Mind the Gap: The Underrepresentation of Female Participants and Authors in Virtual Reality Research

Tabitha C. Peck Member, IEEE, Laura E. Sockol, and Sarah M. Hancock

Abstract—A common goal of human-subject experiments in virtual reality (VR) research is evaluating VR hardware and software for use by the general public. A core principle of human-subject research is that the sample included in a given study should be representative of the target population; otherwise, the conclusions drawn from the findings may be biased and may not generalize to the population of interest. In order to assess whether characteristics of participants in VR research are representative of the general public, we investigated participant demographic characteristics from human-subject experiments in the Proceedings of the IEEE Virtual Reality Conferences from 2015-2019. We also assessed the representation of female authors. In the 325 eligible manuscripts, which presented results from 365 human-subject experiments, we found evidence of significant underrepresentation of women as both participants and authors. To investigate whether this underrepresentation may bias researchers’ findings, we then conducted a meta-analysis and meta-regression to assess whether demographic characteristics of study participants were associated with a common outcome evaluated in VR research: the change in simulator sickness following head-mounted display VR exposure. As expected, participants in VR studies using HMDs experienced small but significant increases in simulator sickness. However, across the included studies, the change in simulator sickness was systematically associated with the proportion of female participants. We discuss the negative implications of conducting experiments on non-representative samples and provide methodological recommendations for mitigating bias in future VR research.

Index Terms—Virtual reality, gender, research methods, meta-analysis, simulator sickness, bias

1 INTRODUCTION

The contexts in which Virtual Reality (VR) research can be applied are wide-ranging and often address areas of key social importance. For example, the use of VR technology has shown promise as a way to improve real-world outcomes in fields such as driver education [48], surgical training [70], and the treatment of physical and psychological illnesses [22, 64]. Accordingly, the target user populations for VR applications are also diverse. Potential users may vary on many elements of identity, including age, gender, race/ethnicity, sexual orientation, and physical ability.

User-centered design practices, informed by research using human participants, are often core to the development and evaluation of VR applications [1]. From 2015 through 2019, 82% of papers published in the Proceedings of the IEEE Virtual Reality Conferences (hereafter referred to as the IEEE VR Proceedings) presented findings from human-subject experiments. The target user populations of these studies varied widely: some studies were designed to evaluate claims about the use of VR in the general population (e.g., [5, 6]), while others were designed for application within specific populations, such as individuals with multiple sclerosis [4]. Although the sample of participants in a single experiment is unlikely to be perfectly representative of all characteristics of the target population (e.g., age, gender, race/ethnicity, etc.), it is important that the sample be generally representative of the population for whom the findings are intended to generalize [1, 69]. The validity of research findings, and the degree to which these findings generalize to their intended real-world applications, may be limited by the use of non-representative samples [2].

Gender is an important demographic characteristic to consider in user research. For example, when conducting research with human subjects for the National Institutes of Health, the inclusion of female participants must be reviewed for acceptable representation, or justification must be provided for failure to adequately represent female participants [62]. In our analyses, we investigated whether the representation of female participants in studies presented in the IEEE VR Proceedings is consistent with the demographic characteristics of the general population, the target user population for many VR applications. We also evaluated the gender of authors of these manuscripts in order to assess whether female authors are underrepresented in VR research and whether the degree to which studies include female participants is related to the gender of the study’s authors. Finally, we conducted a meta-analysis in order to evaluate potential bias in research findings that may be introduced by a lack of gender diversity among participants.

It is important to acknowledge that gender represents only one facet of identity relevant to evaluating the representativeness of participants...
in VR research. In this manuscript, we focus on gender, as information regarding participants’ self-reported gender identity is commonly reported in publications of VR research. Gender refers to socially-constructed expectations for behaviors, characteristics, and roles considered appropriate for women and men, while sex is related to biological and physical characteristics typically associated with males and females [63]. Information regarding other participant characteristics, such as racial/ethnic identity, is much less commonly included in manuscripts. Furthermore, when manuscripts do characterize participant race/ethnicity, there are substantive differences in the ways in which this information is reported. It will be important for future research to investigate comparable questions with regard to other characteristics of participants.

2 BACKGROUND

2.1 Gender Differences in Virtual Reality Research

It is particularly important to ensure that research participants include representative diversity when the experiences of users from different groups are known to vary. Research on gender differences in VR suggests that the experiences of male and female users may differ in key ways, with important implications for the design and use of VR hardware and software.

Gender differences in users’ experiences with VR technology have been demonstrated in numerous and unrelated areas. In studies of redirected walking techniques, researchers have observed gender differences in perceptual thresholds [60, 85]. Men and women may interact with VR hardware in different ways. For example, when interacting with virtual hand-held objects, women show greater variability in controller grip and are more likely to switch between a loose and firm grip [11]. Studies have also found gender differences in user experiences and performance in virtual environments (VEs). Male and female users report differences in subjective spatial immersion and feelings of involvement [35]. In another study, men’s performance decreased as spatial tasks became more challenging, while women’s performance remained constant regardless of task difficulty [78]. User preferences may also differ; for example, women prefer same-gender self-avatar hands, while men show no gender preference for these avatars [73]. In other cases, gender differences are not found. For example, Kitson et al. were unable to detect gender differences in turning likelihood for point-to-origin tasks [41].

Simulator sickness is an intriguing area of research in which findings regarding gender differences have been mixed. Several studies report significant gender differences, finding that women are more susceptible to simulator sickness than men [26, 77]. These findings have been attributed to various causes, including women having a larger field-of-view compared to men, which is associated with greater sickness [37], anatomical or hormonal differences between sexes [77], or a tendency for men to under-report feelings of sickness [37]. However, other studies have found that men and women report comparable levels of simulator sickness [30, 85].

Given these findings, it is important for researchers to assess for the presence of gender differences in their studies. In practice, however, many studies do not compare male and female participants in order to determine if this characteristic should be accounted for in analyses. Furthermore, when studies do assess gender differences, many do not have adequate statistical power. In order to conduct analyses of gender differences, studies must have an adequate sample size, including sufficient representation of female participants. Failure to do so results in increased likelihood of a Type II error, in which the researcher fails to identify a true significant finding in their data. For example, Williams and Peck found no evidence for gender differences in simulator sickness in a redirected walking task [85]. However, their sample included only 16 participants. Although their sample was adequately powered to detect large and medium effects, their analyses would have required a total sample size of 68 participants in order to detect a small effect of gender (calculated using G*Power version 3.1.9.4). Thus, the current state of the literature may not be adequate to accurately characterize the presence of gender differences in important outcomes.

2.2 Implications for Design & Use

Designing without consideration for the target user population can have negative and sometimes life-threatening consequences. Consider the case of evaluating safety in automotive design. Until 2011, car manufacturers evaluated the safety of their vehicles with crash-test dummies designed to represent the average male size and anatomical proportions. When these vehicles were involved in accidents, women had a 47% higher chance of injury compared to men in the same accident [14]. Even today, the crash-test dummies used to evaluate whether vehicles are safe for women were not designed to accurately represent the average woman’s body, but are instead scaled-down versions of male dummies.

This example highlights the important consequences of inadvertent bias in research. Bias is an unequal and unfair weighting in favor of one group compared to another. In the context of research, bias negatively affects the validity of results [13]. The presence of bias in research findings can contribute to design bias: the development and dissemination of hardware and software whose characteristics systematically do not meet the needs of a subset of target users. Unintentional bias in product design can limit the groups of people who are able to easily, comfortably, or safely use a product [36]. If the experience of using a device is frustrating or confusing, people are less likely to adopt and continue to use it [61]. The presence of experimental and design bias has been demonstrated with regard for many underrepresented groups, including women [84], members of racial and ethnic minority groups [25], children [21], the elderly [29], transgender individuals [15], and people with physical disabilities [18].

An example of design bias in VR research is present in the interpubic distance (IPDs) accommodated by commercial head-mounted displays (HMDs). Wearing an HMD with a mismatched IPD is known to increase eyestrain and discomfort [49] and cause errors in depth perception [80]. Differences in IPD by gender, race, and age have been well documented [23]. However, commercial HMDs accommodate a higher percentage of men’s IPDs compared to women’s IPDs. For example, the HTC Vive Pro accommodates only 61% of women’s IPDs, yet supports 81% of men’s IPDs (based on the gender IPD specifications from the ANSUR dataset [23]). This may have important implications for user experiences. For example, a recent study found that the Oculus Rift is more nauseating for women than men [57]. It is possible that differences in the degree to which the design of HMDs accommodate female users’ needs, such as IPDs, contribute to these differences in user experiences. We can safely assume that users are less likely to continue using a device that makes them feel ill. Thus, the very design of HMDs is likely to dissuade widespread use by women.

One potential factor that may contribute to the presence of design bias is that women are less likely to be involved in the HMD design process as designers or testers [59]. This highlights another negative implication of inadvertent bias in VR research: the resulting bias in the design of VR technology may contribute to the underrepresentation of women as researchers and developers in the field. If the design of VR hardware and software does not take female users’ needs into account, these users will be less likely to engage with the technology and develop their interests in the field. Adolescent females are already half as likely as males to aspire to occupations in STEM fields [66]. This has been attributed to a wide range of causes, including parental encouragement and expectations [52], stereotypes about programmers and programming [65], socioeconomic factors [52], sexual harassment [67], imposter syndrome [50], stereotype threat [7], and implicit bias [74]. In addition to these potential causes, it is also important to consider the role of existing hardware and software in providing opportunities for girls and women to develop their interests in VR and technology.

Importantly, the underrepresentation of women in research and development roles in VR is likely to have negative consequences for research quality. Gender diverse design teams have been shown to be more innovative than gender homogeneous teams [46]. Similarly, culturally diverse teams have been shown to be more creative [76]. Additional research shows that collaborative research teams maintain high performance when diversity (including gender, ethnicity, career stage, etc.) is fostered [19]. This suggests that research groups and
development teams that are lacking in representative diversity may not achieve the levels of innovation, creativity, and productivity that would be possible with a less homogenous group.

3 General Analysis

We first investigated the gender distribution of participants and authors to determine if females were underrepresented in VR research. We hypothesized that:

H1: Female participants are underrepresented in human-subject VR research compared to male participants.

H2: Female authors are underrepresented in VR research compared to male authors.

H3: There are a greater number of female participants in studies in papers with female authors compared to papers with exclusively male authors.

3.1 Method

Studies were identified through review of all journal and conference papers from the IEEE VR Proceedings from 2015 through 2019. We chose the Proceedings of IEEE Virtual Reality Conferences since that conference is considered to be “the premier international venue for the presentation of research results in the broad area of virtual reality.” The full text of each manuscript was reviewed for eligibility by at least two of this paper’s authors. To be eligible for inclusion in the participant analyses, studies were required to meet the following criteria:

1. Presented results of an empirical study of human subjects.
2. Reported the number and gender (or number proportion) of participants.

In addition to participant gender, we also collected data on author gender. The gender of each author was recorded based on Internet searches of sources including personal websites, LinkedIn accounts, and university websites. Previous research using Internet searches to classify author gender suggests a misclassification rate of approximately 0.3% [33]; although this approach may not result in perfectly accurate classification of author gender (and does not account for authors with non-binary gender identities), these inaccuracies are unlikely to affect our findings.

To be eligible for inclusion in the author analyses, papers were required to meet the following criteria:

1. Identifiable author gender of each author (for author analyses).
2. Identifiable first-author gender (for first-author analyses).

3.2 Results

A flow chart depicting the process by which studies were assessed for inclusion in the participant analyses (green) and author analyses (orange) is presented in Figure 2. We reviewed 325 full-text manuscripts for eligibility, representing 429 independent studies. Of these studies, 319 empirical studies of human subjects reported information about the number and gender of included participants and were included in the participant analyses. The gender of two first authors and at least one author on three other papers could not be identified, yielding 320 manuscripts included in the author analyses and 323 manuscripts included in the first-author analyses.

Participant Analyses: Analyses of participant gender included 9,557 participants from 319 studies (one non-binary participant was excluded from analyses). Of these, 3,684 participants were female (39%) and 5,873 (61%) were male (see Figure 1). The difference in the number of male and female participants was analyzed using a Wilcoxon signed-rank test. We assume that the general population is comprised of approximately 50% women and girls [86]. Thus, significantly fewer female participants compared to male participants would be evidence that women and girls are underrepresented in VR research. Studies included significantly more male participants (Mdn = 13) than female participants (Mdn = 7), $r = -.57, p < .0001$. This supports H1: female participants are underrepresented in VR research compared to male participants.

We also investigated the stability of the representation of female participants across the included conference proceedings. A chi-square test was conducted to compare the proportion of female participants between conference years. The proportion of female participants ranged from 36% in 2015 through 42% in 2016 (see Table 1). There were no differences in the proportion of female participants across the included conference years, $\chi^2(4) = 7.30, p = .12$.

Author Analyses: Analyses of author gender included 1,425 authors from 320 manuscripts, with an average of 4.45 authors per manuscript. Of the included authors, 228 (16%) were female, and
Table 1: Proportion of female participants and authors in studies published in the IEEE VR Proceedings from 2015 through 2019, by year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Proportion Female Participants</th>
<th>Proportion Female Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>35.77%</td>
<td>17.56%</td>
</tr>
<tr>
<td>2016</td>
<td>42.05%</td>
<td>21.33%</td>
</tr>
<tr>
<td>2017</td>
<td>40.13%</td>
<td>15.19%</td>
</tr>
<tr>
<td>2018</td>
<td>38.33%</td>
<td>16.99%</td>
</tr>
<tr>
<td>2019</td>
<td>38.48%</td>
<td>13.73%</td>
</tr>
</tbody>
</table>

1,197 (84%) were male (see Figure 3). Over half (54%) of the included manuscripts included no female authors (see Table 2).

Similar to the participant analyses, manuscripts included a significantly higher number of male authors ($Mdn = 3$) compared to female authors ($Mdn = 0$), $r = -0.81$, $p < .0001$. The proportion of female authors ranged from 14% in 2019 through 21% in 2016 (see Table 1). We found no significant difference in the proportion of female authors between conference years, $\chi^2(4) = 5.96$, $p = .20$. These findings support H2: female authors are underrepresented in VR research compared to male authors.

**First-Author Analyses:** Of the 323 manuscripts included in the analyses of the gender of first authors, 50 manuscripts (15%) had female first authors. This proportion was significantly lower than the proportion of manuscripts with male first authors, $r = -0.69$, $p < .0001$. We also observed a trend indicating that manuscripts with female first authors were marginally more likely to include female co-authors than manuscripts with male first authors, $\chi^2(1) = 3.22$, $p = .07$. The odds ratio indicates that manuscripts were 1.43 (95% CI 0.94 - 2.16) times more likely to include additional female co-authors when the first author was a woman.

**Participant-Author Analyses:** Finally, we investigated whether author gender was related to the representation of female participants. Studies with at least one female author had significantly greater representation of female participants (41% female) compared to studies with only male authors (34% female), $\chi^2(1) = 39.38$, $p < .0001$ (see Figure 4). The odds ratio indicates that women and girls were 1.32 (95% CI 1.21 - 1.44) times more likely to participate in studies with at least one female author. Similarly, there were significantly more female participants in studies with female first authors (40% female) compared to studies with male first authors (37% female), $\chi^2(1) = 6.22$, $p = .01$. The odds ratio indicates that women and girls were 1.12 (95% CI 1.02 - 1.23) times more likely to participate in studies published in manuscripts with female first authors, compared to those in manuscripts with male first authors. This supports H3: there are a greater number of female participants in studies published in manuscripts with female authors compared to participants in studies published in manuscripts whose authors are exclusively male.

**3.3 Discussion**

The general analyses of participant and author gender support our hypotheses that female participants and authors are significantly underrepresented compared to the general population in VR research presented in the IEEE VR Proceedings.

We also found that the representation of female participants in these studies was related to the gender of the manuscript’s authors. There are a number of potential explanations for this finding. One possibility is that female researchers may be seen as role models [8] or mentors [20], which could aid recruitment of female participants. Another possibility is that female experimenters may place a higher emphasis on recruiting female participants than their male counterparts. This is particularly important, as previous research has demonstrated that gender and race of the experimenter may influence findings [55, 83]. Future research should further investigate the effects of female and underrepresented minority researchers on participant recruitment.

One limitation of our analyses is that we did not assess the gender of manuscripts’ last author. This position often indicates the senior researcher on a project. It is possible that a similar relationship exists between the gender of the last author and the proportion of female co-authors, first-author gender, and proportion of female participants. Given the degree to which the senior researcher plays a leadership role in recruitment and other elements of research design, future research should investigate associations between last author gender and other author and participant demographics.

Our results demonstrate that women authors are underrepresented compared to the general population. The same underrepresentation of women appears in conference attendance: however, the gender of attendees at IEEE VR conferences appears to only have been collected in 2018 (19% female attendees) and 2019 (18% female attendees) [42]. The percentage of female conference attendees closely corresponds to the percentage of female authors in both years, 2018 (17% female authors) and 2019 (14% female authors) (see Figure 5). The underrepresentation of female conference attendees and authors also mirrors the underrepresentation of women in computer science. According to the 2018 Taulbee survey, only 19% of doctoral degrees in the field were awarded to women [88]. Similar trends in the underrepresentation of women have been seen on journal editorial boards in the mathematical sciences, where only 8.9% of editorships are held by women even though 15% of the tenure/tenure-track positions are held by women [79].

Research participants are often recruited from the general public or from students and other members of the community at institutions where the research is being conducted. In contrast to the low representation of women within computer science, the general population of the world is roughly 50% women [86], and women comprise the majority (57%) of students on college campuses [54]. Thus, obtaining a sample of participants with adequate gender diversity should be a realistic and attainable goal for all studies making claims about the general population. But, do we really need to consider the gender of participants in our studies to perform quality research?
We also observed a trend indicating that manuscripts with female first authors were marginally more likely to include female co-authors than those in manuscripts with male first authors (34% female), compared to those in manuscripts with female first authors (40% female), compared to studies with male first authors (37% female). This proportion was significantly lower than the proportion of manuscripts with female first authors, which was 32.05%.

We also found that the representation of female participants in these studies was related to the gender of the manuscript’s authors. There are multiple independent studies that may be systematically related to effect sizes. In a meta-analysis, moderator (t-test). These analyses evaluate the relationship between study-level characteristics, referred to as moderators, and study-level outcomes, represented by each study’s effect size. For example, in a meta-analysis, a researcher could investigate whether there are differences in reaction time between groups using a t-test.

A flow chart depicting the process by which studies were assessed for inclusion in the meta-analysis is presented in Figure 2 (purple). We reviewed 325 full-text manuscripts for eligibility, representing 429 independent studies. Of these studies, 407 studies that did not meet inclusion criteria were excluded. A total of 22 independent studies, from 21 manuscripts, were included in the meta-analysis.

### 4.1 Method

#### 4.1.1 Search Procedure and Selection of Studies

Studies were identified through review of all journal and conference papers from the IEEE VR Proceedings from 2015 through 2019. The full text of each manuscript was reviewed for eligibility by at least two of this paper’s authors. To be eligible for inclusion in the meta-analysis, studies were required to meet the following criteria:

1. Presented results of an empirical study of human subjects.
2. Study participants were exposed to an immersive VE using an HMD. We excluded studies in which there was no VR exposure, augmented reality studies, non-immersive VR studies, projection mapping studies, and studies of CAVE-like systems.
3. Assessed simulator sickness before and after VR exposure using the SSQ.
4. Reported data for the calculation of an effect size representing the average within-participant change in SSQ among participants exposed to a VE.
5. Reported the number or proportion of female participants exposed to a VE.

### 4.1.2 Data Extraction

We developed a coding manual (available upon request) for extracting characteristics of studies and data for the calculation of effect sizes from each manuscript. Data for each study was extracted according to the coding manual by the student author (S.M.H.) and reviewed for accuracy by senior authors (T.C.P. L.E.S.). Characteristics of studies, authors, and samples were assessed according to the coding manual. These variables included manuscript type (conference vs. TVCG), gender of the paper’s first author (female vs. male), proportion of female participants, and average participant age.

In a meta-analysis, effect sizes from included studies must be independent from one another. In order to ensure that our data met this statistical assumption of independence, a single effect size was calculated for each study. If a study reported SSQ results separately for different groups of participants, the effect size was calculated using combined data from all participants. If a study used the same sample of participants in more than one experiment, the effect size was calculated using results from the first experiment. If a study assessed the SSQ at more than one post-VR exposure time point, the effect size was calculated using the first assessment conducted following a VR exposure.

---

**4 Meta-Analysis**

In order to evaluate whether the underrepresentation of female participants introduces potential bias into research findings, we conducted a meta-analysis and meta-regression to assess whether the degree to which studies include female participants is systematically associated with the studies’ findings. For our meta-analysis, we chose to investigate a common outcome assessed in a wide range of VR studies: the degree to which individuals exposed to a VE using an HMD experience simulator sickness.

We chose simulator sickness to provide an illustrative example of our overall concern regarding the relationship between participants’ demographic characteristics and study outcomes. One strength of selecting this outcome is that it is assessed in studies investigating a wide variety of research questions using different paradigms. Another strength is the existence of a widely-used, standardized measure of simulator sickness: the Simulator Sickness Questionnaire (SSQ) [38]. Finally, as described above, research findings regarding gender differences in simulator sickness have been mixed. Thus, this meta-analysis and meta-regression provide an interesting opportunity to investigate and characterize these varied findings.

We hypothesized that:

- **H4**: The change in simulator sickness from pre- to post-VR exposure is systematically associated with the proportion of female participants in a given study.
- **H5**: The change in simulator sickness from pre- to post-VR exposure is not systematically associated with other characteristics of studies (manuscript type, first-author gender, and average participant age).

---

**4.2.2 Effect Sizes**

Effect sizes were calculated using the first assessment conducted following a VR exposure. We computed standardized mean differences in SSQ effect sizes using the formula:

$$d = \frac{\bar{Y}_{female} - \bar{Y}_{male}}{SD_{total}}$$

where $\bar{Y}_{female}$ and $\bar{Y}_{male}$ represent the mean SSQ scores for female and male participants, respectively, and $SD_{total}$ is the pooled standard deviation of SSQ scores.

---

**4.3 Results**

In total, we included 22 independent studies from 21 manuscripts. The proportion of female participants exposed to a VE was significantly lower than the proportion of female attendees at the IEEE VR conference. The proportion of female attendees was 41%, compared to 32.05% in the Proceedings.

---

**4.4 Discussion**

The results of our meta-analysis suggest that female authors are underrepresented in IEEE VR Proceedings, compared to the representation of female participants in these studies. This underrepresentation is consistent with findings from other studies in the field of virtual reality.

---

**Appendix**

**Table 1:** Proportion of manuscripts with female first authors, compared to those in manuscripts with female first authors.

<table>
<thead>
<tr>
<th>Manuscript Type</th>
<th>Percentage of Female First Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscripts</td>
<td></td>
</tr>
<tr>
<td>Manuscripts</td>
<td></td>
</tr>
<tr>
<td>Manuscripts</td>
<td></td>
</tr>
<tr>
<td>Manuscripts</td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 5:** The proportion of female participants (purple, long-dash) and authors (orange, short-dash) in the IEEE VR Proceedings, and the proportion of female attendees (green, line) at IEEE VR. Data regarding attendee gender was only available for 2018 and 2019.

---

**References**


---

**Author Notes**

PECK ET AL.: MIND THE GAP: THE UNDERREPRESENTATION OF FEMALE PARTICIPANTS AND AUTHORS IN VIRTUAL REALITY RESEARCH
4.1.3 Data Analysis

We conducted a meta-analysis and meta-regression using Comprehensive Meta-Analysis v. 3.0, using the analytic plan described below. Results of these analyses are reported in subsection 4.2.

Effect sizes were calculated using Cohen’s d to represent the standardized mean within-subject change in simulator sickness, so that positive effect sizes indicate an increase in simulator sickness from pre- to post-exposure:

\[ d = \frac{M_{\text{Post}} - M_{\text{Pre}}}{SD_{\text{Diff}}/\sqrt{2(1 - r)}} \]

where

\[ SD_{\text{Diff}} = \sqrt{SD_{\text{Pre}}^2 + SD_{\text{Post}}^2 - 2rSD_{\text{Pre}}SD_{\text{Post}}} \]

and r is the pretest-posttest correlation of the outcome measure. As this information is not commonly reported, following the recommendations of Lipsey and Wilson, this value was estimated using the test-retest reliability of the SSQ (r = .90) [28, 51]. This correlation represents the average similarity in participant scores on the SSQ in the absence of an intervention.

Overall Effect Size: We first conducted a meta-analysis to estimate the average within-participant change in SSQ across all studies. Because our goal was to characterize the relationship between potential moderators and effect sizes within our sample of studies, the overall effect size was estimated using a fixed effects model. In our model, we assume that variability among effect sizes is due to a combination of subject-level sampling error within individual studies and systematic variability in effect sizes between studies [51]. As a result, in our analyses, the overall effect size can be interpreted as an estimate of the true change in SSQ scores among the included studies.

After estimating the overall effect size, outliers were identified using the Sample-Adjusted Meta-Analytic Deviance Statistic (SAMD) [34]. Outliers were identified for exclusion on the basis of quantitative and visual review of the distribution of SAMD scores [3]. We then used Duval and Tweedie’s trim-and-fill procedure to assess and correct for publication bias [24].

Analyses of Heterogeneity: Prior to evaluating the relationship between study characteristics and effect sizes, we first needed to establish sufficient variability in effect sizes among studies to justify analyses of moderation. The heterogeneity of effect sizes was assessed using the Q statistic and I² value [12]. The Q statistic evaluates whether the variability in observed effect sizes is greater than would be expected on the basis of within-study variability; significant Q statistics indicate the presence of between-study heterogeneity. The I² value represents the proportion of heterogeneity in effect sizes that can be attributed to between-study variance. We used Higgins and colleagues’ suggested cutoff I² values of 25, 50 and 75 to indicate low, medium, and high levels of heterogeneity, respectively [32].

Moderator Analyses: If both the Q statistic and I² value indicated significant heterogeneity in effect sizes, we conducted analyses of moderation to investigate the relationship between study characteristics and effect sizes. The relationship between effect size and continuous moderators (proportion of female participants and average participant age) was assessed using meta-regression. The relationship between effect size and categorical moderators (paper type: TVCG vs. conference paper; first author gender: female vs. male) was assessed using mixed-model ANOVA.

4.2 Results

Overall Effect Size: Results of the meta-analysis are presented in Table 3. The meta-analysis included findings from 22 independent studies, representing data from 592 participants. The distribution of women (37%) and men (63%) was comparable to our findings regarding the gender of participants in all human subjects studies from the 2015 through 2019 IEEE VR Proceedings. A significant increase in simulator sickness from pre- to post-exposure was observed in 16 studies, with effect sizes for individual studies ranging between −0.05 and 3.50. The estimated overall effect size was \( d = 0.30 \) (95% CI 0.27 − 0.34, \( p < .001 \)), indicating a small but significant increase in simulator sickness after exposure to a VE.

Quantitative and visual examination of the distribution of SAMD scores identified one outlier [9]. After removal of this outlier, the estimated average effect size remained significant (\( d = 0.29 \), 95% CI 0.26 − 0.33, \( p < .001 \)).

Duval and Tweedie’s trim-and-fill procedure [24] suggested potential publication bias, with an estimated 7 studies missing with values to the left of the mean (indicating potential unpublished studies with smaller changes in simulator sickness). After the trim-and-fill correction was applied, the overall effect size continued to indicate a small but significant increase in simulator sickness from pre- to post-exposure (\( d = 0.18 \), 95% CI 0.14 − 0.21, \( p < .001 \)).

Analyses of Heterogeneity: Heterogeneity analyses indicated significant variability in effect sizes among included studies. The Q statistic was significant (\( Q(20) = 193.19 \), \( p < .001 \)), indicating significant heterogeneity. The I² value of 90 indicated a large degree of heterogeneity. These analyses provided evidence of sufficient variability in effect sizes between studies. Moderator analyses were therefore conducted to assess whether characteristics of the included studies were systematically associated with effect size.

Moderator Analyses: We conducted a meta-regression to assess whether the proportion of female participants in each study was associated with the study’s observed effect size. The proportion of female participants in the individual studies ranged from 0 to 81%, with a mean of 35% (SD = 21%). Results of the meta-regression indicated that the proportion of female participants was significantly associated with effect size (slope = −0.30, 95% CI −0.49 − −0.11, \( p = .002 \)). This indicates that smaller increases in simulator sickness were observed in studies with a greater proportion of female participants (see Figure 6).

Consistent with H4, the average change in simulator sickness from pre- to post-exposure systematically varies according to the proportion of female participants in a study.

Consistent with H5, no other moderators were systematically associated with effect size. There was no difference in the change in simulator sickness between TVCG papers (\( k = 12 \), \( d = 0.36 \)) and conference papers (\( k = 9 \), \( d = 0.40 \)); \( Q(1) = 0.09 \), \( p = .768 \). There was also no difference in effect size between studies with female first authors (\( k = 3 \), \( d = 0.55 \)) and studies with male first authors (\( k = 18 \), \( d = 0.35 \)); \( Q(1) = 2.34 \), \( p = .126 \). Finally, each study’s average participant age was not associated with effect size (\( k = 18 \), slope = .003, \( p = .423 \)).
We conducted a meta-analysis and meta-regression using Comprehendstudies, representing data from 592 participants. The distribution of...e to investigate the relationship between study characteristics and...d the presence of between-study heterogeneity. The...eration to test-retest reliability of the SSQ (\(d\)) was not associated with effect size (\(r = 18, p < 0.001\)).

### 4.3 Discussion

This meta-analysis provides evidence that the underrepresentation of female participants in VR studies may result in biased findings. The overall effect size indicates that, on average, participants in VR studies using HMDs experience small but significant increases in simulator sickness after VR exposure. This holds true even after excluding studies with extreme findings and correcting for publication bias. However, this outcome was systematically related to the proportion of female participants included in each study: smaller increases in simulator sickness following VR exposure were observed in studies with a greater proportion of female participants.

This finding is surprising, as some studies have found evidence that women are more susceptible to motion sickness in VEs using both HMDs [57] and CAVE-like systems [26]. However, it is important to note that the study-level findings of this meta-analysis do not necessarily indicate that individual female participants reported smaller changes in simulator sickness. There are a number of potential explanations for our finding, ranging from methodological artifact to true differences in susceptibility to simulator sickness between men and women. It is possible that other, perhaps unidentified, characteristics of studies may account for this pattern. For example, the representation of female participants may not be comparable across different kinds of VR research which vary in the degree to which they elicit simulator sickness. A limitation of this meta-analysis is that we were unable to directly assess other study characteristics that may have contributed to the findings. For example, the small number of studies included in different research categories (e.g., avatars, cybersickness, walking and navigation) prevented us from controlling for potential differences between research topics when investigating the effect of gender. In the context of mixed findings from individual studies that have evaluated gender differences in simulator sickness, the results of this meta-analysis suggest that further research is necessary.

Our concerns regarding the underrepresentation of female participants in VR research are strengthened by the finding that the proportion of female participants was the only study characteristic that was systematically associated with effect size. We found no evidence that effect size was systematically associated with other characteristics of studies, including characteristics of the publication (manuscript type), authors (first author gender), or sample (average participant age). This suggests that there may be true variability in this outcome related to gender, and that conclusions drawn from samples with inadequate gender diversity may not accurately characterize simulator sickness in the general population. Although we assessed simulator sickness as an outcome in our meta-analysis, there is reason to believe that the underrepresentation of female participants may also be problematic for other outcomes. For example, a recent study comparing manual and automatic speed adjustment reported that there was no evidence that the automatic navigation had any negative effects on users. However, only 22% of participants in the study were women [56]. There is clear evidence for small but significant sex differences in spatial navigation [58]. Given these differences, it is possible that a study with a more representative sample might yield different findings.

In the studies included in the meta-analysis, many authors characterized their findings in terms of “individual” or “user” experiences. Our findings suggest that characterizing the results of typical VR studies as relevant to “user” experiences may be inaccurate. Furthermore, when female participants are included, many studies do not assess gender differences in findings. The underrepresentation of female participants and failure to explicitly address potential gender differences is rarely discussed as a limitation in studies for which this may present a concern.

The meta-analysis provides a snapshot of five years of data from IEEE VR, which may not be representative of other conferences. Future
work should investigate the representation of women and girls at other prominent VR conferences. Inclusion of data from a longer time frame could allow for the evaluation of changes in the representation of female participants in VR research over time. Future meta-analyses should also evaluate additional outcomes, such as presence, that may be systematically related to the proportion of female participants included in studies.

5 Conclusion

Women are significantly underrepresented as both participants and authors in VR research compared to the general population. Critically, this underrepresentation results in biased research findings. When VR hardware and software are designed without consideration for women, the design of these products may dissuade women from adopting or utilizing new technology. More critically, VR technology for real-world applications, such as training and treatment, may not be as effective. For example, VR is regularly used to provide exposure-based therapy for post-traumatic stress disorder (PTSD) [22]. However, the prevalence of PTSD is more than twice as high among women compared to men [39]. Virtual exposure therapies developed using research with predominantly male participants may not be as effective for the very patients most likely to benefit from these interventions. Furthermore, by hindering women’s ability to develop and support interests in VR technology, biases in VR research and design may decrease the likelihood that women pursue careers in VR research or development. The passive exclusion of potentially interested and talented female scholars from the field deprives us of the knowledge their scholarship would generate.

If current trends continue, inaccurate conclusions drawn from biased research will continue to promote the development of technology that does not take the needs of female users into account. As a result, girls and women will continue to have fewer opportunities to develop and explore interests in VR technology, and we will continue to see mostly male authors and attendees at IEEE VR and other VR conferences. What we will not be able to see are the ways in which future VR research will be constrained by this gender bias. Lau and colleagues demonstrated that academic teams comprised of both men and women were more innovative than teams with less gender diversity; 75% of the lowest-performing groups in their study were single-gender teams [46]. We may not be able to quantify the degree to which the field has been harmed by our collective failure to recruit and retain female colleagues, but this study provides evidence that innovation and performance in VR research has likely been impaired by the absence of these women in our ranks.

The positive news is that we can begin to address this problem and produce higher-quality scholarship with a simple solution: researchers conducting experiments in VR must commit to using samples that are representative of the population for which the technology is being designed. If researchers wish to make claims about technology use by the general population, these claims must be justified by adequate diversity among participants; for example, adequate representation of male and female participants (and evaluating the experiences of transgender and gender nonconforming participants, when relevant), representative racial and ethnic diversity, and international diversity. Demographic information must be included when reporting characteristics of participants, including age, gender, and race/ethnicity, so that readers can accurately interpret the studied population and future meta-analyses of participant demographics can be performed. We provide example demographic questionnaires in Section 6. As suggested by the National Institutes of Health, researchers should also explicitly investigate systematic differences in findings associated with participant demographic characteristics [62].

As editors and reviewers, we must hold our colleagues’ research to these standards. When reviewing papers that make general claims but do not use a sample representative of the population, we must provide appropriate constructive criticism and advice to strengthen the research and produce higher-quality VR scholarship.

6 Appendix

To support future meta-analyses and collection of data from representative samples of the population, studies should collect and report participant gender. For collecting data regarding participant gender, the Human Rights Campaign suggests using two questions to assess gender identity and transgender status.

1. What is your gender?
   - Female
   - Male
   - Non-binary/ third gender
   - Prefer to self-describe _____________
   - Prefer not to say

2. Transgender is an umbrella term that refers to people whose gender identity, expression or behavior is different from those typically associated with their assigned sex at birth. Other identities considered to fall under this umbrella can include non-binary, gender fluid, and genderqueer as well as many more.
   - Do you identify as transgender?
     - Yes
     - No
     - Prefer not to say

For collecting data regarding participant race/ethnicity, we recommend allowing participants to self-identify, as in the example below.

1. What is your race?
   - Prefer to self-describe _____________
   - Prefer not to say

2. What is your ethnicity?
   - Prefer to self-describe _____________
   - Prefer not to say

For researchers who prefer to use standardized questions, the questions below are based on the United States census and modified through consultation with individuals from four continents. Participants should be permitted to select multiple responses.

1. What is your race?
   - Black, African, or African descent
   - East Asian
   - Native or Indigenous
   - Polynesian
   - South Asian
   - White
   - Prefer to self-describe _____________
   - Prefer not to say

2. What is your ethnicity?
   - Hispanic, Latinx, or Spanish Origin
   - Not Hispanic, Latinx, or Spanish Origin
   - Prefer to self-describe _____________
   - Prefer not to say
Acknowledgments
The authors wish to thank the diverse group of people consulted on this paper. Consultant diversity included: gender, race, ethnicity, age, sexual orientation, nationality, and academic discipline. Additionally, the authors wish to thank the Davidson Research Initiative for supporting this work.

References


