This Lecture Makes Me Sick: On Confounding Factors Influencing the Simulator Sickness Questionnaire (SSQ)

Tim Weissker

Abstract—The Simulator Sickness Questionnaire (SSQ) has become a standard tool for quantifying the severity and distribution of discomfort symptoms in virtual reality (VR) research. Despite its straightforward administration, the use of the SSQ also comes with significant challenges, including response subjectivity, strict threshold values based on a military reference population, and confounding factors influencing the results. To demonstrate the adverse interplay of these issues, we asked three cohorts of students to fill in a SSQ after having attended a 90-minute lecture of our teaching program. Although students were not exposed to any form of VR experience, the resulting SSQ scores were indistinguishable from VR studies and extended far beyond the originally defined threshold of a “bad simulator”, with 88.1% of TS scores being larger and 25.4% even exceeding thrice this value. We compare our results to alternative scoring systems of the SSQ proposed in the literature and suggest implications for future experimental designs involving the quantification of sickness symptoms. In summary, our results motivate to exert caution when interpreting the results of the SSQ in the context of a VR study; participants might just have attended a lecture prior to the experiment.

Index Terms—Virtual Reality, Simulator Sickness, Discomfort, Cybersickness, Nausea

1 Introduction

Cybersickness [5, 28, 34], simulator sickness [15, 19], virtual reality sickness [12, 23], and visually-induced motion sickness [18, 21] all describe a range of unpleasant symptoms that may occur as negative side effects during or after the exposure to a virtual reality (VR) experience. While the exact definitions of these terms differ across publications, the scientific consensus is that the emergence of these symptoms is “one of the most significant usability issues in VR” [1], therefore limiting the widespread adoption of VR interfaces. As a result, the identification of factors that minimize sickness is a highly relevant research focus that has direct implications for a wide range of use cases. However, the accurate quantification of sickness in experiments still remains an open challenge, especially given that the resulting feeling of unpleasantness is *polysymptomatic* (i.e., it consists of many symptoms) and *polygeneic* (i.e., the manifestation of symptoms differs between individuals) [28].

Despite substantial efforts to derive objective measurements of sickness based on, for example, postural instability or physiology, the most prevalent measurement method remains the collection of subjective ratings in the form of questionnaires before, during, and/or after the experience [5, 12, 28]. The Simulator Sickness Questionnaire (SSQ), originally introduced by Kennedy et al. for military flight simulators in 1993 [19], has emerged as one of the most cited questionnaires to quantify the severity and distribution of sickness-related symptoms in VR research [3, 5, 12, 14, 28, 34]. It consists of 16 symptoms and asks participants to rate their perceived severity of these symptoms on a 4-point Likert scale from *0 (None)* to *3 (Severe)*. While the SSQ is an appealing measurement instrument due to its simplicity, standardized rating scale, and the resulting widespread use in the community, it is also often criticized for yielding suboptimal results and leading to potential misinterpretations (e.g., [3, 6, 12, 28]). Given the often abstract nature of these critiques, we were motivated by the overarching research question of how often-cited limitations of the SSQ affect the practical interpretation of its results and, as a first step, set out to quantify the influence of external factors unrelated to a VR experience on the scores of the SSQ. We report on the results of an experiment in which we asked three cohorts of students to fill in a SSQ after having attended one of our 90-minute lectures as part of our teaching program. Afterwards, we interpreted the obtained results in light of the original threshold values as introduced by Stanney et al. [32] and compared them to alternative SSQ scoring systems designed to mitigate specific issues identified in earlier work. In summary, our work led to the following contributions:

- The summary and comparison of four different SSQ scoring systems compatible with the original formulation of the questionnaire and its rating scale
- The demonstration that attending a lecture in a real-world auditorium can result in SSQ scores that are indistinguishable from VR studies in the literature, considerably higher than the original threshold describing a “bad simulator”, and even higher than reported after being exposed to the *Trier Social Stress Test* in a related publication [6]
- A summary of implications for future experimental designs involving the quantification of sickness symptoms

Our results motivate to exert particular caution when interpreting the results of the SSQ, since high symptom ratings may stem from external factors unrelated to the VR experience.

2 Related Work

The Simulator Sickness Questionnaire (SSQ) was originally developed by Kennedy et al. [19] in the context of military flight simulators and has been adopted as the most widely used measurement method to quantify sickness symptoms in modern virtual reality research [3, 5, 12, 14, 28, 34]. As of December 2025, the original paper counts more than 7700 citations according to Google Scholar. Participants are presented a list of 16 potential symptoms they might have experienced after using a particular system and asked to rate how much each of these symptoms affects them on a 4-point Likert scale from *0 (None)* to *3 (Severe)*. Based on a principal-factors analysis with varimax rotation, the authors derived a simple scoring system to compute a *Total Score (TS)* based on all symptoms as well as subscores representing *Nausea (N)*, *Oculomotor [Disturbance] (O)*, and *Disorientation (D)* based on subsets of symptoms.

2.1 Challenges of Using the SSQ

Despite its widespread use in VR research, the SSQ is not an optimal measurement tool to quantify sickness after a VR exposure. Bimberg et al. [3] summarized a range of pitfalls when working with the SSQ and highlighted differences in its administration and evaluation that restrict the comparability of studies in the literature. In the following, we summarize related work regarding three specific issues with the SSQ that are particularly relevant in the context of this paper.

- *Tim Weissker is with the Visual Computing Institute at RWTH Aachen University. E-mail: me@tim-weissker.de*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2.1.1 Response Subjectivity

As with all questionnaire-based methods, the symptom ratings on the SSQ are subjective and rely on an honest assessment and report by the participants, which is mentioned as a limitation in several publications [3, 5, 12, 14, 28]. In particular, the perceived severity of symptoms can differ between participants, and participants might even attribute the same sensation to different symptoms on the SSQ, as illustrated by the confusion between *dizziness* and *nausea* mentioned by Chang et al. [12]. Challenges regarding the interpretation of terminology are also reflected in the driving simulator studies of Chang et al. [11] and Curry et al. [13], who observed participants with high SSQ scores even though they had previously responded *No* to the question "Are you motion sick?". The work of Biocca [4] also underlined that participants might understate the severity of symptoms in an attempt to appear stronger. Finally, the work of Young et al. [36] found that symptom ratings after the exposure to a VR system were higher when the SSQ was already administered as a pre-exposure questionnaire, which is beneficial as a baseline measurement but also introduces demand characteristics. This finding is in line with the original SSQ publication by Kennedy et al. [19], who noted that difference scores have poor reliability and, therefore, advocated for a more general screening of healthy participants before an experiment.

2.1.2 Strict Threshold Values

Four years after the publication of the original questionnaire, Stanney et al. [32] performed a meta analysis of SSQ data obtained after the use of 2D flight simulators. Based on an analysis of about 9000 responses across more than 30 flight trainers, the authors derived threshold values to classify simulators based on the total score (TS). They concluded that values above 15 represent *concerning symptoms*, while values above 20 fall in their poorest category called a *bad simulator*. However, the participant samples used for the development as well as initial usage of the SSQ are often criticized for not being representative of the general population, given that they consisted of mostly male subjects from the military [3, 6, 7, 14, 28, 34]. Moreover, it appears questionable if data from early 2D flight simulators with small displays is suitable to form baseline values to which modern immersive VR experiences should be compared. Indeed, the reported SSQ scores in the literature across the years often seem to be in critical ranges when interpreted as originally intended. Starting with Stanney et al. [32] themselves in 1997, another meta analysis of college students experiencing "virtual environment (VE) systems" yielded an average TS score of 29 (min: 16, max: 55), which is claimed to be mostly due to technical differences rather than demographics. The meta analysis of Balk et al. [2] compared nine driving simulation studies from 2003 to 2012 and found mean total scores ranging from 8.62 to 27.25, with individual maxima up to 134.64. The HMD condition of a virtual factory visit in the seminal work of Sharples et al. [30] comparing different display systems in 2008 had an average total score of 29.92 ($\sigma = 21.59$). Across both validation studies of the *Fast Motion Sickness Scale* by Keshavarz and Hecht [20], the average total score was 41.05 ($\sigma = 33.43$). These and even higher values from early VR research can still be observed with modern HMD devices. The sickness testbed introduced by Calandra et al. [10] in 2024, for example, featured four different testing scenarios (a tower defense game, a waypoint navigation task, a track race, and a roller coaster ride), in which total scores ranged from 13.86 to 39.60 with a field-of-view restrictor and from 50.16 to 95.48 without any form of mitigation technique.

2.1.3 Confounding Factors

Given the generality of the symptoms probed on the SSQ, it is plausible that they may also arise as a result of other factors than the VR experience under investigation. For example, symptoms like *General discomfort* or *Fullness of head* represent complex cognitive conditions that can be easily affected by time of day, activities before the experiment, and general psychological state. To underline the influence of confounding factors, the review of Rebenitsch and Owen [28] mentioned that even closing one's eyes for a longer time can already register as increases in SSQ scores. Similarly, the work of Smart et al. [31] observed individual participants with high SSQ ratings after merely focusing on a static visual target in a real-world room. The study of Bouchard et al. [6] employing an in-person variant of the *Trier Social Stress Test* highlighted that many symptoms on the SSQ can also be induced by exposing participants to a stressful conversation scenario in real life. The authors furthermore demonstrated a strong correlation between SSQ scores and participants' anxiety, indicating another potential confounding factor in VR experiments. Similarly, the work of Grassini et al. [17] found that the nausea subscale of the SSQ correlated with participants' neuroticism, which reveals potential influences of general personality traits as well.

2.1.4 Summary

Taken together, these results motivate taking a closer look when interpreting the SSQ and considering alternative factors that might have contributed to a particular result. Motivated by the work of Bouchard et al. [6] indicating that SSQ symptoms may be invoked by stressful situations in real life, our work in this paper goes one step further and demonstrates that even non-stressful everyday situations in the life of a student, such as attending a lecture, can result in strong responses on the SSQ.

2.2 Alternative Scoring Methods of the SSQ

In addition to the military reference population used for their derivation, another common point of criticism regarding the scoring formulas of the SSQ relates to the fact that five symptoms contribute to more than one of the subscores [2, 3, 6, 7]. For example, the symptom *Nausea* not only contributes to the Nausea (N) but also to the Disorientation (D) subscore and, therefore, affects the total score twice as much [3, 19]. To address these issues, several researchers have proposed alternative methods for scoring the SSQ. The aforementioned work on driving simulator data by Balk et al. [2], for example, proposed a revised scoring model for the three subscores N, O, and D that associates each symptom with only one subscore and excludes the five symptoms *Burping*, *Fatigue*, *Headache*, *Blurred vision*, and *Fullness of head*. However, several inconsistencies emerged when including the total score as well as participants' dropout rate into the model, leading the authors to conclude that focusing on individual symptom scores may be more insightful than overall SSQ scores. The study of Bouchard et al. in 2007 [7] included 44% of participants suffering from an anxiety disorder and revealed a two-factor model in which each symptom contributed to either a *Nausea* or an *Oculomotor* score exclusively. This two-factor structure with exclusive symptom contributions has been confirmed with a new sample in 2021 [6]. Stone III [33] and Kim et al. [23] each introduced shorter versions of the SSQ called the *Cybersickness Questionnaire (CSQ)* and the *Virtual Reality Sickness Questionnaire (VRSQ)*, respectively. Both publications selected different subsets of nine symptoms and each proposed a two-factor model with exclusive symptom contributions based on the participants' ratings. In case of the CSQ, the two factors are labeled *Dizziness* and *Difficulty focusing*, which we abbreviate by *DZ* and *DF* in this paper. For the VRSQ, the authors make use of the original terms *Oculomotor* and *Disorientation*. In an empirical study by Sevinc and Berkman [29] comparing different scoring approaches of the SSQ, the results indicated that both the CSQ [33] and the VRSQ [23] demonstrated better psychometric qualities for assessing modern VR applications compared to the original SSQ scoring as well as the modified scoring by Bouchard et al. [7].

Taken together, these results indicate that alternative scoring approaches might be beneficial over the original SSQ formulas as introduced by Kennedy et al. [19]. Inspired by the work of Sevinc and Berkman [29], we will therefore compare different scoring systems when interpreting the results of our experiment in this paper. Following their analysis, we will focus on the original SSQ scoring [19], the modified scoring by Bouchard et al. [7], the CSQ [33], and the VRSQ [23] as all of these can be directly computed based on the original test materials without requiring additional data or alternative rating scales. The scoring procedure of these four scales is summarized in Table 1.

Table 1: The 16 symptoms of the SSQ to be rated on a Likert scale from 0 to 3 as well as the different scoring approaches focused on in this paper, namely the original *SSQ* scoring by Kennedy et al. [19], the modified scoring by Bouchard et al. [7] labeled SSQ_B, the CSQ by Stone III [33], and the VRSQ by Kim et al. [23]. This table is similar to the overview provided by Sevinc and Berkman [29], but it corrects two mistakes regarding the scoring of *Difficulty concentrating* on the original SSQ and the scoring of *Burping* on the SSQ_B. The letters N, O, and D refer to the original cluster names *Nausea*, *Oculomotor*, and *Disorientation*, respectively. The subscores *Dizziness* and *Difficulty focusing* of the CSQ are abbreviated by DZ and DF, respectively.

	SSQ [19]			SSQ_B* [7]		*CSQ*** [33]		*VRSQ [23]*	
	N	O	D	N	O	DZ	DF	O	D
General discomfort	1	1		1				1	
Fatigue		1			1			1	
Headache		1			1	0.50			1
Eyestrain		1			1		0.58	1	
Difficulty focusing		1	1		1		0.89	1	
Increased salivation	1			1					
Sweating	1			1					
Nausea	1		1	1		0.84			
Difficulty concentrating	1	1			1				
Fullness of head			1		1		0.55		1
Blurred vision		1	1		1		0.81		1
Dizzy (eyes open)			1	1		0.89			
Dizzy (eyes closed)			1	1		0.99			1
Vertigo			1	1		0.54			1
Stomach awareness	1			1					
Burping	1			1					
(Weighted) Sum	[A]	[B]	[C]	[D]	[E]	[F]	[G]	[H]	[I]
Subscore Formulas	N = [A] · 9.54 O = [B] · 7.58 D = [C] · 13.92			N = [D] O = [E]		DZ = [F] DF = [G]		O = ([H]/12) · 100 D = ([I]/15) · 100	
Total Score Formula	TS = ([A]+[B]+[C]) · 3.74			TS = [D]+[E]		*None*		TS = (O + D)/2	

*Bouchard et al. [7] did not provide exact scoring rules based on their two-factor model, leaving some ambiguities. While Sevinc and Berkman [29] stated that the scores [D] and [E] are likely arithmetic means of the individual symptom ratings, we argue that the total score range from 0 to 42 as reported by Bouchard et al. [7] can only be achieved when summing up the symptom scores instead of taking their mean. This process is confirmed by a later publication of Bouchard et al. in 2021 [6].
**The scoring of the CSQ requires clamping the raw symptom ratings at a value of *2 (Moderate)*, resulting in all responses of *3 (Severe)* being replaced by *2 (Moderate)*.

3 Experiment: The SSQ After Attending a Lecture

Based on our overview of related work, it appears that the clear interpretation of SSQ results is complicated by the subjectivity of responses, strict threshold values, and potential confounding factors. As the work of Bouchard et al. [6] demonstrated that stressful real-life situations can register on the SSQ without making use of any form of immersive hardware, we thought about less stressful scenarios that might, nonetheless, also evoke some of the symptoms probed on the SSQ. Thinking about personal experiences in the life as a student, we supposed that even regular everyday activities like attending a lecture might already produce measurable responses on the SSQ, as the conventional university learning format of listening to a complex 90-minute presentation demands several cognitive resources for a prolonged time. Being now in the position of the lecturer, we set out to test this empirically and obtain measurements to be compared with established results of the SSQ in the literature.

3.1 Experimental Procedure

Our teaching program includes an annual master's course on virtual reality, which teaches advanced and novel concepts in VR research in an attempt to prepare students for a master's thesis or PhD in the field. The course is designed in a modular fashion and covers a large variety of topics on the design and evaluation of immersive interactive systems. In each lecture, the lecturer presents a projected slide show that mostly contains text and images related to a specific area of VR research. In the iterations of the course from 2023 to 2025, we each selected an individual 90-minute lecture of the course and, at the end, asked students to volunteer to fill in the SSQ by stating to be interested in evaluating potential negative side effects that our teaching induces. Students explicitly gave their consent to the processing of their data for research purposes on the form. The SSQ had not been introduced as part of the course before its administration, so it is unlikely that participants attempted to manipulate its results. To prevent any form of demand characteristics as highlighted by Young et al. [36], we did not perform a baseline measurement at the beginning of the lecture. In 2023, 2024, and 2025, we selected the lectures on *Collaborative VR*, *Acoustics*, and *Experimental Design*, respectively, in order to rule out potential effects related to only a single topic of the course. Students were debriefed in a later lecture on *Cybersickness*, where we formally introduced the SSQ and presented descriptive results of our experiment in that year. Given that our course offers online video recordings to increase its outreach and accessibility, we only processed the responses from on-site attendees to eliminate potential adverse effects introduced by distance learning.

Based on an internal checklist provided by the Interfaculty Ethics Committee of RWTH Aachen University, no explicit ethical approval was required for this procedure. In particular, the evaluation was integrated into the everyday teaching program at the faculty, with the only difference to normal conditions being that the SSQ was added to the regular teaching evaluation form. This modification did not elicit strong emotions or stress, induce pain or harm, conceal risks, capture biological material, or pose security-relevant threats.

3.2 Participants

Across the three years, we collected a total of 59 SSQ responses from on-site attendees of our lectures. In 2023, the sample consisted of 23 participants (13 male, 7 female, 2 other, 1 unknown) with a mean age of 24.74 ($\sigma = 2.98$). In 2024, the sample was slightly smaller, consisting of 14 participants (11 male, 3 female) with a mean age of 26.79 ($\sigma = 9.41$). In 2025, a total of 22 participants (14 male, 7 female, 1 unknown) with a mean age of 25.50 ($\sigma = 2.31$) completed the questionnaire. As a consequence of these demographics, it is important

Table 2: Descriptive parameters of all 59 SSQ responses from 2023 to 2025, out of which the scores based on the original SSQ scoring formulas are visualized as boxplots in Figure 1. For each score, the table lists the theoretically possible maximum value (Max_T) as well as the observed mean (M), standard deviation (σ), 25%, 50%, and 75% percentile values (P_i), and maximum (Max). The minimum observed value for all scores was 0.00.

Score		Max_T	M	σ	P_{25}	P_{50}	P_{75}	Max
SSQ	N	200.34	29.42	19.43	19.08	28.62	38.16	95.40
	O	159.18	47.79	28.61	26.53	45.48	68.22	136.44
	D	292.32	45.06	33.04	27.84	41.76	55.68	125.28
	TS	235.62	47.23	28.51	31.79	41.14	61.71	138.38
SSQ_B	N	27.00	1.88	1.97	0.00	1.00	3.00	7.00
	O	21.00	6.61	4.08	4.00	6.00	9.00	19.00
	TS	48.00	8.49	5.41	5.00	8.00	12.00	26.00
CSQ	DZ	7.52	0.62	1.00	0.00	0.00	0.94	3.31
	DF	5.66	2.27	1.42	1.13	2.33	3.46	5.66
VRSQ	O	100.00	32.63	19.59	16.67	33.33	41.67	83.33
	D	100.00	12.88	13.09	0.00	6.67	20.00	53.33
	TS	100.00	22.75	14.74	12.50	20.00	30.84	68.33

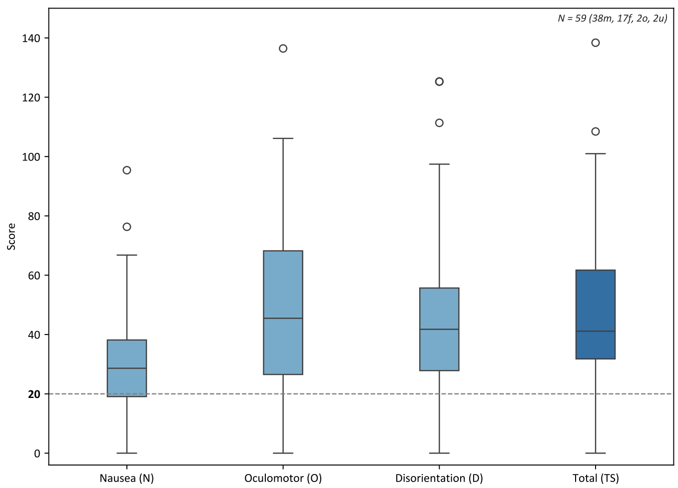

Fig. 1: Distribution of scores representing nausea (N), oculomotor (O), disorientation (D), and total (TS) based on all 59 SSQ responses from 2023 to 2025, computed with the original formulas introduced by Kennedy et al. [19]. A score of more than 20 is considered a "bad simulator" according to Stanney et al. [32].

to note that our overall sample across the three years is skewed in terms of gender distribution (64.4% male participants), which we attribute to the similarly skewed demographics of enrolled students at our faculty.

3.3 Results

Table 2 provides a descriptive overview of the resulting scores representing all 59 responses when applying the original scoring formulas by Kennedy et al. [19], the modified scoring system by Bouchard et al. [7], the CSQ formulas [33], and the VRSQ formulas [23]. Figure 1 presents boxplots illustrating the distribution of the original SSQ scores *Nausea (N)*, *Oculomotor (O)*, *Disorientation (D)*, and *Total (TS)* in relation to the threshold value of 20 used to classify a "bad simulator" by Stanney et al. [32]. To provide more context for the presented aggregate scores, Table 3 provides a sorted list of the individual symptoms based on their mean response score from *0 (None)* to *3 (Severe)*.

3.4 Discussion

Our results demonstrate that merely attending a 90-minute lecture as a student can already produce a variety of symptoms that are also probed on the SSQ. The three most prominent symptoms, all with mean scores

Table 3: Symptoms sorted by mean score (M) across all 59 responses; standard deviations are denoted by σ. The four rightmost columns list the subscore(s) that each symptom contributes to.

Sorted Symptoms	M	σ	*SSQ*	SSQ_R	*CSQ*	*VRSQ*
Difficulty concentrating	1.627	0.919	N, O	O	-	-
Difficulty focusing	1.356	1.021	O, D	O	DF	O
Fatigue	1.220	0.958	O	O	-	O
Fullness of head	0.949	0.910	D	O	DF	D
Eyestrain	0.695	0.670	O	O	DF	O
General discomfort	0.644	0.683	N, O	N	-	O
Blurred vision	0.390	0.713	O, D	O	DF	D
Headache	0.373	0.734	O	O	DZ	D
Increased salivation	0.288	0.554	N	N	-	-
Stomach awareness	0.237	0.499	N	N	-	-
Dizzy (eyes open)	0.203	0.479	D	N	DZ	-
Dizzy (eyes closed)	0.136	0.342	D	N	DZ	D
Nausea	0.119	0.454	N, D	N	DZ	-
Sweating	0.119	0.372	N	N	-	-
Vertigo	0.085	0.279	D	N	DZ	D
Burping	0.051	0.220	N	N	-	-

N: Nausea, O: Oculomotor, D: Disorientation, DF: Difficulty focusing, DZ: Dizziness, -: Excluded

larger than *1 (Slight)*, were *Difficulty concentrating*, *Difficulty focusing*, and *Fatigue*. While it is a theoretical option that individual students were indeed sick from having to focus on a static target for a prolonged time as in the work of Smart et al. [31], the top symptoms also describe complex general conditions that are highly plausible to arise as a result of general strain. The three least-experienced symptoms on the list were *Nausea/Sweating* (same mean), *Vertigo*, and *Burping*, which all represent issues that appear to be defined more narrowly and are less likely to emerge from being stationary in an auditorium without being exposed to any form of motion cues. In the following, we discuss how our measured symptom ratings influenced the different aggregate scores of the SSQ suggested in the literature.

3.4.1 Original SSQ Scoring System

When analyzing our data using the original scoring formulas by Kennedy et al. [19], we observed a mean total score of 47.23 ($\sigma = 28.51$), which is more than twice of the threshold value characterizing a "bad simulator" introduced by Stanney et al. [32]. To underline this result, a one-sample t-test indicated that the mean total score is significantly larger than the threshold value of 20 with a large effect size, $t(58) = 7.336, p < 0.001, d = 0.955$. In particular, 52 TS scores (88.1%) were above this threshold, while 15 (25.4%) of them even exceeded a value of 60 – three times the original threshold. Making use of the verbal interpretation of Stanney et al. [32] stating that "individuals who report symptoms which tally higher than 15 on the SSQ, would be experiencing sufficient discomfort that, unless there were other incentives, they would not voluntarily seek additional exposure because of the level of discomfort", our lectures would be considered a severe source of malaise that is unethical to present to human subjects. Looking at the subscores, it appears that the total score is most strongly driven by oculomotor- and disorientation-related symptoms, even though the smaller nausea scores are still mostly above the threshold of 20 ($M = 29.42$, $\sigma = 19.43$). A paired-samples t-test confirmed that the nausea scores were significantly smaller than both the oculomotor ($t(58) = 7.83$, $p < 0.001$, $d = 1.019$, large effect) and the disorientation scores ($t(58) = 5.17, p < 0.001, d = 0.673$, medium effect). However, these significance tests have to be treated with caution, as they operate on different maxima of the subscales (see Table 2).

As introduced in Section 2.1, SSQ scores far beyond the originally defined thresholds are commonly observed in the VR research literature, which is likely due to the military reference population and the rudimentary 2D display hardware that were used for the initial experiments with the questionnaire. Nonetheless, we were surprised to have obtained scores that are indistinguishable from data observed after the exposure to VR systems over the years (see Section 2.1.2), given that our experimental setup did not involve any form of immersive hardware and that the digital content was limited to static text and images in a slide show.

3.4.2 Alternative SSQ Scoring Systems

Given that the modified scoring system by Bouchard et al. [7] also makes use of the full original symptom list, the TS score of the SSQ_B system is similarly affected by the measured symptom ratings as the TS score in original scoring system. To provide an additional anchor for putting our results into context, we also compared our data to the in-person *Trier Social Stress Test* study in a later publication by Bouchard et al. [6], who also reported their results using the SSQ_B scoring system. We found that the mean SSQ_B total score after attending our lectures ($M = 8.49$ out of 48, 17.7%) was above the reported mean total score after the deliberately stress-inducing interview scenario ($M = 5.01$ out of 48, 10.4%). An independent samples t-test based on Bouchard et al.'s reported mean ($M = 5.01$), standard deviation ($\sigma = 4.51$), and sample size ($N = 91$) indicated that their ratings of the stress test were significantly lower than the ratings of our lecture with a medium effect size, $t(148) = 4.264, p < 0.001, d = 0.699$. This further demonstrates a notable impact of our lecture on the symptom ratings. Looking at the subscores, it appears that the high scores after our lectures were mostly due to oculomotor-related ($M = 6.61$ out of 21, 31.55%) as opposed to nausea-related symptoms ($M = 1.88$ out of 27, 7.0%), which was also the case with the original SSQ scoring system. This difference was also statistically significant ($t(58) = 10.6$, $p < 0.001$, $d = 1.38$, large effect), even though the nausea subscale has a higher maximum (see Table 2).

Both the CSQ and the VRSQ scoring systems make use of only a subset of the original symptom list, offering novel opportunities to disregard factors that might not contribute to a clear interpretation of the results. Regarding the CSQ system, attending our lectures strongly affected the *Difficulty focusing (DF)* subscore, which turned out to be the only scale in our experiment for which the entire range of possible scores from 0.00 to 5.66 was observed ($M = 2.27$ out of 5.66, 40.1%). This is likely due to the fact that the CSQ combines the symptom ratings *2 (Moderate)* and *3 (Severe)* into the same maximum rating (see Table 1). The subscore representing *Dizziness (DZ)*, on the other hand, was far less affected by attending our lectures ($M = 0.62$ of 7.52, 8.2%), which is also reflected by the fact that all symptoms contributing to DF were rated higher on average than all symptoms contributing to DZ (see Table 3). A paired-samples t-test confirmed that this difference was statistically significant ($t(58) = -9.29$, $p < 0.001$, $d = -1.21$, large effect), even though the DZ score has a higher theoretical maximum. As the only scoring system in our analysis, the CSQ explicitly argues against a joint total score representing overall sickness, leaving us with DZ and DF as the only indicators of participant wellbeing.

Based on the VRSQ scoring system, oculomotor-related symptoms ($M = 32.63$ out of 100) scored higher than disorientation-related symptoms ($M = 12.88$ out of 100), which was also observed for the original SSQ scoring. This difference was also statistically significant ($t(58) = 9.76$, $p < 0.001$, $d = 1.27$, large effect). Compared to the CSQ, however, fewer of the top-rated symptoms after our lecture were eliminated in the scoring procedure (see Table 3). In particular, the VRSQ only eliminated one of the eight most frequently observed symptoms (*Difficulty concentrating*, rank 1), while the CSQ eliminated three of them (*Difficulty concentrating*, *Fatigue*, *General discomfort*, ranks 1, 3, and 6). As a result, the effects of attending our lecture influenced the VRSQ scoring more strongly than the CSQ scoring.

Overall, it appears that none of the alternative scoring approaches could fully eliminate the influences of external factors unrelated to the use of immersive VR systems. Nonetheless, similarly to the *Nausea (N)* score of the original system, all alternatives revealed a subscore that was significantly less affected by attending our lectures than the others with large effect sizes. These were the *Nausea (N)* score of the SSQ_B, the *Dizziness (DZ)* score of the CSQ, and the *Disorientation (D)* score of the VRSQ. While these subscores might be interesting candidates to study in future work, we explicitly argue against narrowing the scope of the SSQ without further scientific validation. Based on our data of this paper, there appears to be a slight advantage of the CSQ system in separating the higher- from the lower-rated symptoms experienced after our lectures, but further validation is required as well to manifest and generalize this initial explorative finding.

3.4.3 Limitations

While we believe that there are many potential confounding factors that might lead to high scores on the SSQ, our experiment presented in this paper only focused on the one fixed scenario of attending a lecture as part of a university course. As a consequence, analyzing the generalizability of our results to other scenarios is still subject to future work. It has to be noted that our lectures' duration of 90 minutes is considerably longer than many VR sessions of experiments in the research literature, which could be a potential confounding variable. Moreover, it cannot be excluded that the observed SSQ scores might stem from activities before the lecture, as prior literature surveys have shown that symptoms typically last for 1 to 2 hours but can persist for days in extreme cases [37]. Furthermore, we obtained data from students voluntarily deciding to attend our course, which resulted in a young and gender-skewed sample reflecting the enrollment statistics of our faculty. As a consequence, the results might be unrepresentative of the general adult population. Finally, as we administered the SSQ after three different lectures across three years, it was not possible to ensure perfectly identical conditions for all participants as commonly observed in empirical user studies. While these limitations have to be taken into consideration when interpreting our data, we believe that similar results may arise in other lectures at other universities. To confirm this belief, we strongly encourage other scholars in the field to replicate our experiment by asking students to fill in the SSQ after one of their lectures.

4 Implications

Given that attending our lectures resulted in high total SSQ scores indistinguishable from VR experiences as well as total SSQ_B scores that were higher than being exposed to the *Trier Social Stress Test* as conducted by Bouchard et al. [6], the most straightforward implication is that our teaching performance is exceptionally poor and ethically questionable. However, the latest official university-wide teaching evaluation in 2025, completed by the same cohort of students participating in our experiment, resulted in an average grade of 1.3 ($\sigma = 0.5$) for the concept of our lectures and 1.2 ($\sigma = 0.5$) for the behavior of the lecturer on a rating scale from 1.0 (best) to 5.0 (worst). As a result, we tend to believe that our course is generally appreciated and not considered a major source of discomfort that students "would not voluntarily seek additional exposure" [32] to. Instead, we believe that our results demonstrate that university learning can be a cognitively demanding task that can result in several complex conditions that exhibit symptoms also present on the SSQ. As a result, we advise being cautious when interpreting high SSQ scores in the context of a VR study as participants (who are often students in VR research [38]) might have attended a lecture prior to joining the study. Based on our insights, the following sections summarize a set of suggestions for VR research and beyond.

4.1 Consider Pre- and Post-Measurements

Although Kennedy et al. [19] as well as Young et al. [36] argue against the administration of the SSQ prior to the VR exposure due to the poor reliability of difference scores and the potential of introducing demand characteristics, several VR researchers have adopted this practice nonetheless to put high post-exposure SSQ scores into context. The meta-analysis of IEEE VR papers from 2015 to 2019 by Peck et al. [26], for example, analyzed 21 manuscripts with pre- and post-SSQ measurements, with the actual number likely being higher as the studied papers also had to meet other criteria. The survey on the usage of the SSQ in VR research by Bimberg et al. [3] also argues in favor of pre-exposure measurements to interpret the SSQ more meaningfully. Potentially, such a procedure could also detect our observed effects of having attended a lecture prior to participating in a VR experiment, assuming that sickness-related symptoms emerge rapidly during or directly after the lecture. However, prior research on cybersickness identified a large variety of potential emergence patterns, including reports of delayed symptoms or symptoms resurfacing after an initial improvement [37]. In these cases, a pre-exposure measurement might not be able to appropriately quantify a user's state before entering a VR environment. We encourage researchers to consider the benefits and drawbacks of pre- and post-measurements and communicate their design decisions explicitly in a publication. If pre-measurements are obtained, the investigation of difference scores can lead to situations in which the SSQ score after the exposure is lower than before, as Bimberg et al. [3] also note. In these cases, the authors advised that the observed improvement should not be attributed to the experimental condition by stating the there was *no negative* instead of a *positive* effect on participants.

4.2 Keep User Studies Concise

While we have argued that lengthy and cognitively-demanding activities before a VR experiment might influence the SSQ results of that experiment, our results also motivate reducing the length and complexity of the experiment itself to increase the interpretability of the SSQ. Based on our lecture results, it can also be expected that the execution of other high-level cognitive tasks (e.g., involving memory or spatial orientation skills) and even an abundance of repetitions of simple tasks are likely to also register on the SSQ. As a result, high SSQ scores might also occur when executing these tasks outside of VR, and it is difficult to isolate different potential contributors after a VR study has been performed. While we acknowledge that it might not always be possible, we encourage researchers to explore ways of performing baseline measurements in the real world and/or streamlining their study designs, for example, by focusing on fewer research questions or hypotheses per study or by employing between-subject designs.

4.3 Consider Alternative Measurement Instruments

Based on the popularity of the SSQ as well as the importance of reducing sickness in all kinds of VR systems, researchers have proposed several alternative measurement instruments to quantify a user's wellbeing. While there is great research interest in estimating a user's wellbeing objectively based on physiological and behavioral measures [5, 12, 34], none of these methods has yet emerged as a widely adopted standard in VR research. Instead, questionnaire-based evaluations are still of high interest due to their simplicity and ease of administration. To improve the expressivity of the SSQ, researchers have proposed extending the original version to include additional dimensions [22] or severity rating steps [25], as well as combining the subjective symptom ratings with more objective physiological measures [8], to name just three examples. Other researchers suggested that symptom ratings of any kind may not be the best way to distinguish people with and without sickness. Therefore, one suggestion is to move away from specific symptom ratings to considerably simpler forms of assessing one's current overall wellbeing with just a single question. This might be done by asking participants about the presence or absence of sickness as a two-alternative forced choice question [11, 13] or by using more specific formulations with more nuanced rating scales. For example, Rebenitsch and Owen suggested asking participants to rate "where they are now" on a scale from 0 (how they felt coming in) to 10 (they want to stop) [27], which has been used repeatedly in the VR literature since (e.g., [9, 16, 24, 35]). However, our personal observations with this scale across studies revealed that the general formulation of this question often leads to participants rating their experience with the tested interaction techniques rather than their personal wellbeing, which makes us believe that our lecture experiment presented in this paper could also strongly register on this scale. The Fast Motion Sickness Scale (FMS) by Keshavarz and Hecht [20], instead, explicitly asks participants to rate the level of *sickness* from 0 (no sickness) to 20 (frank/severe sickness) and, consequently, has been shown to most strongly correlate with the nausea subscale of the SSQ ($r = 0.828$). As this was the least affected subscale in our lecture experiment, we believe that attending our lectures would have registered less strongly on the FMS than it did on the SSQ. However, it appears that a fully isolated measurement of sickness resulting from a VR exposure is still an open research challenge.

4.4 Reconsider University Teaching Formats

Reducing all sickness-related symptoms has been a major focus of research since the emergence of VR systems, with the SSQ presenting a widely-adopted list of symptoms that altogether describes an unpleasant experience that should be avoided as much as possible when designing VR applications. As our experiment in this paper has shown, a wide range of the very same symptoms were reported by students after attending a single 90-minute lecture in their semester. Given the large number of lectures that students attend to pass a single course, and the large number of courses required to pass their degree program, it appears fruitful to reconsider if the conventional frontal teaching format is the best way to move forward in university education. While we acknowledge that computer science is a complex discipline that requires cognitive resources to study thoroughly, our results have motivated us to integrate more interactive elements into our lectures, including group work, dynamic quizzes, and guided discussions. We believe that these elements provide additional stimulations that have the potential to reduce the unpleasantness of having to passively listen to a 90-minute presentation, and we encourage the community to become creative regarding novel and less strenuous teaching formats.

5 Conclusion and Future Work

While our experiment presented in this paper started as a playful activity to educate our students about potential confounding influences on the SSQ, the results were much more pronounced than we had initially expected. In particular, we conclude that the SSQ data obtained from attending our lectures cannot be distinguished from data commonly observed after VR experiments, which demonstrates that the accurate and isolated measurement of adverse symptoms introduced by a VR experience is not guaranteed by the SSQ. Based on our observations, we suggest considering pre-exposure measurements to provide more context for the observed results, encourage keeping study designs concise, motivate the consideration of alternative measurement instruments like the FMS [20], and raise general concerns regarding the conventional university teaching formats. All in all, we encourage the community to explore the reasons behind an observed effect instead of solely relying on numeric statistical insights to form conclusions.

Future work on the accurate quantification of sickness in VR is vast, with the most appealing vision being an objective descriptor of sickness purely based on physiological and/or behavioral measurements. However, given the complexity of sickness and sickness-related symptoms, it is unclear if this vision can be fully achieved in the future. While we cannot monitor physiological or behavioral data of the students in our lecture due to ethical reasons, we intend to explore alternative sickness-related questionnaires over the next years to analyze how their results relate to our SSQ data presented in this paper. Furthermore, we are interested in gathering the SSQ scores from different lectures given by the research community across the globe, potentially revealing the effects of region, number of students, or lecture topics on the emergence of unpleasant symptoms by students.

Acknowledgments

I would like to thank *Pauline Bimberg* for coming up with the random idea of checking my students' level of sickness-related symptoms during my lectures, "just to be sure". This work has received funding from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under grant 528403131 (Project "Put me There").

References

[1] I. B. Adhanom, D. Zielasko, R. Islam, G. Kim, and E. S. Rosenberg. The Sixth Workshop on Immersive Sickness Prevention. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 813–814. IEEE, 2025. doi: 10.1109/VRW66409.2025.00166 1

[2] S. A. Balk, D. B. Bertola, and V. W. Inman. Simulator Sickness Questionnaire: Twenty Years Later. In *Driving Assessment Conference*, vol. 7,

pp. 257–263. University of Iowa, 2013. doi: 10.17077/drivingassessment.1498 2
[3] P. Bimberg, T. Weissker, and A. Kulik. On the Usage of the Simulator Sickness Questionnaire for Virtual Reality Research. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 464–467, 2020. doi: 10.1109/VRW50115.2020.00098 1, 2, 5, 6
[4] F. Biocca. Will Simulation Sickness Slow Down the Diffusion of Virtual Environment Technology? *Presence: Teleoperators and Virtual Environments*, 1(3):334–343, 1992. doi: 10.1162/pres.1992.1.3.334 2
[5] N. Biswas, A. Mukherjee, and S. Bhattacharya. "Are you feeling sick?" – A systematic literature review of cybersickness in virtual reality. *ACM Computing Surveys*, 56(11), 2024. doi: 10.1145/3670008 1, 2, 6
[6] S. Bouchard, M. Berthiaume, G. Robillard, H. Forget, C. Daudelin-Peltier, P. Renaud, C. Blais, and D. Fiset. Arguing in Favor of Revising the Simulator Sickness Questionnaire Factor Structure When Assessing Side Effects Induced by Immersions in Virtual Reality. *Frontiers in Psychiatry*, 12, 2021. doi: 10.3389/fpsyt.2021.739742 1, 2, 3, 5
[7] S. Bouchard, G. Robillard, and P. Renaud. Revising the factor structure of the Simulator Sickness Questionnaire. *Annual Review of CyberTherapy and Telemedicine*, 5:117–122, 2007. 2, 3, 4, 5
[8] S. Bruck and P. A. Watters. The factor structure of cybersickness. *Displays*, 32(4):153–158, 2011. doi: 10.1016/j.displa.2011.07.002 6
[9] H. Brument, M. Marchal, A.-H. Olivier, and F. Argelaguet Sanz. Studying the Influence of Translational and Rotational Motion on the Perception of Rotation Gains in Virtual Environments. In *Proceedings of the 2021 ACM Symposium on Spatial User Interaction*, SUI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3485279.3485282 6
[10] D. Calandra and F. Lamberti. A Testbed for Studying Cybersickness and its Mitigation in Immersive Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(12):7788–7805, 2024. doi: 10.1109/TVCG.2024.3448203 2
[11] C.-H. Chang, F.-C. Chen, W.-C. Kung, and T. A. Stoffregen. Effects of Physical Driving Experience on Body Movement and Motion Sickness During Virtual Driving. *Aerospace Medicine and Human Performance*, 88(11):985–992, 2017. doi: 10.3357/AMHP.4893.2017 2, 6
[12] E. Chang, H. T. Kim, and B. Yoo. Virtual Reality Sickness: A Review of Causes and Measurements. *International Journal of Human–Computer Interaction*, 36(17):1658–1682, 2020. doi: 10.1080/10447318.2020.1778351 1, 2, 6
[13] 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*, 36(12):1161–1167, 2020. doi: 10.1080/10447318.2020.1726108 2, 6
[14] S. Davis, K. Nesbitt, and E. Nalivaiko. A Systematic Review of Cybersickness. In *Proceedings of the 2014 Conference on Interactive Entertainment*, IE2014. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2677758.2677780 1, 2
[15] N. Dużmańska, P. Strojny, and A. Strojny. Can Simulator Sickness Be Avoided? A Review on Temporal Aspects of Simulator Sickness. *Frontiers in Psychology*, 9, 2018. doi: 10.3389/fpsyg.2018.02132 1
[16] 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)*, pp. 201–210, 2016. doi: 10.1109/3DUI.2016.7460053 6
[17] S. Grassini, K. Laumann, and A. K. Luzi. Association of Individual Factors with Simulator Sickness and Sense of Presence in Virtual Reality Mediated by Head-Mounted Displays (HMDs). *Multimodal Technologies and Interaction*, 5(3), 2021. doi: 10.3390/mti5030007 2
[18] R. S. Kennedy, J. Drexler, and R. C. Kennedy. Research in visually induced motion sickness. *Applied Ergonomics*, 41(4):494–503, 2010. doi: 10.1016/j.apergo.2009.11.006 1
[19] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, 1993. doi: 10.1207/s15327108ijap0303_3 1, 2, 3, 4, 5
[20] B. Keshavarz and H. Hecht. Validating an Efficient Method to Quantify Motion Sickness. *Human Factors*, 53(4):415–426, 2011. doi: 10.1177/0018720811403736 2, 6
[21] B. Keshavarz, B. E. Riecke, L. J. Hettinger, and J. L. Campos. Vection and visually induced motion sickness: how are they related? *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.00472 1
[22] D. Kim, D. Parker, and M. Park. A New Procedure for Measuring Simulator Sickness – The RSSQ. *Human Interface Technology Laboratory, University of Washington, Seattle, WA*, 2004. 6
[23] H. K. Kim, J. Park, Y. Choi, and M. Choe. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics*, 69:66–73, 2018. doi: 10.1016/j.apergo.2017.12.016 1, 2, 3, 4
[24] M. Krüger, T. Gerrits, T. Römer, T. Kuhlen, and T. Weissker. IntenSelect+: Enhancing Score-Based Selection in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2829–2838, 2024. doi: 10.1109/TVCG.2024.3372077 6
[25] S. Nichols, S. Cobb, and J. R. Wilson. Health and Safety Implications of Virtual Environments: Measurement Issues. *Presence: Teleoperators and Virtual Environments*, 6(6):667–675, 1997. doi: 10.1162/pres.1997.6.6.667 6
[26] 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 Transactions on Visualization and Computer Graphics*, 26(5):1945–1954, 2020. doi: 10.1109/TVCG.2020.2973498 5
[27] L. Rebenitsch and C. Owen. Individual variation in susceptibility to cybersickness. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, p. 309–317. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2642918.2647394 6
[28] L. Rebenitsch and C. Owen. Review on cybersickness in applications and visual displays. *Virtual Reality*, 20:101–125, 2016. doi: 10.1007/s10055-016-0285-9 1, 2
[29] V. Sevinc and M. I. Berkman. Psychometric evaluation of Simulator Sickness Questionnaire and its variants as a measure of cybersickness in consumer virtual environments. *Applied Ergonomics*, 82:102958, 2020. doi: 10.1016/j.apergo.2019.102958 2, 3
[30] S. Sharples, S. Cobb, A. Moody, and J. R. Wilson. Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems. *Displays*, 29(2):58–69, 2008. Health and Safety Aspects of Visual Displays. doi: 10.1016/j.displa.2007.09.005 2
[31] L. J. Smart, R. J. Pagulayan, and T. A. Stoffregen. Self-induced motion sickness in unperturbed stance. *Brain Research Bulletin*, 47(5):449–457, 1998. doi: 10.1016/S0361-9230(98)00103-8 2, 4
[32] K. M. Stanney, R. S. Kennedy, and J. M. Drexler. Cybersickness is Not Simulator Sickness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 41(2):1138–1142, 1997. doi: 10.1177/107118139704100292 1, 2, 4, 5
[33] W. B. Stone III. *Psychometric evaluation of the Simulator Sickness Questionnaire as a measure of cybersickness*. PhD thesis, Iowa State University, 2017. doi: 10.31274/etd-180810-5050 2, 3, 4
[34] N. Tian, P. Lopes, and R. Boulic. A review of cybersickness in head-mounted displays: raising attention to individual susceptibility. *Virtual Reality*, 26(4):1409–1441, 2022. doi: 10.1007/s10055-022-00638-2 1, 2, 6
[35] T. Weissker, M. Franzgrote, and T. Kuhlen. Try This for Size: Multi-Scale Teleportation in Immersive Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2298–2308, 2024. doi: 10.1109/TVCG.2024.3372043 6
[36] S. D. Young, B. D. Adelstein, and S. R. Ellis. Demand Characteristics in Assessing Motion Sickness in a Virtual Environment: Or Does Taking a Motion Sickness Questionnaire Make You Sick? *IEEE Transactions on Visualization and Computer Graphics*, 13(3):422–428, 2007. doi: 10.1109/TVCG.2007.1029 2, 3, 5
[37] D. Zielasko, B. Rehling, D. Clement, and G. Domes. Carry-Over Effects Ruin Your (Cybersickness) Experiments and Balancing Conditions Is Not a Solution. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 1–5, 2024. doi: 10.1109/VRW62533.2024.00007 5, 6
[38] D. Zielasko and T. Weissker. Stay Vigilant: The Threat of a Replication Crisis in VR Locomotion Research. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, VRST '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3611659.3615697 5