [45] for details).
6) Reported required information for the analysis, including the mean and standard deviation of the delta- or post-SSQ for each condition and the sex distribution of participants.
7) The average SSQ score of the experiment is above 10.

Although there is no single number that can serve to characterize the overall incidence of sickness using the SSQ, Stanney et al. categorized symptoms induced by simulated experiences with average SSQ scores below 10 as minimal or negligible symptoms based on 9,000 samples collected from military flight trainers on different simulators [46]. As a result, we excluded studies reporting an average SSQ score below 10, as a typical cybersickness study should induce more than minimal symptoms to allow for meaningful comparisons and avoid floor effects. If a paper studied physically induced motion sickness, we excluded it since we focus our analysis on visually induced cybersickness. Among all the samples we collected, we excluded conditions (4 in total) that did not use HMDs and kept the data from conditions that used HMDs.

[1]https://ieeexplore.ieee.org
[2]https://dl.acm.org/
[3]https://www.frontiersin.org/journals/virtual-reality
[4]https://link.springer.com/journal/10055

Figure 2 shows the detailed selection process for the identified papers. We utilized a two-round selection process during screening. In the first pass, we identified and excluded all entries that were not full-length papers. Four co-authors (TN, VC, CHP, IBA) individually screened all search results to flag posters, extended abstracts, workshop papers, or short papers based on the first inclusion criterion. Each reviewer's results were double-checked by another reviewer. Next, six co-authors (TN, VC, CHP, DZh, HT, IBA) read the full text of the remaining papers and selected those that met the remaining inclusion criteria. At this stage, we excluded one paper that was indexed but not published by ACM DL, as its full-text PDF file could not be retrieved online. Again, each reviewer's results were cross-checked by another reviewer.

We include 50 papers that meet our inclusion criteria, which include 11 papers from Springer Virtual Reality, 9 papers from TVCG (including special issues), 6 papers from CHI, 6 papers from Frontiers in VR, 3 papers from IEEE VR, 1 paper each from 3DUI, ISMAR, SUI, TAP, and ToG, and 8 papers from other venues. A total of 55 independent human-subject studies were reported in these papers that met our inclusion criteria.

D. Categorization of Experiments

We categorized the experiments into three groups: between-subjects, within-subjects (long), and within-subjects (short), and listed them in Table I.

1) Between-Subjects: The between-subjects category includes studies that explicitly used a between-subjects design. It also includes studies with only one experimental condition, like those focused on building datasets for cybersickness measurement, since each participant experienced only a single VR session.

2) Within-Subjects: The remaining studies employed a within-subjects design and were further divided based on the length of their washout period.

a) Short Washout Periods: We define "*short*" washout periods as those that allow the entire experimental session for a single participant to be completed within a single lab visit. Theoretically, such a washout period could vary from zero to several hours. However, to the best of our knowledge, experiments rarely report using washout periods longer than 20 minutes for single-visit sessions. As shown in Table I, in all included experiments that fell into this category and reported their washout periods, the durations were less than or equal to 15 minutes. Even though there is some uncertainty regarding the exact length of the washout periods, as a few experiments extended resting time if participants did not meet certain criteria (e.g., reporting an FMS score of 1 or an SSQ score of 0) or did not report the exact duration, the general trend suggests it is very rare for researchers to design experiments with washout periods that average significantly over 15 minutes. Nevertheless, there are large variations in the length of washout periods across the included experiments. The washout periods of included experiments range from completely optional to 1, 5, or 10–15 minutes, which may have varying impacts on experiment outcomes. We discuss these variations further in Section VIII.

b) Long Washout Periods: In contrast, we define "*long*" washout periods as those that require participants to experience conditions on different days, allowing them at least one night to recover from previous cybersickness symptoms. It is worth noting that six out of 13 experiments reported that different conditions were tested on different days; this implies that, despite being significantly longer than the *short* washout periods mentioned above, some washout periods may have been less than 24 hours. Four experiments reported that conditions were separated by at least 24 hours, and three other experiments required participants to revisit the lab after 5–10 days. However, long washout periods do not necessarily eliminate carryover effects, as participants may adapt to VR over time [16], [17].

Although many of the reviewed studies employed mixed designs, often treating cybersickness as a within-subjects factor and other variables (e.g., sex) as between-subjects factors, we focus exclusively on the design relevant to the cybersickness factor in this analysis.

E. Effect Size Calculation

a) Standard Approach: For our meta-analysis, we chose to investigate the SSQ outcome–the most commonly used cybersickness measure. We focused on the effect size (Cohen's d) of SSQ increase in each condition of every experiment. Cohen's d is calculated using the distribution from pre- and post-exposure using equation 18. Cohen's d serves as the standardized mean change, which is the foundational and validated metric for quantifying within-group temporal shifts in meta-analyses [87]. By calculating this metric for every condition, we were able to standardize the impact of experiment design across studies, allowing us to accurately estimate how different washout periods influence carryover.

$$d = \frac{M_{post} - M_{pre}}{\frac{SD_{diff}}{\sqrt{2(1-r)}}} \tag{18}$$

where

$$SD_{diff} = \sqrt{SD_{pre}^2 + SD_{post}^2 - 2rSD_{pre}SD_{post}} \tag{19}$$

SD_{diff} is the standardized mean difference (SMD). The factor r is the pre-post test correlation value (Pearson correlation coefficient).

b) Limitations: The correlation factor r is difficult to determine directly. Ideally, with sufficient data, it can be estimated using Equation 20:

$$r = \frac{\sum(Pre_i - M_{pre})(Post_i - M_{post})}{\sqrt{\sum(Pre_i - M_{pre})^2 \sum(Post_i - M_{post})^2}} \tag{20}$$

However, this value is rarely reported in the existing literature. Peck et al. [20] claimed to estimate r as 0.90, citing the test-retest reliability of the SSQ based on a paper by Golding [88]. In reality, Golding reported the test-retest reliability of the Motion Sickness Susceptibility Questionnaire (MSSQ) — not the Simulator Sickness Questionnaire (SSQ). Importantly, the MSSQ does not measure cybersickness symptoms; rather, as Golding described, it is designed to "predict individual differences in motion sickness caused by a variety of stimuli."

TABLE I
THIS TABLE PRESENTS KEY INFORMATION FOR EACH INDEPENDENT EXPERIMENT INCLUDED IN OUR META-ANALYSIS, INCLUDING THE EXPERIMENTAL DESIGN, THE NUMBER OF CELLS, THE DURATION OF VR EXPOSURE, WASHOUT PERIODS, AGGREGATED EFFECT SIZES, VARIANCES, SAMPLE SIZES, THE PROPORTION OF FEMALE PARTICIPANTS, AND THE TYPE OF SSQ DATA USED IN THE ANALYSIS. NOTE THAT THE ONLY ELIGIBLE DATA POINT REPORTED IN [47] WAS IDENTIFIED AS AN OUTLIER AND EXCLUDED FROM THE FINAL ANALYSIS.

Experiment	Design	Number of Cells	Duration *(* may be shorter due to cybersickness)*	Washout Period	Effect Size	Variance	Sample Size	Proportion of Female Participants	SSQ *(Delta- or post-SSQ*
[48]	Between	2	Unspecified	N/A	0.964	0.03	47	43%	Post
[49]	Between	1	Unspecified	N/A	1.065	0.06	26	54%	Post
[50]	Between	3	130 seconds	N/A	1.088	0.031	54	57%	Post
[51]	Between	2	10 min	N/A	1.312	0.047	40	63%	Post
[52]	Between	1	30 min	N/A	1.042	0.057	27	41%	Post
[8]	Between	2	45 min	N/A	1.727	0.066	40	50%	Delta
[53]	Between	8	9 min	N/A	0.826	0.006	240	58%	Delta
[54] Exp 1	Between	2	10 min	N/A	1.525	0.08	30	50%	Post
[54] Exp 2	Between	2	10 min	N/A	1.259	0.08	24	50%	Post
[55]	Between	1	15 min	N/A	2.125	0.142	23	9%	Post
[56]	Between	1	14 min*	N/A	1.402	0.066	30	50%	Delta
[57]	Between	3	15 min	N/A	1.003	0.01	153	35%	Delta
[58]	Between	1	7.5 hours	N/A	1.314	0.085	22	55%	Post
[16]	Between	1	20 min	N/A	1.387	0.103	19	42%	Delta
[59]	Between	2	25 min	N/A	1.486	0.14	16	25%	Post
[60]	Between	4	30 min*	N/A	1.478	0.027	81	53%	Delta
[61]	Between	1	18 min	N/A	1.871	0.11	25	48%	Post
[62]	Between	1	40 min	N/A	1.566	0.07	32	44%	Post
[63]	Between	4	40 min*	N/A	2.728	0.114	44	39%	Delta
[64]	Between	2	27 min	N/A	0.957	0.011	129	50%	Post
[65]	Between	1	14 min*	N/A	1.166	0.042	40	53%	Post
[66]	Between	1	56 min	N/A	1.419	0.067	30	50%	Post
[67]	Between	1	15 min	N/A	1.241	0.049	36	44%	Post
[17]	Between	2	20 min*	N/A	1.323	0.018	103	43%	Post
[68]	Between	4	20 min*	N/A	1.297	0.009	201	41%	Post
[69]	Between	2	10-25 min	N/A	0.856	0.033	41	20%	Delta
[3]	Within (Long)	2	21 min	Different days	1.712	0.053	24	33%	Delta
[70] Exp 1	Within (Long)	2	50 min	5-10 days	1.142	0.055	30	50%	Post
[70] Exp 2	Within (Long)	2	35 min	5-10 days	1.233	0.059	30	27%	Post
[71]	Within (Long)	2	Unspecified	>24 hours	0.892	0.032	22	50%	Post
[72]	Within (Long)	4	10 min	>24 hours	1.551	0.019	30	0%	Post
[73]	Within (Long)	3	9 min	About 24 hours	0.707	0.03	14	50%	Post
[74]	Within (Long)	2	Unspecified	>24 hours	0.924	0.025	28	50%	Post
[75]	Within (Long)	3	20 min	Different days	1.467	0.035	20	50%	Post
[47]	Within (Long)	2	15 min	Different days	2.94	0.111	48	50%	Delta
[76]	Within (Long)	5	15 min	Different days	0.779	0.015	19	42%	Delta
[77]	Within (Long)	2	Unspecified	About a week	0.86	0.022	31	48%	Post
[78] Exp 1	Within (Long)	3	20 min	Different days	0.585	0.02	20	50%	Delta
[78] Exp 2	Within (Long)	3	20 min	Different days	0.989	0.025	20	50%	Delta
[79]	Within (Short)	2	5 min*	>1 min	0.707	0.052	12	33%	Post
[13]	Within (Short)	2	25 min	10 min	0.578	0.022	28	50%	Delta
[41]	Within (Short)	7	<3 min	(Optional)	0.886	0.007	28	36%	Post
[80]	Within (Short)	4	3 min	5 min	0.531	0.01	30	23%	Delta
[37]	Within (Short)	2	<10 min	5 min	1.043	0.035	22	45%	Post
[38]	Within (Short)	3	146 seconds	>5 min	0.862	0.019	24	63%	Post
[40]	Within (Short)	2	7-8 min	(Until rested and heart rate at resting level)	0.803	0.041	16	38%	Post
[10]	Within (Short)	3	5 min	>1 min and until FMS=1	0.964	0.013	38	45%	Delta
[81]	Within (Short)	3	3-5 min	Unspecified	1.209	0.024	24	29%	Post
[82]	Within (Short)	2	15 min	10-15 min	1.121	0.029	28	32%	Post
[12] Exp 2A	Within (Short)	5	3 min	5 min	0.79	0.024	11	0%	Delta
[12] Exp 3	Within (Short)	3	223 seconds	(Until SSQ=0)	0.892	0.009	51	35%	Delta
[83]	Within (Short)	2	<7 min	Unspecified	1.069	0.049	16	0%	Delta
[84]	Within (Short)	2	412 seconds	15 min	1.033	0.024	32	47%	Post
[85]	Within (Short)	4	5 min	>5 min and until FMS=1)	1.009	0.007	54	37%	Delta
[86]	Within (Short)	2	35 min	Unspecified	1.232	0.028	32	53%	Post

Thus, its test-retest reliability reflects the stability of individual susceptibility to motion sickness, not the correlation between pre- and post-exposure symptom scores measured by the SSQ.

Because estimating r requires both pre- and post-SSQ

TABLE II
THIS TABLE PROVIDES FURTHER DETAILS ON THE YEAR OF PUBLICATION, INDEPENDENT VARIABLES, AND RELATED DEPENDENT VARIABLES FOR EACH INCLUDED EXPERIMENT. THE NAMES OF THE INDEPENDENT VARIABLES WERE DIRECTLY COPIED FROM THE RESULTS REPORTED IN EACH PAPER.

Experiment	Year	Independent Variables	Dependent Variables
Bessa et al. [48]	2016	360Video; 3D-360Video	SSQ
Schmitz et al. [49]	2018	N/A	SSQ
Tyrrell et al. [50]	2018	Neck pain; Vestibular; Control	SSQ; Simulator sickness visual analogue scale
Mittelstaedt et al. 2019 [51]	2019	Bike/HMD; Gamepad/HMD	SSQ
Szpak et al. [52]	2019	N/A	SSQ
Nie et al. [8]	2020	blur disabled; blur enabled	SSQ
Peng et al. [53]	2020	visual-only w/o head bobbing; visual-only with head bobbing; audio; 2-sided tapping; 2-sided vibration synchronized; 2-sided vibration random; backside vibration synchronized; backside vibration random;	SSQ; Discomfort score
Yildirim [54] Exp 1	2020	Rift; Vive	SSQ
Yildirim [54] Exp 2	2020	Rift; Vive	SSQ
Islam et al. [55]	2020	N/A	SSQ; Fast Motion Sickness Scores
Teixeira et al. 2022 [56]	2022	N/A	SSQ; Fast Motion Sickness Scores; Vection Strength
Sepich et al. [57]	2022	No-Task; 0-Back; 2-Back	SSQ
Chen and Weng [58]	2022	N/A	SSQ
Adhanom et al. 2022 [16]	2022	N/A	SSQ; Discomfort score
Cortes et al. [59]	2023	VRS; NoVRS	SSQ; 10-point scales on six symptoms
Venkatakrishnan et al. 2023 [60]	2023	ND; AD; VD; CD	SSQ; 10-point scale on comfort
Achanccaray et al. [61]	2023	N/A	SSQ
Thorp et al. [62]	2024	IVR	SSQ
Venkatakrishnan et al. 2024 [63]	2024	SIL-Low; SIL-High; STD-1-back; STD-2-back	SSQ; 10-point scale on comfort
Mittelstaedt et al. 2024 [64]	2024	Static; Dynamic	SSQ; three items asking for sickness symptoms
Teixeira et al. 2024 [65]	2024	N/A	SSQ; Vection Strength
Palmisano et al. [66]	2024	N/A	SSQ; Fast Motion Sickness Scores
Setu et al. [67]	2024	N/A	SSQ; Fast Motion Sickness Scores
Doty et al. [17]	2024	Day 1; Full FOV	SSQ; Discomfort score
Kelly et al. [68]	2024	None; FOV; Snap; Both	SSQ; Discomfort score
Nisiotis et al. [69]	2024	Normal; Time Dilation	SSQ
Fernandes and Feiner [3]	2016	No restriction; Restriction	SSQ; Discomfort score
Benz et al. [70] Exp 1	2019	HMD	SSQ
Benz et al. [70] Exp 2	2019	HMD	SSQ
Adhanom et al. 2020 [71]	2020	FX; FV	SSQ; Discomfort score
Wibirama et al. [72]	2020	FPS player; FPS spectator; racing player; racing spectator	SSQ
Ranasinghe et al. [73]	2020	No-Olf; High-Olf; Mde-Olf	SSQ; Fast Motion Sickness Scores
Adhanom et al. 2021 [74]	2021	RN; RY	SSQ
Vlahovic et al. 2021 [75]	2021	BS; OU; SS	SSQ
Park et al. [47]	2022	HMD	SSQ
Pohlmann et al. [76]	2023	C1-NR; C2-A; C3-V; C4-AV; C5-NC	SSQ; Misery Scale
Weser et al. [77]	2023	Walking-2; Manual wheeling	SSQ
Vlahovic et al. 2024 [78] Exp 1	2024	BS; OU; SS	SSQ; CSQ; VRSQ
Vlahovic et al. 2024 [78] Exp 2	2024	FN; DB; SPT	SSQ; CSQ; VRSQ
Hu et al. [79]	2019	P; O	SSQ
Al Zayer et al. [13]	2019	RN; RY	SSQ; Discomfort score
Wolf et al. [41]	2020	forward-jumping; teleportation; scaled-1.4; scaled-1.8; scaled-2.2; scaled-2.6; scaled-3.0	SSQ
Tian et al. [80]	2020	SB; SV; DB; DV	SSQ
Adhikari et al. [37]	2021	Gamepad; HJ	SSQ; Fast Motion Sickness Scores (0-100)
Nguyen-Vo et al. [38]	2021	NaviBoard; Walking; Controller	SSQ
Xu et al. [40]	2021	Young Adults VR-Cer; Young Adults VR-Unc	SSQ
Ang and Quarles 2022 [10]	2022	Flat; Regular Bump; Perlin Noise	SSQ; Fast Motion Sickness Scores (0-10)
Matviienko et al. [81]	2022	Straight road; Turns; Slopes	SSQ; Fast Motion Sickness Scores
Paroz and Potter [82] Exp 1	2022	Control; Air flow	SSQ
Lin et al. [12] Exp 2A	2023	30; 60; 90; 120; 180	SSQ
Lin et al. [12] Exp 3	2023	TP; JS; Ours	SSQ; Fast Motion Sickness Scores
Al-Ashwal et al. [83]	2024	Clear; Stormy	SSQ
Mayor et al. [84]	2024	FNW; MTRW	SSQ
Ang and Quarles 2024 [85]	2024	Control; FOV; Flat; Smooth	SSQ; Fast Motion Sickness Scores (0-10)
Gabes et al. [86]	2024	Active; Passive	SSQ

scores, Peck et al. were only able to include 22 studies out of the 407 they sampled. This is unsurprising, given that when Kennedy et al. introduced the SSQ, they stated that "the SSQ scoring system ... is intended only for applications to postexposure symptoms" [14]. Although many studies do collect pre- and post-exposure SSQ scores, it is not always the case. We estimate r directly using Equation 20 and data from prior studies available to us. Based on the data from a previous experiment [5], we estimated $r = 0.5$. While this value may vary across studies, it is important to note that r affects all

effect sizes uniformly, as shown in Equation 18, and therefore does not influence the relative differences between studies.

c) Solution: Fortunately, for delta-SSQ scores (pre-post difference), we can calculate Cohen's *d* directly using equation 12 because both M_{pre} and SD_{pre} in equation 18 are 0 in this case [87].

In our search results, we found that only a subset of studies reported delta-SSQ scores, while the rest reported only post-SSQ scores. We carefully considered whether it would be appropriate to treat post-SSQ scores as delta-SSQ scores.

When Kennedy et al. introduced the SSQ, they found minimal variance in pre-exposure SSQ scores among 1,119 healthy U.S. Air Force pilots [14]. Based on this finding, they suggested that the mean and standard deviation of pre-exposure SSQ scores could reasonably be assumed to be zero. However, recent works have questioned the validity of this assumption when generalizing beyond the pilot population. For instance, Brown et al. [89] argued that the zero-baseline assumption may not hold for the general population, leading to increased interest in reporting delta-SSQ scores. Nonetheless, many researchers have continued to follow the original recommendation, resulting in a mix of reporting practices.

This inconsistency poses a challenge for meta-analysis, as it is generally advised not to combine post-intervention scores and change scores (delta) when calculating standardized mean differences (SMD) [90]. However, a previous empirical study analyzing 21 meta-analyses in osteoarthritis research found no meaningful difference between using post-intervention values and change scores [91]. Based on such evidence, Higgins et al. recommended that both types of data may be combined in a single analysis under certain conditions [92].

Following this guidance, we included studies that reported either delta-SSQ or post-SSQ scores (see Section IV-C). When both delta-SSQ and post-SSQ were available, we used delta-SSQ. In cases where only pre- and post-SSQ scores were reported (and not delta-SSQ), we did not compute delta-SSQ ourselves using Equation 18, due to the uncertainty surrounding the correlation factor *r* discussed earlier. Instead, we used only the post-SSQ scores in such cases.

F. Data Collection

1) Unit of Analysis: The unit of analysis refers to the fundamental entity from which data is collected and analyzed. In this meta-analysis, we defined the unit of analysis at the condition level. For each condition within a study, we extracted a single effect size from the reported results.

As noted by Li et al., the reporting of crossover trials (i.e., within-subjects designs) varies greatly, and essential details for performing paired analyses are often missing in published studies [93]. As a result, researchers frequently treat crossover trials as if they were parallel-group trials (i.e., between-subjects designs), assuming independence between groups [94]. While this approach introduces a unit-of-analysis error, it is generally considered conservative, as it tends to underweight the influence of crossover trials. Thus, although not ideal, this type of error is viewed as less serious than other unit-of-analysis violations [92].

2) Data Collection and Extraction: As mentioned in Section IV-E, we extracted the means and standard deviations of the delta or post-SSQ scores for each condition across the 55 experiments included in Section IV-C for our meta-analysis. In cases where authors reported data for the same condition split into subgroups, such as by presentation order (e.g., [3]) or participant classification (e.g., sick vs. not sick in [37]), we recombined the data using Equations 7 and 8, following the method outlined by Higgins et al. [92].

One experiment only reported the Cohen's *d* between pre- and post-SSQ scores [47]. In cases involving between-subjects designs where sex-specific data was unavailable at the condition level, we assumed an even distribution of male and female participants across conditions. For example, if a between-subjects experiment reported that 60% of the participants were female but did not specify the sex distribution for each condition, we assumed that all conditions shared the same distribution of 60%. While this assumption introduces a limitation, we consider it a necessary and practical compromise to enable data synthesis.

3) Descriptive Statistics: The sample-size-weighted average SSQ scores for between-subjects, within-subjects (long), and within-subjects (short) conditions are 36.12, 32.88, and 24.26, respectively.

V. Results

The complete dataset collected in Section IV-F and the corresponding analysis code are available in our GitHub repository[5].

A. Basic Meta-Analysis

We first estimated the overall effect size for the entire population. To get a thorough presentation on outliers, we used the `influence` function in the R package `metafor` [95] to calculate and plot the influence diagnostics. Upon careful review of the externally standardized residuals, DFFITS values, Cook's distances, covariance ratios, and a few other diagnostics, all metrics revealed the same outlier. As a result, we excluded it from the analysis. After we removed this outlier, the estimated average effect size was $d = 1.06$ (95% CI $1.00 - 1.12, p < .001$), indicating a large effect size and a significant increase in the SSQ score after VR exposure from cybersickness studies.

B. Assessing Statistical Heterogeneity

The heterogeneity analyses showed significant variability in effect sizes among all experiments. The test for heterogeneity was significant ($Q(133) = 297.52, p < .001$). The I^2 statistic was $51.41\% > 50\%$, indicating that our data is heterogeneous, and a random effect model of meta-analysis should be used.

[5]https://github.com/nietongyu/CybersicknessGuideline

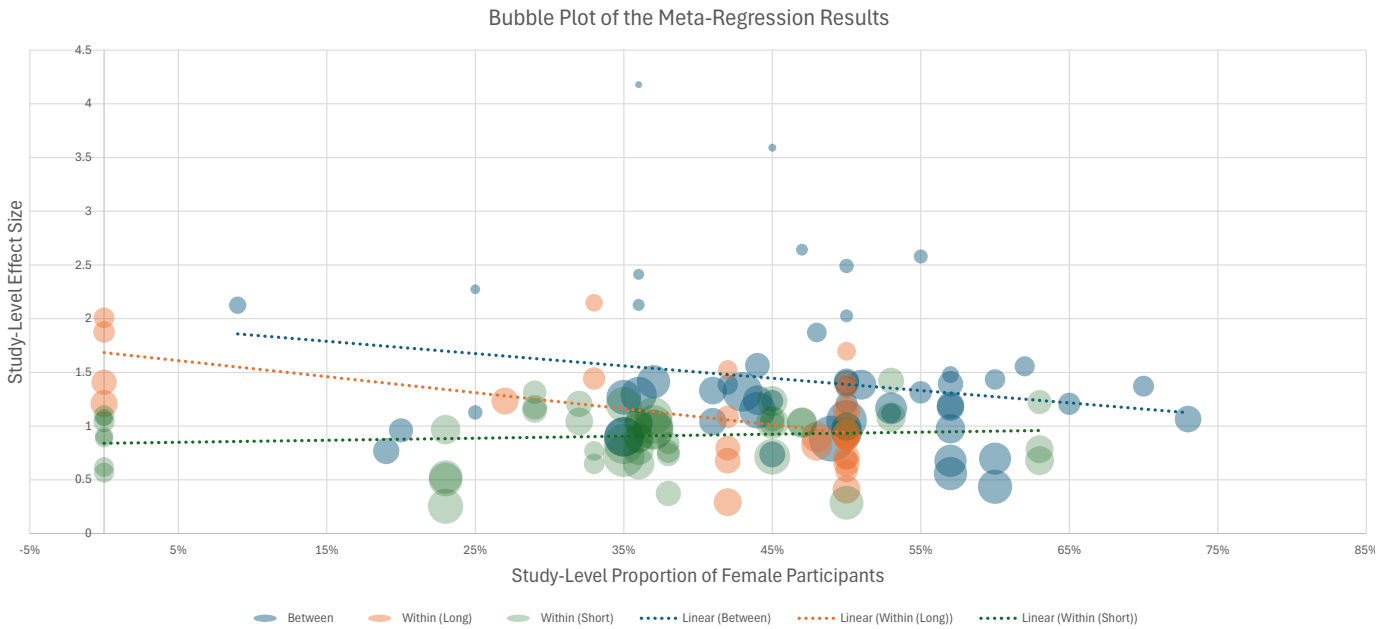

Fig. 3. In this bubble plot, we present results of the meta-regression. Each circle represents the effect size in SSQ score change for one condition in an independent study. The sizes of circles are proportional to the inverse of variances ($1/Variance$). We found a significant difference among different study designs. Between-subjects designs resulted in significantly higher effect size compared to within-subjects designs with both long and short washout periods. Within-subjects designs with long washout periods resulted in significantly higher effect size compared with within-subjects designs with short washout periods.

C. Moderator Analyses

1) Full Analysis: We conducted a meta-regression to test if three different study designs: between, within (long), and within (short), can significantly affect study's effect size. We also included the proportion of female participants as a factor because a previous meta-analysis has found it significantly impact experiments' effect size of SSQ changes [20]. The test of moderators of the meta-regression indicated that study design had a significant impact on the observed effect size ($QM(3) = 24.95, p < .001$). Model results showed that between-subjects design studies have significantly higher effect size ($1.42, 95\%$ CI $1.20 - 1.63, p < .001$) than within (long)'s ($1.22, 95\%$ CI $0.85 - 1.59, p = .01$) and within (short)'s ($1.06, 95\%$ CI $0.71 - 1.42, p < .001$) effect size. As a result, **H1** and **H2** are supported. We also observed a significant association between the proportion of female participants and effect size (slope: $-0.45, 95\%$ CI $-0.87 - -0.04, p = .03$). This result supports **H3**.

2) Analysis with Only Delta-SSQs: In addition, to enhance the generalizability of our approach and validate the practice of treating post-SSQ scores as equivalent to delta-SSQ scores, we conducted a separate analysis on the delta-SSQ data alone. This subset consists of only eight between-subjects experiments, four within (long) experiments, and seven within (short) experiments, containing 25, 13, and 23 condition-level data points correspondingly. On the subset alone, the test of moderators of the meta-regression again indicated that study design had a significant impact on the observed effect size ($QM(3) = 9.12, p = .03$. Model results showed that between-subjects design studies have significantly higher effect size ($1.28, 95\%$ CI $0.88 - 1.68, p < .001$) than within (short)'s ($0.90, 95\%$ CI $0.23 - 1.57, p = .006$). While the within (long)'s effect size is also smaller than the between-subjects effect size, the difference is not significant ($1.01, 95\%$ CI $0.33 - 1.69, p = .07$). As a result, only **H2** was supported while **H1** was not. However, considering the very limited number of within (long) experiments reporting delta-SSQ and the p-value being only slightly above .05, we believe the conclusion drawn regarding **H1** using the full dataset remain valid.

We did not observe any significant association between the proportion of female participants and effect size (slope: $-0.19, 95\%$ CI $-0.95 - 0.57, p = .62$). This result did not support **H3**.

VI. The Impact of Carryover Effects

A. Symptoms Carried to the Second Condition

Recall the results we get in Section III, Equation 17. Using the estimated effect size we have in Section V-C, we can estimate the proportion ($k = M_e/M$) of cybersickness symptoms carried over from a previous experimental session to the subsequent one.

1) Conservative Estimation of Effect Size Differences: Since the equations we derived in Section III are based on a study with two conditions, while a large proportion of within-subjects studies in our dataset have more than two conditions, the effect size change we observed in Section V-C is larger than the change for the studies with only two conditions. However, the estimation of carryover effects depends on different assumptions for three condition designs. We consider two extreme conditions: the upper bound condition where the means of three order sessions are M, $M + M_e$, and

$M+2M_e$; and the lower bound condition where the means are M, $M+M_e$, and $M+M_e$. The upper bound case assumes the impact of the first session is fully there at the third session while the lower bound case assumes the impact of the first session is fully washed out. Realistically, the impact should be somewhat in the middle between these two cases. Using Equation 8, we can derive the combined standard deviation SD_c for both cases.

$$SD_{c\_up} = \sqrt{SD^2 + \frac{2nM_e^2}{3(n-1)}} \tag{21}$$

$$SD_{c\_bottom} = \sqrt{SD^2 + \frac{2nM_e^2}{9(n-1)}} \tag{22}$$

Given the above results, we estimate that studies with three conditions have roughly twice the impact of carryover effects on standard deviation per condition than studies with two conditions. We confirmed this estimate through a simple simulation. The simulation code is available in our GitHub repository. As a result, we adjusted our estimations accordingly by weighing experiments with two conditions as 1×2 (two conditions) and those with more than two conditions as $2\times j$ (j is the number of conditions). We adjusted the estimations using Equation 23.

$$\begin{aligned} d_{adj} &= d_{between} - \frac{2 \cdot n_{exp}}{2 \cdot n_{cond=2} + \sum_{j=3}^{max} j \cdot n_{cond=j} \cdot 2}(d_{between} - d_{long}) \\ &= d_{between} - \frac{n_{exp}}{n_{cond=2} + \sum_{j=3}^{max} j \cdot n_{cond=j}}(d_{between} - d_{long}) \end{aligned} \tag{23}$$

In Section V-C, the meta-regression's result shows $d_{between} = 1.42$, $d_{long} = 1.22$, and $d_{short} = 1.06$. Based on Table I, the experiments with long washout periods include seven studies with two conditions, four with three conditions, and one each with four and five conditions. Consequently, we estimate the effect size for the two-condition studies as follows:

$$d_{long\_adj} = d_{between} - \frac{13}{28} * (d_{between} - d_{long}) = 1.33 \tag{24}$$

For short washout periods, the dataset includes eight experiments with two conditions, four with three conditions, two with four conditions, and one experiment each with five and seven conditions.

$$d_{short\_adj} = d_{between} - \frac{16}{40} * (d_{between} - d_{short}) = 1.28 \tag{25}$$

2) Long Washout Period's Carryover Effects: Based on the effect size for two-condition studies we estimated above, we have:

$$\begin{aligned} c_{1\_long} &= \frac{1}{d_{long\_adj}^2} = 0.57 \\ c_2 &= \frac{1}{d_{between}^2} = 0.50 \end{aligned} \tag{26}$$

If we plug them into Equation 17, we can get an estimation of the carryover effect that each participant experiences after the first condition.

$$k_{long} = \frac{M_{e\_long}}{M} = 0.56 \text{ or } -0.36 \tag{27}$$

Recall from Equation 17 that M_e is directly proportional to M. Both solutions theoretically account for the observed effect size differences. However, to select the valid solution, we evaluate them against established boundaries of cybersickness adaptation. A value of $k_{long} = 0.56$ implies a 56% *increase* in cybersickness symptoms after a long washout period. Extensive quantitative results from multi-day VR exposure consistently reports an adaptation effect (symptom reduction) of approximately -30% [16]–[19], [96], [97]. To the best of our knowledge, the only documented instance of positive adaptation is the Climb 2 condition in Kelly et al. [19]. However, Climb 2 involves primarily vertical locomotion with minimal continuous optical flow, a distinctly different profile from the standard locomotion interfaces evaluated here. Therefore, we reject the extraneous root $k_{long} = 0.56$ because it violates the established bounds of multi-day adaptation, leaving $k_{long} = -0.36$ as the unique, physiologically plausible solution. As a result, the estimated carryover effect is around -36% of the cybersickness level induced in the first session.

This negative carryover effect is congruent with previous observations on adaptation to cybersickness, where participants' average SSQ total scores reduced 25.5% from Day 1 to Day 2 (SSQ-TS: 47.41 to 35.31) in a controlled study [17]. Palmisano and Constable found that the disorientation subscore of the SSQ decreased from 52 to 28 between Day 1 and Day 2 [98]. The average SSQ total scores were also reduced by 35% from Day 1 to Day 2 (SSQ-TS: 28.82 to 18.89, extracted using PlotDigitizer from Figure 2 in the paper). This result confirmed the validity of our approach despite our rough estimation.

3) Short Washout Period's Carryover Effects: Similar to the previous section, we can estimate how much cybersickness is carried over to the second condition after a short washout period. Using the meta-regression results, we have:

$$c_{1\_short} = \frac{1}{d_{short\_adj}^2} = 0.61 \tag{28}$$

Plugging c_{1_short} and c_2 into Equation 17, we can get an estimation of the symptoms that get carried over to the second condition.

$$k_{short} = \frac{M_{e\_short}}{M} = 0.79 \text{ or } -0.43 \tag{29}$$

Similarly, evaluating the short-term carryover yields two mathematical roots. A solution of $k_{short} = -0.43$ indicates that a short break not only clears initial symptoms but also generates a 43% reduction in subsequent cybersickness susceptibility. Through a systematic literature review, Rebenitsch and Owen concluded that resolving cybersickness symptoms requires significant time, often spanning hours [27]. Furthermore, Zielasko et al. monitored participants for 90 minutes after VR exposure and mapped the gradual decay of cybersickness symptoms [99]. Achieving a -43% adaptation effect almost immediately post-exposure contradicts these known

symptom decay curves. Therefore, we treat $k_{short} = -0.43$ as an extraneous root representing an unlikely physiological recovery rate. As a result, the estimated carryover effect suggests that 79% of the cybersickness symptoms induced in the first session are carried over to the following session, indicating a short break (< 15 minutes) is not enough to wash out cybersickness symptoms.

This result is also congruent with previous findings. Schmitz et al. administered SSQ directly after (SSQ_POST1) and 10 minutes after (SSQ_POST2) VR exposure and found that SSQ_POST2 is 60% of the SSQ_POST1 (48.92 to 29.53) [49]. Mittelstaedt et al. found that participants' SSQ scores remained 73% (Bike/HMD, 54.0 to 39.5) and 70% (Gamepad/HMD, 52.6 to 37.0) of their original score at the end of VR exposure after performing some tasks for at most 25 minutes [51]. Overall, these results serve as evidence that our calculation on the carryover effect is accurate.

4) Estimations Using Delta-SSQ Data: We also performed the same estimation for k_{long} and k_{short} using the delta-SSQ results from Section V-C2. The resulting estimates are $k_{long} = -0.39$ and $k_{short} = 0.85$, which are consistent with the estimates produced using the full dataset. This consistency confirms the validity of our earlier assumption of treating post-SSQ scores as equivalent to delta-SSQ scores.

B. Impact on Experiment Design and Statistical Power

1) Sample Size Comparison: The Pitman Asymptotic Relative Efficiency (ARE) represents the ratio of sample sizes per condition required to achieve the same power under a within- and a between-subjects design [100].

$$\begin{aligned}\frac{n_w}{n_b} &= \frac{\sigma_w^2}{\sigma_b^2} \\ &= \frac{2(1-r)\sigma^2}{4\sigma^2} \\ &= \frac{1-r}{2}\end{aligned} \tag{30}$$

Here, n_w and n_b are the number of participants for within- and between-subjects design, σ_w and σ_b are the asymptotic variances for each design, and r is the Intraclass Correlation Coefficient (ICC) [101]. We used multiple within-subjects studies' results and determined that the SSQ's ICC is $r = 0.5$. Without any carryover effect, we can see that a within-subjects design requires only one fourth of participants per condition compared with between-subjects'. For an experiment with two conditions, if a within-subjects study needs n participants, a between-subjects study needs $8n$ participants in total to achieve the same power. However, as we will show in Section VI-B2 below, with the carryover effect of short washout periods, this superiority is hard to achieve.

2) With Carryover Effects: We parameterize the outcome means of a within-subject study with two conditions (E and C) below in a way similar to a previous work [101]. M_1^C and M_1^E are the means of the first experiment session while M_2s are the means of the second experiment session, with the superscript denoting the order of conditions applied.

$$\begin{aligned}M_1^C &= \mu \\ M_1^E &= \mu + \theta \\ M_2^{CC} &= \mu + \tilde{\tau} \\ M_2^{EC} &= \mu + \tilde{\tau} + \lambda_0 \\ M_2^{EE} &= \mu + \tilde{\tau} + \theta \\ M_2^{CE} &= \mu + \tilde{\tau} + \theta - \lambda_1\end{aligned} \tag{31}$$

Suppose the alternative hypothesis is that $\theta > 0$. Variables $\lambda_0 = M_2^{EC} - M_2^{CC}$ and $\lambda_1 = M_2^{EE} - M_2^{CE}$ are the expected carryover effects under the two order groups. Using results from sections VI-A2 and VI-A3, we expect:

$$\lambda_0 = \lambda_1 = \theta(M_e/M) \tag{32}$$

This is because we expect the symptoms carried over from the first condition to the second condition are proportional to the symptoms induced in the first condition. As a result, the difference between the carryover effect in the orders that start with E and those that start with C is proportional to $M_1^E - M_1^C = \theta$. For example, for the order effect of cybersickness, we have $\lambda_0 = \lambda_1 = -0.36\theta$. And for the carryover effect, we have $\lambda_0 = \lambda_1 = 0.79\theta$.

Shi and Ye derived the Pitman ARE between within- and between-subject designs with order/carryover effects taken into consideration [101]:

$$\frac{\tilde{n}_w}{n_b} = \frac{\tilde{\sigma}_w^2}{\sigma_b^2} \cdot \left(1 - \frac{\lambda_0 + \lambda_1}{2\theta_{alt}}\right)^{-2} \tag{33}$$

a) Effect of Short Washout Period: If we plug in all the data we have, the Pitman ARE for short washout period is:

$$\begin{aligned}\frac{\tilde{n}_w}{n_b} &= \frac{1-0.5}{2} \cdot \left(1 - \frac{(0.79+0.79)\theta}{2\theta}\right)^{-2} \\ &= 4.73\end{aligned} \tag{34}$$

This result shows that, given the carryover effect we obtained and cross-checked, $4.73\times$ more participants per condition are needed when using a short washout period to achieve the same power compared to using between-subjects design, which is close to the total number of participants needed by a between-subjects design.

b) Effect of Long Washout Period: In their work on carryover effects, Shi and Ye argue that a negative carryover effect ($\lambda_0 = \lambda_1 = -0.36 < 0$) can result in a bloated type I error rate (P_{H_0} in Equation 35), making the statistical test invalid [101]. However, this conclusion may not generalize to our case because of two of its assumptions. First, the negative carryover effect is indeed proportional to the symptoms developed in previous exposure. To the best of our knowledge, there is no empirical evidence supporting this assumption for cybersickness adaptation [17]. Second, this impact on type I error rate only happens when there is a difference between the two conditions, in which case the estimated difference is bloated. However, if there is no actual difference between the conditions, the impact on type I error rate is minimal. In the

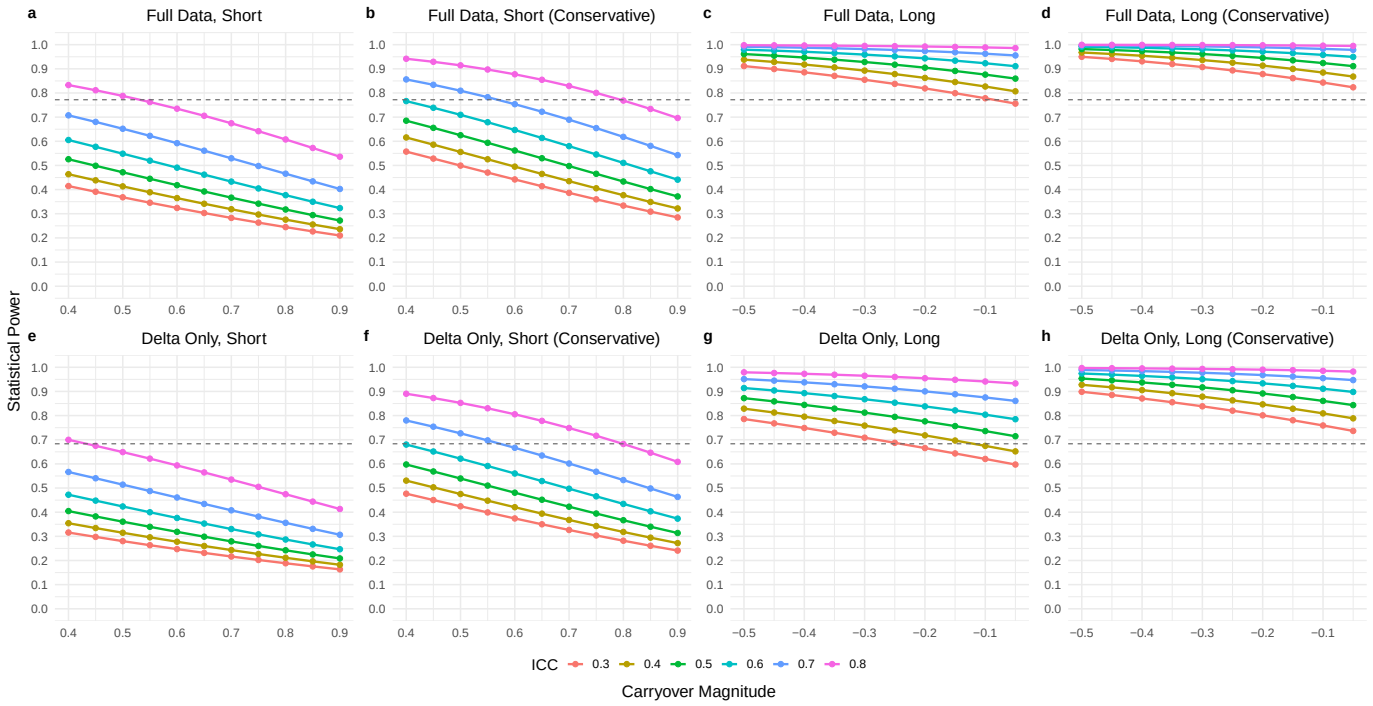

Fig. 4. Simulated statistical power of within-subjects designs across varying carryover magnitudes and intraclass correlation coefficients (ICCs). The y-axis represents statistical power. The x-axis indicates the magnitude of carryover effect, which ranges from 0.4 to 0.9 for short washout periods and -0.5 to -0.1 for long washout periods. Each solid colored line corresponds to a specific ICC value ranging from 0.3 to 0.8, illustrating how higher correlations generally preserve statistical power. Figures (a, c, e, g) were created using the effect sizes without conserved estimation, which were reported in Section V-C Due to the impact of nested carryover effects, these results are more suitable for reference when creating experiments with more than two conditions. Figures (b, d, f, h) were created using the effect sizes with conserved estimation reported in Section VI-A1, which are more suitable for reference when designing experiments with only two conditions. The horizontal dashed gray line in each panel represents the baseline statistical power of an equivalent between-subjects design. The intersection points where the solid lines fall below this dashed threshold indicate the exact carryover magnitudes at which a within-subjects design loses its statistical advantage over a between-subjects design.

future, further research is needed to quantify the impact of a negative carryover effect on type I error rate.

$$P_{H_0}(T_{cr} > z_{1-\alpha}) \approx \Phi\left(-z_{1-\alpha} - \sqrt{n} \cdot \frac{\lambda_0 + \lambda_1}{2\tilde{\sigma}_{cr}}\right) \tag{35}$$

3) Analytical Power Estimation : One of the primary goals of this paper is to demonstrate the differences in statistical power between within-subjects (long), within-subjects (short), and between-subjects designs. We can now achieve this by utilizing the results from Section VI-A.

a) Basics: In a basic within-subjects design with two conditions, control (C) and experimental (E), and proper counterbalancing, the statistical power is determined by the effect size, sample size, and significance level [102]. Since the significance level is typically set to $\alpha = .05$, for a given number of participants, an experiment's power is primarily affected by the effect size *between conditions*, often calculated as Cohen's d (see Equation 12). Based on Equation 12, we know that $d \propto M_{diff}/SD_{diff}$. In Section III-A, we demonstrated the impact of the carryover effect magnitude, k, on M_{diff}. Conversely, SD_{diff} for within-subjects studies is influenced by both the *condition-level pre-post* standard deviations and the Intraclass Correlation Coefficient (*icc*), as shown in Equation 36. For within-subjects studies, a higher *icc* results in smaller within-subjects variance, thereby increasing power compared to an equivalent between-subjects study with the same number of participants per condition.

$$SD_{diff} = \sqrt{SD_C^2 + SD_E^2 - 2 \cdot icc \cdot SD_C \cdot SD_E} \tag{36}$$

To account for the uncertainty in our estimations of the carryover magnitude (k) and the *icc*, we evaluated various combinations of these parameters to provide a comprehensive view. This allows readers to make informed judgments based on their own experience and historical data. The power analysis was implemented in R using `power.t.test`; the code is available in our GitHub repository.

b) Setup: We conducted an analytical power estimation to evaluate the comparative statistical power between independent and paired study designs under varying levels of data contamination. In this estimation, we defined a sample size of 30 and assumed two conditions: a control condition C with a mean SSQ of 50, and an experimental condition E with a mean SSQ of 30. We calculated the standard deviation for both conditions using Equation 12 with the estimated condition-level effect sizes in Section VI-A to reflect the impact of carryover effects.

c) Simulation Method: For the between-subjects estimations, we calculated the statistical power for an independent samples t-test using the standard deviations of the two groups. For the within-subjects estimations, we calculated statistical

power based on the mean difference, M_{diff}, using Equation 2 and the standard deviation of the difference scores, SD_{diff}, using Equation 36. The mean difference between conditions is directly affected by carryover effects. Rather than adding an effect to a generated dataset, our method accounts for these effects by systematically updating the expected mean difference, as shown in Section III-A. Specifically, we reduced the mean difference by a factor proportional to the carryover magnitude (k), as calculated using Equation 2.

Due to individual differences, the carryover magnitude k may vary across the populations from which researchers recruit participants. Moreover, while the intraclass correlation coefficients (ICC) of the SSQ are centered around 0.5, they also show variation across experiments, likely due to individual differences in cybersickness susceptibility. To help researchers make informed decisions when they can estimate these parameters from past experiments or pilot testing, or when future research provides more empirical estimates for both k and ICC, we applied this calculation across a sequence of possible carryover magnitudes ($k \in [0.4, 0.9]$ for short washout period and $k \in [-0.05, -0.5]$ for long washout period) alongside a range of intraclass correlation coefficients ($icc \in [0.3, 0.8]$). This allowed us to map the exact thresholds where the power of a paired design drops below that of an independent design, while accounting for the uncertainties within our estimations.

d) Results: Figure 4 shows the results of our power analyses for different combinations of carryover effects and ICC. The simulated power analysis revealed that within-subjects designs are highly vulnerable to positive carryover effects, which heavily penalize the expected mean difference in within (short) experiments for both of our estimations from full data (Figure 4a, b) and delta-only data (Figure 4e, f). As the carryover magnitude increased from 0.4 to 0.9, statistical power for the within (short) design degraded rapidly. Crucially, the within-subjects design only outperformed the between-subjects baseline (represented by the dashed horizontal line) under highly optimal conditions. Based on our estimations, the power of the between-subjects design was 0.77 for the full dataset and 0.68 for the delta-only subset. Specifically, when carryover was minimal ($c \leq 0.45$) and the intraclass correlation was exceptionally high ($r = 0.8$). For any $ICC \leq 0.7$, or at carryover magnitudes exceeding 0.5, the between-subjects design consistently yielded superior statistical power.

In contrast, within (long) experiments (Figure 4c, d, g, h) demonstrated significant robustness. Because the simulated carryover effects in these sequences were negative ($c \in [-0.5, -0.1]$), they increased the expected mean difference rather than diminishing it. Consequently, the within-subjects design maintained high power. Across almost the entire range of negative carryover magnitudes and ICC combinations, the paired design outperformed the independent baseline. The only exception was observed at the weakest correlation ($ICC = 0.3$) combined with the smallest negative carryover magnitude ($c = -0.1$), where power dropped marginally below the between-subjects threshold.

Across all modeled scenarios, higher ICC values systematically buffered the within-subjects design against power loss. This highlights that tightly correlated paired data is essential to justify a within-subjects approach when order effects are present. Furthermore, while utilizing delta-only estimations (Figure 4e–h) rather than full data estimations (Figure 4a–d) resulted in a universal, absolute reduction in overall statistical power, the relative behavioral trends remained identical. Finally, the specific tipping points at which the independent design became more statistically powerful than the paired design were structurally preserved regardless of whether the normal or conservative effect size estimates were applied.

VII. General Discussion

A. Within-Subjects Designs

Contrary to the common assumption that within-subjects designs universally enhance statistical power by accounting for inter-individual variability, our findings suggest that this is not always the case within the context of cybersickness research.

1) Short Washout Period: Our results showed that a < 15 minute washout period can result in approximately 79% of the symptoms from the first session carrying over to the second session. Under this level of carryover, our power analysis demonstrated that within-subject designs with short washout periods can significantly reduce statistical power due to interference effects between conditions, making them undesirable for researchers.

A within-subject design with a short washout period can be an efficient approach when researchers have high confidence in their results. This method significantly shortens the experiment duration and reduces the number of participant visits, which may increase recruitment and lower dropout rates. If researchers find a significant difference between conditions using this design, the use of such a design is not necessarily a reason to reject the paper, since the results are conservative. However, other potential issues, such as non-linear symptom development as discussed in [21], may still be confounding factors if confirmed. In general, unless researchers are very confident in the experiment's outcome, they should use caution when testing all conditions on the same day.

2) Long Washout Period: A promising compromise is the use of within-subject designs with sufficiently long washout periods to mitigate carryover effects. These designs preserve the advantage of controlling for individual variability while minimizing condition-to-condition contamination. As expected, our meta-analysis showed that such designs offered the highest statistical power, combining the strengths of both within- and between-subject approaches.

However, this design also involves trade-offs. It typically requires more time and effort per participant and remains vulnerable to residual order effects, which must be carefully counterbalanced. Notably, counterbalancing does not fully eliminate these effects, and their presence can still undermine result reliability if not properly modeled. One study explored cybersickness abatement from repeated exposures using the classic FOV restriction method and full FOV, and found that repeated exposure to the FOV restrictor led to lower cybersickness scores later when the same participant was exposed to the full FOV [17], indicating that even long washout periods can still expose participants to adaptation effects.

Nevertheless, when a long washout period is included, this design consistently outperforms both short-washout within-subject designs and between-subject designs in terms of power.

One potential concern is that symptom adaptation may exaggerate differences between conditions if the degree of adaptation is proportional to the severity of symptoms experienced in the first condition. Researchers should weigh the added effort of using washout periods longer than 24 hours against these risks when choosing a study design.

B. Between-Subjects Designs

A between-subjects design offers several advantages for research studies. It generally has higher power than a within-subjects design with a short washout period, as it avoids carryover effects. Furthermore, because each participant is exposed to only one condition, they are less likely to infer the hypothesis of the study, which could potentially reduce bias. This design also facilitates participant recruitment, as the shorter time commitment per individual can lower dropout rates and, in some cases, reduce the overall experiment duration compared to a within-subjects design. While between-subjects designs may introduce greater inter-participant variability, this diversity can enhance the representativeness of the sample. However, a key drawback is the need for a larger sample size to ensure reliable results.

C. Which Method to Choose?

When selecting a study design for cybersickness research, prioritizing the avoidance of carryover effects is crucial.

- If the study design is simple and participant availability is limited, a within-subject design with a longer washout period is recommended to minimize carryover effects.
- If a large participant pool is available and/or experiment time is limited, a between-subject design is preferable to ensure sufficient statistical power and avoid order effects.
- For complex study designs (i.e. > 3 conditions), a between-subject design is ideal to prevent participant fatigue and dropouts. This is especially important in cybersickness research, where severe symptoms in one session may lead to dropouts, wasting study time.
- A within-subject design with a short washout period should only be considered as a last resort—when both participant availability and experiment time are limited, and there is strong confidence in the study results.

A position paper has also provided valuable guidelines on design choices for cybersickness experiments [21]. A more advanced, empirically validated cybersickness experiment protocol with shorter VR exposure is also desired to minimize carryover effects if it can be developed, even though attempts like this are sparse (like [6]) and have not been widely used.

D. Effects of Sex

There is a wide body of research exploring the effects of sex on cybersickness, although findings across studies remain mixed. One recent analysis finds additional heterogeneity introduced into experimental outcomes [103], and some works find that women tend to experience more cybersickness than men [104], [105], while others have found no significant difference [106]–[108]. One study found incidence higher for women, while the severity of the sickness remains similar [109]. Another tied visually induced motion sickness much more closely to motion sickness susceptibility rather than sex [110]. Nevertheless, the effect of sex is generally considered small: studies with sufficient statistical power tend to detect an effect, whereas underpowered studies often do not [103].

In Section V-C, we observed an effect similar to that reported by Peck et al. [20] using the full dataset and a slightly different methodological approach. However, within the delta-SSQ subset, although the trend still suggests that a higher proportion of female participants reduces effect sizes, the effect was non-significant. We attribute this partially to the fact that the delta-SSQ subset contains less than half the number of experiments compared to the full dataset. Our results suggest that a link between participant sex distribution and observed cybersickness outcomes may exist. Therefore, researchers should not overlook sex balance in study samples, as ensuring diversity is crucial to the validity and generalizability of their findings.

E. Data Reporting

It is strongly recommended to collect both pre- and post-SSQ scores [45]. While some participants may be symptom-free before the experiment, the pre-SSQ remains essential for establishing a baseline and prescreening participants to ensure the study's validity [89]. This is particularly important in within-subject designs with short washout periods. Some researchers use the post-SSQ from the previous session as the pre-SSQ for the next; however, since symptoms may naturally subside during rest, these scores no longer accurately reflect the participant's state at the beginning of the next session. Therefore, a new pre-SSQ measurement should be administered to ensure accurate and reliable results. In the future, researchers should report SSQ scores accurately, along with sex distribution, to enable a clearer understanding of carryover effects through aggregated data.

VIII. Limitations

There are several limitations to this meta-analysis and the accompanying power analysis. First, the short washout periods we sampled range from 0 to 15 minutes, which introduces a few primary limitations. On one hand, there is a substantial gap between a 15-minute washout period and the practice of separating experimental conditions across different days. The carryover effects for durations within this gap remain underexplored due to limited data, making it difficult to draw conclusions regarding short washout periods longer than 15 minutes. The generalizability of our findings to washout periods exceeding 15 minutes should be reevaluated as more data becomes available, ideally from at least eight to ten experiments. If future research demonstrates that a washout period of $T_{washout} > 15$ minutes can offer comparable statistical power to a within-subjects (long) design, researchers might consider adopting it, provided the longer duration is

acceptable. However, if the $T_{washout}$ required to achieve power similar to a $T_{washout} > 24$-hour design is prohibitively long (e.g., > 30 minutes), researchers must weigh the extra resources required, such as increased participant compensation, against the potential benefits. On the other hand, there is significant variation in washout periods across experiments that test multiple sessions on the same day. In some cases, the washout period varies within a single experiment due to design constraints, such as conditions being met before a break can officially end. This variability introduces additional noise and limits the accuracy of our results. Additionally, we could not specify how different durations of short washout periods affect experiment outcomes due to limited sample sizes and missing information in many studies. Our current conclusions are based on the 16 experiments as a collective group. Therefore, results may worsen when using a very short washout period, such as one minute.

Second, we assumed that cybersickness symptoms develop and accumulate linearly. While this assumption simplifies the analysis, we acknowledge that symptom development may follow a non-linear trajectory and grow faster when participants are already sick [21]. Nonetheless, this linearity assumption is conservative and unlikely to significantly distort our results.

Third, we hypothesized that carryover effects are proportional to the symptoms induced by prior exposure. While this may be a reasonable assumption for positive carryover effects, its applicability to the negative carryover effects (adaptation) remains uncertain [17].

Finally, although we discussed how experimental design influences effect sizes, other factors also clearly affect outcomes, such as experiment duration and the method used to induce cybersickness. We assumed these factors were distributed evenly across the three experimental designs, though this may not always hold true. Additionally, such factors are not consistently reported in the literature or are difficult to quantify and incorporate into our model. As noted in Section IV-F3, the weighted average SSQ scores for within-subject designs with short washout periods were lower than those of the other two designs, suggesting that researchers may favor this approach when the stimulus is relatively mild. However, based on our observations, stronger stimuli also tend to increase data dispersion due to individual differences. This is evident in Figure 3, where studies with larger effect sizes also show greater variance, as indicated by smaller bubble sizes, and are thus weighted less in the analysis.

Given these limitations, we acknowledge the potential for confounding effects. Further investigation will be possible if more detailed data become available in future studies.

IX. Conclusion

Cybersickness remains a significant challenge for VR researchers, and experimental design choices can affect the validity and reliability of their findings. This meta-analysis demonstrates that short washout periods in within-subject studies can significantly compromise statistical power due to carryover effects, while long washout periods and between-subject designs are more reliable. We estimate that short breaks leave up to 79% of symptoms unresolved, while long breaks result in an adaptation effect of approximately -36%. Based on these findings, we recommend avoiding short washout periods when possible. Researchers should favor long washout periods or between-subject designs to increase their chances of detecting a true effect. Ensuring sex balance and consistently reporting both pre- and post-SSQ scores will further strengthen the validity and reproducibility of future studies.

Acknowledgments

The authors would like to thank Victoria Interrante, Jerald Thomas and Guanghan Zhao for their encouragement, support, and feedback on this project at different stages. This work was partially supported by the University of Minnesota Doctoral Dissertation Fellowship program and the University of Minnesota President's Postdoctoral Fellowship Program.

References

[1] C. M. O'Connor, D. R. Norris, G. T. Crossin, and S. J. Cooke, "Biological carryover effects: linking common concepts and mechanisms in ecology and evolution," *Ecosphere*, vol. 5, no. 3, pp. 1–11, 2014.

[2] N. Dużmańska, P. Strojny, and A. Strojny, "Can simulator sickness be avoided? a review on temporal aspects of simulator sickness," *Frontiers in psychology*, vol. 9, 2018.

[3] A. S. Fernandes and S. K. Feiner, "Combating VR sickness through subtle dynamic field-of-view modification," in *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, Mar. 2016, pp. 201–210.

[4] Z. Cao, J. Jerald, and R. Kopper, "Visually-Induced Motion Sickness Reduction via Static and Dynamic Rest Frames," in *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2018, pp. 105–112.

[5] F. Wu and E. S. Rosenberg, "Adaptive Field-of-view Restriction: Limiting Optical Flow to Mitigate Cybersickness in Virtual Reality," in *Proc. of ACM VRST*, 2022, pp. 1–11.

[6] G. Zhao, J. Orlosky, S. Feiner, P. Ratsamee, and Y. Uranishi, "Mitigation of VR Sickness During Locomotion With a Motion-Based Dynamic Vision Modulator," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 10, pp. 4089–4103, 2023.

[7] C. Groth, M. Magnor, S. Grogorick, M. Eisemann, and P. Didyk, "Cybersickness Reduction via Gaze-Contingent Image Deformation," *ACM Transactions on Graphics*, vol. 43, no. 4, 2024.

[8] G.-Y. Nie, H. B.-L. Duh, Y. Liu, and Y. Wang, "Analysis on Mitigation of Visually Induced Motion Sickness by Applying Dynamical Blurring on a User's Retina," *IEEE TVCG*, vol. 26, no. 8, pp. 2535–2545, 2020.

[9] S. Ang and J. Quarles, "Reduction of cybersickness in head mounted displays use: A systematic review and taxonomy of current strategies," *Frontiers in Virtual Reality*, vol. 4, 2023.

[10] ——, "You're in for a Bumpy Ride! Uneven Terrain Increases Cybersickness While Navigating with Head Mounted Displays," in *IEEE Conf. on Virtual Reality and 3D User Interfaces*, 2022, pp. 428–435.

[11] S. H. Park, B. Han, and G. J. Kim, "Mixing in Reverse Optical Flow to Mitigate Vection and Simulation Sickness in Virtual Reality," in *Proc. of CHI*. ACM, 2022, pp. 1–11.

[12] Z. Lin, X. Gu, S. Li, Z. Hu, and G. Wang, "Intentional Head-Motion Assisted Locomotion for Reducing Cybersickness," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 8, pp. 3458–3471, Aug. 2023.

[13] M. Al Zayer, I. B. Adhanom, P. MacNeilage, and E. Folmer, "The effect of field-of-view restriction on sex bias in VR sickness and spatial navigation performance," in *Proc. of CHI*. ACM, 2019, pp. 1–12.

[14] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aerosp. Psychol.*, vol. 3, no. 3, pp. 203–220, 1993.

[15] B. Keshavarz and H. Hecht, "Validating an Efficient Method to Quantify Motion Sickness," *Hum. Factors*, vol. 53, no. 4, pp. 415–426, 2011.

[16] I. Adhanom, S. Halow, E. Folmer, and P. MacNeilage, "VR Sickness Adaptation With Ramped Optic Flow Transfers From Abstract To Realistic Environments," *Frontiers in Virtual Reality*, vol. 3, May 2022.

[17] T. A. Doty, J. W. Kelly, S. B. Gilbert, and M. C. Dorneich, "Cybersickness Abatement from Repeated Exposure to VR with Reduced Discomfort," *IEEE Trans. Vis. Comput. Graph.*, pp. 1–12, 2024.

[18] S. Palmisano, B. Arcioni, and P. J. Stapley, "Predicting vection and visually induced motion sickness based on spontaneous postural activity," *Experimental Brain Research*, vol. 236, no. 1, pp. 315–329, 2018.

[19] J. W. Kelly, T. A. Doty, M. C. Dorneich, and S. B. Gilbert, "Cybersickness Abatement From Repeated Exposure Generalizes Across Experiences," *IEEE Trans. Vis. Comput. Graph.*, vol. 32, no. 4, pp. 2974–2985, 2026.

[20] T. C. Peck, L. E. Sockol, and S. M. Hancock, "Mind the Gap: The Underrepresentation of Female Participants and Authors in Virtual Reality Research," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 5, pp. 1945–1954, 2020.

[21] D. Zielasko, B. Rehling, D. Clement, and G. Domes, "Carry-over effects ruin your (cybersickness) experiments and balancing conditions is not a solution," in *IEEE VRW*, 2024, pp. 1–5.

[22] N. Biswas, A. Mukherjee, and S. Bhattacharya, ""Are you feeling sick?" – A systematic literature review of cybersickness in virtual reality," *ACM Comput. Surv.*, vol. 56, no. 11, Jun. 2024.

[23] T. van Gemert, N. C. Nilsson, T. Hirzle, and J. Bergström, "Sicknificant steps: A systematic review and meta-analysis of vr sickness in walking-based locomotion for virtual reality," in *Proc. of CHI*. ACM, 2024, pp. 1–36.

[24] D. Zielasko and T. Weissker, "Stay vigilant: The threat of a replication crisis in vr locomotion research," in *Proc. of ACM VRST*, 2023.

[25] N. Tian, P. Lopes, and R. Boulic, "A review of cybersickness in head-mounted displays: raising attention to individual susceptibility," *Virtual Reality*, vol. 26, no. 4, pp. 1409–1441, Dec. 2022.

[26] T. Hirzle, M. Cordts, E. Rukzio, J. Gugenheimer, and A. Bulling, "A Critical Assessment of the Use of SSQ as a Measure of General Discomfort in VR Head-Mounted Displays," in *2021 CHI Conference on Human Factors in Computing Systems*. ACM.

[27] L. Rebenitsch and C. Owen, "Review on cybersickness in applications and visual displays," *Virtual Reality*, vol. 20, no. 2, pp. 101–125, 2016.

[28] X. Li, D.-B. Luh, R.-H. Xu, and Y. An, "Considering the consequences of cybersickness in immersive virtual reality rehabilitation: A systematic review and meta-analysis," *Applied Sciences*, vol. 13, no. 8, 2023.

[29] A. J. Martingano, E. Brown, S. H. Telaak, A. P. Dolwick, and S. Persky, "Cybersickness variability by race: Findings from 6 studies and a mini meta-analysis," *J. Med. Internet Res.*, vol. 24, no. 6, p. e36843, 2022.

[30] D. Saredakis, A. Szpak, B. Birckhead, H. A. D. Keage, A. Rizzo, and T. Loetscher, "Factors associated with virtual reality sickness in head-mounted displays: A systematic review and meta-analysis," *Frontiers in Human Neuroscience*, vol. 14, 2020.

[31] M. C. Howard and E. C. Van Zandt, "A meta-analysis of the virtual reality problem: Unequal effects of virtual reality sickness across individual differences," *Virtual Real.*, vol. 25, no. 4, p. 1221–1246, Dec. 2021.

[32] K. M. Stanney and R. S. Kennedy, "The psychometrics of cybersickness," *Communications of the ACM*, vol. 40, no. 8, pp. 66–68, 1997.

[33] S.-H. Liu, N.-H. Yu, L. Chan, Y.-H. Peng, W.-Z. Sun, and M. Y. Chen, "PhantomLegs: Reducing Virtual Reality Sickness Using Head-Worn Haptic Devices," in *IEEE VR*, 2019, pp. 817–826.

[34] S. Freitag, B. Weyers, and T. W. Kuhlen, "Examining rotation gain in cave-like virtual environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 4, pp. 1462–1471, 2016.

[35] P. A. Howarth and S. G. Hodder, "Characteristics of habituation to motion in a virtual environment," *Displays*, vol. 29, no. 2, pp. 117–123, 2008.

[36] G. Wang and A. Suh, "User adaptation to cybersickness in virtual reality: A qualitative study," in *27th European Conference on Information Systems: Information Systems for a Sharing Society*. AIS, 2019.

[37] A. Adhikari, A. M. Hashemian, T. Nguyen-Vo, E. Kruijff, M. v. d. Heyde, and B. E. Riecke, "Lean to Fly: Leaning-Based Embodied Flying can Improve Performance and User Experience in 3D Navigation," *Frontiers in Virtual Reality*, vol. 2, Sep. 2021.

[38] T. Nguyen-Vo, B. E. Riecke, W. Stuerzlinger, D.-M. Pham, and E. Kruijff, "NaviBoard and NaviChair: Limited Translation Combined with Full Rotation for Efficient Virtual Locomotion," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 1, pp. 165–177, 2021.

[39] R. Rouhani, N. Umatheva, J. Brockerhoff, B. Keshavarz, E. Kruijff, J. Gugenheimer, and B. E. Riecke, "Towards benchmarking VR sickness: A novel methodological framework for assessing contributing factors and mitigation strategies through rapid VR sickness induction and recovery," *Displays*, vol. 84, p. 102807, 2024.

[40] W. Xu, H.-N. Liang, K. Yu, and N. Baghaei, "Effect of Gameplay Uncertainty, Display Type, and Age on Virtual Reality Exergames," in *Proc. of CHI*. ACM, 2021.

[41] D. Wolf, K. Rogers, C. Kunder, and E. Rukzio, "JumpVR: Jump-Based Locomotion Augmentation for Virtual Reality," in *Proc. of CHI*. ACM, 2020, pp. 1–12.

[42] I. B. Adhanom, D. Zielasko, R. Islam, G. Kim, and E. S. Rosenberg, "The sixth workshop on immersive sickness prevention," in *IEEE VRW*, 2025, pp. 813–814.

[43] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *bmj*, vol. 372, 2021.

[44] A. P. Field and R. Gillett, "How to do a meta-analysis," *Br. J. Math. Stat. Psychol.*, vol. 63, no. 3, pp. 665–694, 2010.

[45] P. Bimberg, T. Weissker, and A. Kulik, "On the Usage of the Simulator Sickness Questionnaire for Virtual Reality Research," in *IEEE VRW*, 2020, pp. 464–467.

[46] K. M. Stanney, R. S. Kennedy, and J. M. Drexler, "Cybersickness is not simulator sickness," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* SAGE Publications, 1997, pp. 1138–1142.

[47] S. Park, L. Kim, J. Kwon, S. J. Choi, and M. Whang, "Evaluation of visual-induced motion sickness from head-mounted display using heartbeat evoked potential: a cognitive load-focused approach," *Virtual Reality*, vol. 26, no. 3, pp. 979–1000, Sep. 2022.

[48] M. Bessa, M. Melo, D. Narciso, L. Barbosa, and J. Vasconcelos-Raposo, "Does 3D 360 video enhance user's VR experience? An Evaluation Study," in *XVII Intl. Conf. on Human Computer Interaction*. ACM, 2016.

[49] P. Schmitz, J. Hildebrandt, A. C. Valdez, L. Kobbelt, and M. Ziefle, "You Spin my Head Right Round: Threshold of Limited Immersion for Rotation Gains in Redirected Walking," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1623–1632, Apr. 2018.

[50] R. Tyrrell, H. Sarig-Bahat, K. Williams, G. Williams, and J. Treleaven, "Simulator sickness in patients with neck pain and vestibular pathology during virtual reality tasks," *Virtual Real.*, vol. 22, no. 3, pp. 211–219, 2018.

[51] J. M. Mittelstaedt, J. Wacker, and D. Stelling, "VR aftereffect and the relation of cybersickness and cognitive performance," *Virtual Real.*, vol. 23, no. 2, pp. 143–154, Jun. 2019.

[52] A. Szpak, S. C. Michalski, D. Saredakis, C. S. Chen, and T. Loetscher, "Beyond Feeling Sick: The Visual and Cognitive Aftereffects of Virtual Reality," *IEEE Access*, vol. 7, pp. 130 883–130 892, 2019.

[53] Y.-H. Peng, C. Yu, S.-H. Liu, C.-W. Wang, P. Taele, N.-H. Yu, and M. Y. Chen, "Walkingvibe: Reducing virtual reality sickness and improving realism while walking in vr using unobtrusive head-mounted vibrotactile feedback," in *Proc. of CHI*. ACM, 2020, pp. 1–12.

[54] C. Yildirim, "Don't make me sick: investigating the incidence of cybersickness in commercial virtual reality headsets," *Virtual Reality*, vol. 24, no. 2, pp. 231–239, Jun. 2020.

[55] R. Islam, Y. Lee, M. Jaloli, I. Muhammad, D. Zhu, P. Rad, Y. Huang, and J. Quarles, "Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals," in *IEEE ISMAR*, 2020, pp. 400–411.

[56] J. Teixeira, S. Miellet, and S. Palmisano, "Unexpected Vection Exacerbates Cybersickness During HMD-Based Virtual Reality," *Frontiers in Virtual Reality*, vol. 3, Apr. 2022.

[57] N. C. Sepich, A. Jasper, S. Fieffer, S. B. Gilbert, M. C. Dorneich, and J. W. Kelly, "The impact of task workload on cybersickness," *Frontiers in Virtual Reality*, vol. 3, Aug. 2022.

[58] S. Chen and D. Weng, "The temporal pattern of vr sickness during 7.5-h virtual immersion," *Virtual Real.*, vol. 26, no. 3, pp. 817–822, 2022.

[59] C. A. T. Cortes, C.-T. Lin, T.-T. N. Do, and H.-T. Chen, "An EEG-based Experiment on VR Sickness and Postural Instability While Walking in Virtual Environments," in *IEEE VR*, 2023, pp. 94–104.

[60] R. Venkatakrishnan, R. Venkatakrishnan, B. Raveendranath, D. M. Sarno, A. C. Robb, W.-C. Lin, and S. V. Babu, "The Effects of Auditory, Visual, and Cognitive Distractions on Cybersickness in Virtual Reality," *IEEE Trans. Vis. Comput. Graph.*, pp. 1–16, 2023.

[61] D. Achanccaray and H. Sumioka, "A physiological approach of presence and vr sickness in simulated teleoperated social tasks," in *IEEE Intl. Conf. on Systems, Man, and Cybernetics*, 2023, pp. 4562–4567.

[62] S. O. Thorp, L. M. Rimol, S. Lervik, H. R. Evensmoen, and S. Grassini, "Comparative analysis of spatial ability in immersive and non-immersive virtual reality: the role of sense of presence, simulation sickness and cognitive load," *Front. Virtual Real.*, vol. 5, 2024.

[63] R. Venkatakrishnan, R. Venkatakrishnan, B. Raveendranath, R. Canales, D. M. Sarno, A. C. Robb, W. Lin, and S. V. Babu, "The Effects of Secondary Task Demands on Cybersickness in Active

Exploration Virtual Reality Experiences," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 5, pp. 2745–2755, 2024.
[64] J. Mittelstaedt, G. Huelmann, C. Marggraf-Micheel, A. Schiller, C. Seehof, and D. Stelling, "Cybersickness with passenger vr in the aircraft: Influence of turbulence and vr content," *Virtual Real.*, vol. 28, no. 2, p. 112, 2024.
[65] J. Teixeira, S. Miellet, and S. Palmisano, "Effects of vection type and postural instability on cybersickness," *Virtual Reality*, vol. 28, no. 2, p. 82, Mar. 2024.
[66] S. Palmisano, L. Stephenson, R. G. Davies, J. Kim, and R. S. Allison, "Testing the 'differences in virtual and physical head pose' and 'subjective vertical conflict' accounts of cybersickness," *Virtual Reality*, vol. 28, no. 1, p. 22, Jan. 2024.
[67] J. N. Setu, J. M. Le, R. K. Kundu, B. Giesbrecht, T. Höllerer, K. A. Hoque, K. Desai, and J. Quarles, "Mazed and confused: A dataset of cybersickness, working memory, mental load, physical load, and attention during a real walking task in vr," in *IEEE ISMAR*, 2024, pp. 1048–1057.
[68] J. W. Kelly, T. A. Doty, S. B. Gilbert, and M. C. Dorneich, "Field of View Restriction and Snap Turning as Cybersickness Mitigation Tools," *IEEE Trans. Vis. Comput. Graph.*, pp. 1–9, 2024.
[69] L. Nisiotis, P. Hadjidemetriou, and N. Nouhi, "Exploring the Time Dilation Gameplay in VR, and Its Effect on Presence, VR Sickness, and Performance," in *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024, pp. 1–6.
[70] T. M. Benz, B. Riedl, and L. L. Chuang, "Projection displays induce less simulator sickness than head-mounted displays in a real vehicle driving simulator," in *Proc. of AutoUI*. ACM, 2019, pp. 379–387.
[71] I. B. Adhanom, N. Navarro Griffin, P. MacNeilage, and E. Folmer, "The effect of a foveated field-of-view restrictor on vr sickness," in *IEEE VR*, 2020, pp. 645–652.
[72] S. Wibirama, P. I. Santosa, P. Widyarani, N. Brilianto, and W. Hafidh, "Physical discomfort and eye movements during arbitrary and optical flow-like motions in stereo 3D contents," *Virtual Reality*, vol. 24, no. 1, pp. 39–51, Mar. 2020.
[73] N. Ranasinghe, P. Jain, D. Tolley, S. Karwita Tailan, C. C. Yen, and E. Y.-L. Do, "Exploring the Use of Olfactory Stimuli Towards Reducing Visually Induced Motion Sickness in Virtual Reality," in *ACM Symposium on Spatial User Interaction*, 2020.
[74] I. B. Adhanom, M. Al-Zayer, P. Macneilage, and E. Folmer, "Field-of-View Restriction to Reduce VR Sickness Does Not Impede Spatial Learning in Women," *ACM Trans. Appl. Percept.*, vol. 18, no. 2, 2021.
[75] S. Vlahovic, M. Suznjevic, N. Pavlin-Bernardic, and L. Skorin-Kapov, "The Effect of VR Gaming on Discomfort, Cybersickness, and Reaction Time," in *Intl. Conf. on QoMEX*, 2021, pp. 163–168.
[76] K. M. T. Pöhlmann, G. Li, M. Mcgill, R. Markoff, and S. A. Brewster, "You spin me right round, baby, right round: Examining the impact of multi-sensory self-motion cues on motion sickness during a vr reading task," in *Proc. of CHI*. ACM, 2023.
[77] V. U. Weser, J. Sieberer, J. Berry, and Y. Tuakli-Wosornu, "Navigation in immersive virtual reality: a comparison of 1:1 walking to 1:1 wheeling," *Virtual Reality*, vol. 28, no. 1, p. 4, Dec. 2023.
[78] S. Vlahovic, L. Skorin-Kapov, M. Suznjevic, and N. Pavlin-Bernardic, "Not just cybersickness: short-term effects of popular VR game mechanics on physical discomfort and reaction time," *Virtual Reality*, vol. 28, no. 2, p. 108, May 2024.
[79] P. Hu, Q. Sun, P. Didyk, L.-Y. Wei, and A. E. Kaufman, "Reducing simulator sickness with perceptual camera control," *ACM Trans. Graph.*, vol. 38, no. 6, Nov. 2019.
[80] N. Tian, R. Clement, P. Lopes, and R. Boulic, "On the Effect of the Vertical Axis Alignment on Cybersickness and Game Experience in a Supine Posture," in *2020 IEEE Conference on Games*, pp. 359–366.
[81] A. Matviienko, F. Müller, M. Zickler, L. A. Gasche, J. Abels, T. Steinert, and M. Mühlhäuser, "Reducing Virtual Reality Sickness for Cyclists in VR Bicycle Simulators," in *2022 CHI Conference on Human Factors in Computing Systems*. ACM.
[82] A. Paroz and L. E. Potter, "Investigating External Airflow and Reduced Room Temperature to Reduce Virtual Reality Sickness," in *33rd Australian Conference on Human-Computer Interaction*. ACM, 2022, pp. 198–207.
[83] W. Al-Ashwal, H. Asadi, M. C. Qazani, S. Mohamed, S. Nahavandi, and B. E. Riecke, "Galvanic Skin Response and AE-LSTM for Anomaly Detection in VR-Induced Motion Sickness," in *2024 IEEE International Systems Conference (SysCon)*, Apr. 2024, pp. 1–7.
[84] J. Mayor, L. Raya, S. Bayona, and A. Sanchez, "A virtual reality perceptual study of multi-Technique redirected walking method," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–10, 2024.
[85] S. Ang and J. Quarles, "SmoothRide: A Versatile Solution to Combat Cybersickness in Elevation-Altering Environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 11, pp. 7152–7161, 2024.
[86] M. Gabes and A. Mühlberger, "Driving simulation: the effects of interactivity and presentation setting," *Front. Virtual Real.*, vol. 5, 2024.
[87] B. J. Becker, "Synthesizing standardized mean-change measures," *Br. J. Math. Stat. Psychol.*, vol. 41, no. 2, pp. 257–278, 1988.
[88] J. F. Golding, "Predicting individual differences in motion sickness susceptibility by questionnaire," *Personality and Individual Differences*, vol. 41, no. 2, pp. 237–248, 2006.
[89] P. Brown, P. Spronck, and W. Powell, "The simulator sickness questionnaire, and the erroneous zero baseline assumption," *Frontiers in Virtual Reality*, vol. 3, pp. 1–14, Sep. 2022.
[90] J. J. Deeks, J. P. Higgins, D. G. Altman, and C. S. M. Group, "Analysing data and undertaking meta-analyses," *Cochrane handbook for systematic reviews of interventions*, pp. 241–284, 2019.
[91] B. R. da Costa, E. Nüesch, A. W. Rutjes, B. C. Johnston, S. Reichenbach, S. Trelle, G. H. Guyatt, and P. Jüni, "Combining follow-up and change data is valid in meta-analyses of continuous outcomes: a meta-epidemiological study," *J. Clin. Epidemiol.*, vol. 66, no. 8, pp. 847–855, 2013.
[92] J. P. Higgins, T. Li, and J. J. Deeks, "Choosing effect measures and computing estimates of effect," *Cochrane handbook for systematic reviews of interventions*, pp. 143–176, 2019.
[93] T. Li, T. Yu, B. S. Hawkins, and K. Dickersin, "Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis," *PloS one*, vol. 10, no. 8, p. e0133023, 2015.
[94] J. P. Higgins, S. Eldridge, and T. Li, "Including variants on randomized trials," *Cochrane handbook for systematic reviews of interventions*, pp. 569–593, 2019.
[95] W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," *J. Stat. Softw.*, vol. 36, no. 3, pp. 1–48, 2010.
[96] R. S. Kennedy, K. M. Stanney, and W. P. Dunlap, "Duration and Exposure to Virtual Environments: Sickness Curves During and Across Sessions," *Presence*, vol. 9, no. 5, pp. 463–472, 2000.
[97] D. Risi and S. Palmisano, "Effects of postural stability, active control, exposure duration and repeated exposures on HMD induced cybersickness," *Displays*, vol. 60, pp. 9–17, 2019.
[98] S. Palmisano and R. Constable, "Reductions in sickness with repeated exposure to HMD-based virtual reality appear to be game-specific," *Virtual Reality*, vol. 26, no. 4, pp. 1373–1389, 2022.
[99] D. Zielasko, B. Rehling, B. von Dawans, and G. Domes, "Do not immerse and drive? prolonged effects of cybersickness on physiological stress markers and cognitive performance," *Virtual Reality*, 2026.
[100] B. Jones and M. G. Kenward, *Design and analysis of cross-over trials*. Chapman and Hall/CRC, 2015.
[101] D. Shi and T. Ye, "Behavioral carry-over effect and power consideration in crossover trials," *Biometrics*, vol. 80, no. 2, p. ujae023, Apr. 2024.
[102] J. Cohen, "Statistical power analysis," *Current directions in psychological science*, vol. 1, no. 3, pp. 98–101, 1992.
[103] J. W. Kelly, S. B. Gilbert, M. C. Dorneich, and K. A. Costabile, "Gender differences in cybersickness: Clarifying confusion and identifying paths forward," in *IEEE VRW*, 2023, pp. 283–288.
[104] K. M. Pöhlmann, G. Li, M. McGill, F. Pollick, and S. Brewster, "Can gender and motion sickness susceptibility predict cybersickness in vr?" in *IEEE VRW*, 2023, pp. 277–282.
[105] G. Gonçalves, M. Melo, and M. Bessa, "Virtual reality games: a study about the level of interaction vs. narrative and the gender in presence and cybersickness," in *Intl. conf. on graphics and interaction*, 2018, pp. 1–8.
[106] M. Melo, J. Vasconcelos-Raposo, and M. Bessa, "Presence and cybersickness in immersive content: Effects of content type, exposure time and gender," *Computers & Graphics*, vol. 71, pp. 159–165, 2018.
[107] P. Kourtesis, R. Amir, J. Linnell, F. Argelaguet, and S. E. MacPherson, "Cybersickness, cognition, & motor skills: The effects of music, gender, and gaming experience," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 5, pp. 2326–2336, 2023.
[108] C. Curry, R. Li, N. Peterson, and T. A. Stoffregen, "Cybersickness in virtual reality head-mounted displays: examining the influence of sex differences and vehicle control," *International Journal of Human–Computer Interaction*, vol. 36, no. 12, pp. 1161–1167, 2020.
[109] A. H. Park and S. Hu, "Gender differences in motion sickness history and susceptibility to optokinetic rotation-induced motion sickness." *Aviat. Space Environ. Med.*, vol. 70, no. 11, pp. 1077–1080, 1999.
[110] D. A. Graeber and K. M. Stanney, "Gender differences in visually induced motion sickness," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 46, no. 26. SAGE Publications, 2002, pp. 2109–2113.

Tongyu Nie is a Ph.D. candidate in Computer Science at the University of Minnesota. His research focuses on the intersection of augmented and virtual reality and human perception, with an emphasis on mitigating and detecting cybersickness.

Ville Cantory is a Ph.D. candidate at the University of Minnesota. His research focuses on perceptually driven computer graphics.

Courtney Hutton Pospick is a researcher at MediView XR who earned her Ph.D from the University of Minnesota and M.S. from the University of Southern California. Her research focuses on the intersection of 3D graphics, cognition, and perception, with an emphasis on 3D-interaction.

Danhua Zhang is a Ph.D. candidate at the University of Minnesota. She is interested in VR/AR user interfaces and application. Her research focuses on embodied virtual agents and 3D user interface design in VR.

Haoyu Tan is a Ph.D. candidate at the University of Minnesota. She is interested in AR/VR user interfaces, spatial interaction techniques, and perception. Her research focuses on AR depth perception and cybersickness mitigation in VR.

Sam Adeniyi is a Ph.D. candidate at the University of Minnesota. He is broadly interested in the intersections between AR/VR rehabilitation, exercise, and healthcare. His Ph.D. work focuses on the application of virtual reality for exercise and rehabilitation, exploring how immersive technologies can enhance recovery outcomes.

Fei Wu is a Software Development Engineer at AWS Annapurna Labs. She is interested in AR/VR, High Performance Computing and ML System. In 2022, she earned her Ph.D. from the University of Minnesota, where her research focused on mitigating cybersickness in VR.

Daniel Zielasko is an Associate Professor of Computer Science at the Technical University of Denmark (DTU). He previously worked as a postdoctoral researcher in the HCI group at University of Trier, Germany, and received his Ph.D. from RWTH Aachen University in 2020. His research centers on spatial user interaction and human perception in extended reality, with a focus on virtual locomotion, spatial orientation, and ergonomics, including cybersickness. He works across disciplines and aims to transfer fundamental insights into application domains such as biological and geological simulation, as well as health, well-being, and environmental contexts.

Isayas Berhe Adhanom is an Assistant Professor of Computer Science at Texas State University. His research focuses on improving the accessibility and user experience of virtual reality systems, with particular emphasis on understanding and mitigating factors that can impact user comfort and well-being. Isayas works at the intersection of Human-Computer Interaction, Cognitive Neuroscience, Computer Graphics, and Artificial Intelligence to develop user-friendly and inclusive next-generation immersive technologies. Isayas earned his Ph.D., in 2022, in Computer Science and Engineering from the University of Nevada-Reno and completed a postdoctoral fellowship at the University of Minnesota's Illusioniring Lab.

Evan Suma Rosenberg is an Associate Professor of Computer Science and Engineering at the University of Minnesota. Their research interests include virtual and augmented reality technologies, 3D user interfaces, and spatial interaction techniques, with a particular emphasis on locomotion and cybersickness. Dr. Suma Rosenberg received a Ph.D. in Computer Science at the University of North Carolina at Charlotte in 2010 and was a Research Assistant Professor at the University of Southern California until 2018. They have been an active participant in the IEEE VR community since 2004 and were elected to the IEEE VGTC Virtual Reality Academy in 2025.