This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2026.3680686

Telling Us What You Experience: Effects of Questionnaire Interface Design on Subjective Measurements in Virtual Reality

Lucas Küntzer, Martin Feick, Max Benzschawel, Naz Al Kassm, Tilo Mentler, Heike Spaderna, Robert J. Teather, Georg Rock

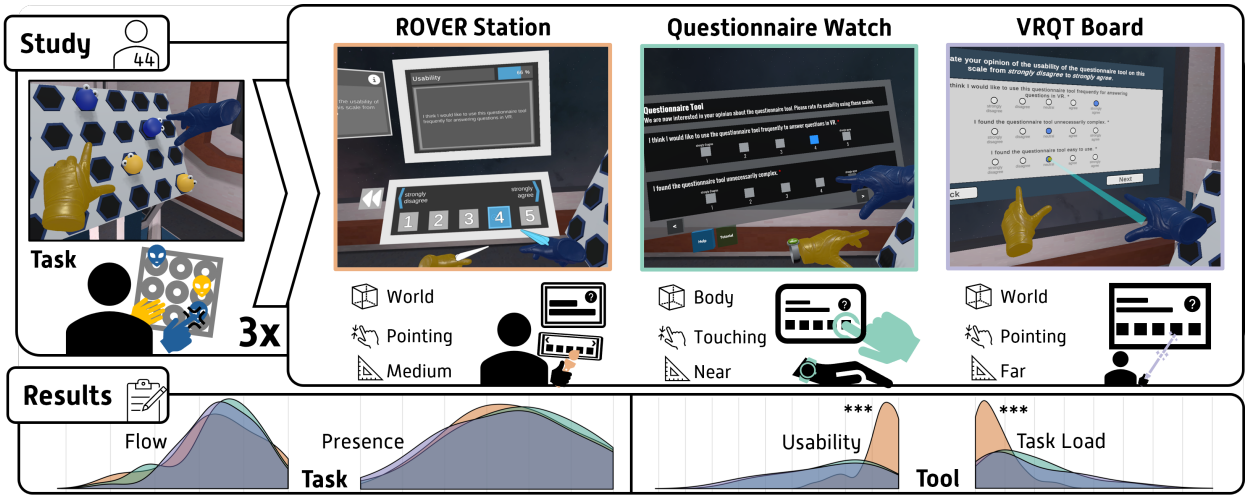

Fig. 1: Participants (n=44) performed a task in VR, followed by self-reports using one of three questionnaire tools in VR (ROVER [30], Q-Watch [3], and VRQT [13]) with interfaces (Station, Watch, Board) featuring different spatial placement (body-anchored, world-anchored), interaction styles (direct touch, laser pointer), and interaction distances (near, medium, far). Flow and Presence ratings regarding the task remained consistent, while distribution of Usability and Task Load ratings show significant differences between tools.

Abstract—Virtual reality (VR) enables immersive, tightly controlled experiments but complicates subjective measurement: moving participants out of the virtual environment (VE) to complete questionnaires changes the measurement context and can increase recall bias. Questionnaires embedded in the VE ($_{IN}$VRQs) enable immediate, time-efficient self-reports and repeated sampling, yet adoption is limited by concerns that $_{IN}$VRQs might bias primary experiential measures and by the implementation burden of usable interfaces. In this within-subject study, 43 participants completed post-task questionnaires using three representative 2D $_{IN}$VRQ interface designs: a body-anchored watch (direct touch), a world-anchored station (handheld pointer), and a world-anchored board at 5 m (laser pointer). Usability and task load differed significantly across interfaces, revealing clear design trade-offs. Interview feedback highlighted corresponding ergonomic themes. In contrast, ratings of presence and flow related to the constant task and VE showed no practically relevant differences across interfaces, supported by equivalence tests. Findings suggest that once basic usability requirements are met, interface choice is unlikely to meaningfully bias experiential measures in similar setups. This study contributes methodological guidance for implementing or selecting $_{IN}$VRQ tools to improve VR study quality, comparability, and reproducibility. Supplemental materials are available at osf.io/528yk (CC BY 4.0).

Index Terms—Virtual Reality, in-VR Questionnaire, inVRQ, Interface Design, Usability, Task Load, Subjective Measurements.

1 Introduction

Virtual Reality (VR) has not only gained popularity in fields such as education, training, and entertainment but also as a promising research medium across disciplines [46]. By leveraging its ability to simulate controlled immersive experiences, VR elicits realistic user responses comparable to screen-based or traditional methods [23, 54].

However, methodological inconsistencies in the reliable and valid assessment of user experience can undermine study comparability and reproducibility. These remain significant barriers to fully realizing VR's research potential, especially in non-technical fields [22, 51, 59].

Traditionally, user experience is measured via self-report questionnaires administered outside of VR ($_{OUT}$VRQ). This can reduce test quality by decreasing test efficiency, interrupting immersion, delaying responses, and reducing participant comfort, particularly in repeated-sampling study designs [4, 37, 40, 52].

Recent literature recommends the integration of questionnaires directly within the virtual environment (VE) [4, 40, 43, 52]. However, $_{IN}$VRQ interfaces vary in positioning (e.g., world-anchored versus body-anchored), visual presentation (2D versus interactive 3D objects), and interaction modalities (controller-based pointing, freehand gestures, gaze, or voice) [4, 41, 42, 55]. These design choices can affect

- *Lucas Küntzer, Max Benzschawel, Tilo Mentler and Georg Rock are with Trier University of Applied Sciences, Germany. E-mail: {kuentzel | mxbn3568 | mentler | rock}@hochschule-trier.de*
- *Martin Feick is with Karlsruhe Institute of Technology, Germany. E-mail: martin.feick@kit.edu*
- *Naz Al Kassm is with Carleton University, Canada. E-mail: naz.alkassm@carleton.ca*
- *Heike Spaderna is with Trier University, Germany. E-mail: spaderna@uni-trier.de*
- *Robert J. Teather is with Monash University, Australia. E-mail: rob.teather@monash.edu*

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

user experience, completion time, and comfort, particularly for VR novices [4,41,42,55]. Given the variety and potential trade-offs among these emerging solutions, researchers face uncertainty regarding the influence of $_{IN}$VRQ design choices on subjective measurements, user perceptions, and overall effectiveness in VR studies [4,42]. The implementation of appropriate $_{IN}$VRQ interfaces requires domain expertise and technical knowledge, which can be particularly challenging for researchers from non-technical fields looking to increase test quality and efficiency in their VR-driven studies [4,26].

This study systematically evaluates the interfaces of three $_{IN}$VRQ tools developed for researchers (see Fig. 1), each representing distinct design choices: the Questionnaire Watch (Q-Watch) [3], the Rating Overlay for Virtual Environments in Research (ROVER) [30], and the Virtual Reality Questionnaire Toolkit (VRQT) [13].

The contribution lies in a tightly controlled and statistically well-powered empirical validation of practical $_{IN}$VRQ tool choices, combining mixed-effects modeling with equivalence testing to support robust inferences about measurement stability. Given converging evidence that $_{IN}$VRQ self-reports match $_{OUT}$VRQ baselines for validity and reliability, this study (n=43) isolates interface effects on post-task user experience ratings using a constant VR stimulus: a repetitive, visuomotor task in a static VE. The results reveal significant differences in Usability and Task Load across the evaluated tools, highlighting the impact of interface design on study participation. Despite pronounced usability differences, measurements of task-related user experience dimensions (Flow and Presence) remained stable, which suggests a robustness of such measurements under the present conditions.

While potentially differing in more immersive or extensive VR scenarios, this indicates that the effects of $_{IN}$VRQ interface design on these ratings are of negligible practical relevance within this task class. The findings provide empirical guidance on implementing or selecting effective $_{IN}$VRQ tools for user experience assessment within VEs, facilitating greater study quality, reproducibility, and comparability in VR-based research.

2 Related Work

Empirical studies suggest clear benefits of embedding questionnaires within VEs, yet they also report inconsistencies when assessing user experience in VR [4, 16, 40, 43, 52]. Dimensions of subjective user experience in VR, such as Presence and Flow [10,20], are of significant interest and can predict overall user experience and performance in VR research [17,18]. Presence refers to the subjective sense of "being there" in a VE [45,50], while Flow describes the psychological state of optimal experience [12]. Flow is characterized by complete absorption in an activity, clear goals, immediate feedback, and a balance between challenge and skill [12].

Presence and Flow are inherently state-dependent and sensitive to individual characteristics, thematic congruity, and task context [9,31]. Violations of alignment between task, environment, and interaction mechanics can reduce users' sense of Presence [9] and Flow experience [32]. Increased perceived workload, frustration, and other interface-related disruptions degrade user experience and attentional continuity, particularly in complex or demanding VR tasks [15,32].

Schwind et al. [43] found no significant differences in mean Presence scores between $_{IN}$VRQs and $_{OUT}$VRQs after exposure to virtual environments (VEs) with varying scene realism (abstract, realistic), suggesting comparable measurement validity. However, they observed reduced variance in Presence ratings with $_{IN}$VRQs, indicating decreased bias caused by breaks in presence (BIPs). Expanding on this, Putze et al. [37] specifically investigated physiological responses to BIPs and found significantly reduced reactions with $_{IN}$VRQs, but only limited evidence for improved measurement reliability. Subsequent studies measuring Presence have produced mixed results, with some finding significant differences between $_{IN}$VRQ interfaces [48,52], while most reported no effects between interface conditions or when compared to $_{OUT}$VRQs [4,19,40,41,43,52].

Test quality of research studies is commonly defined by objectivity, reliability, and validity [34]. However, it also includes other factors such as efficiency, fairness, and respondent burden [34]. Previous research consistently emphasizes the benefits of embedding questionnaires within VEs to sustain immersion, reduce disruptions, and improve rating immediacy, efficiency, and participant comfort, which ultimately enhances study quality while retaining reliability and validity [4,37,40,43,52].

2.1 $_{IN}$VRQ Design

Embedding questionnaires within VEs introduces complexities in design choices and interaction modalities for $_{IN}$VRQ interfaces. There is a lack of consensus on the optimal method for presenting $_{IN}$VRQs, which range from 2D panels to fully 3D interactive objects [4,41,42]. Regal et al. [40] evaluated various positions for their $_{IN}$VRQ interface (world-anchored billboard, body-anchored head-up display, body-anchored hand mount) and found no differences in usability ratings, despite participants' preference for the billboard interface.

Comparing extradiegetic and intradiegetic interfaces, Wagener et al. [52] found intradiegetic interfaces significantly enhanced user experience and Presence, and were strongly preferred by participants, despite increased completion times. Studies aligning intradiegetic 3D $_{IN}$VRQ interface interactions with the primary VR task, such as shooting targets or using 3D objects, demonstrated similar usability and workload compared to 2D interfaces, which participants favored for their familiarity [1,19,41,48].

Alexandrovsky et al. [4] systematically analyzed current practices around embedding 2D questionnaires within the VE, identifying inhibitors and design guidelines for effective $_{IN}$VRQ implementation. Their study's participants preferred $_{IN}$VRQs over $_{OUT}$VRQs, despite slightly increased physical demand and lower, but acceptable usability ratings. Wei et al. [55] compared joystick and ray casting $_{IN}$VRQ selection methods against a $_{OUT}$VRQ baseline, finding $_{OUT}$VRQs to be faster and less demanding in terms of Task Load in their setup.

Interaction and presentation modalities of $_{IN}$VRQs introduce different usability trade-offs, workload implications, and potential biases. Previous studies indicate a preference among participants for 2D $_{IN}$VRQs administered within the VE of the primary VR stimulus or task [4,16,40,43]. However, empirical guidance remains limited for practice-oriented decisions among different 2D $_{IN}$VRQ interface implementations, particularly regarding whether usability and workload differences translate into practically meaningful bias in experiential ratings.

2.2 $_{IN}$VRQ Tools

A systematic literature review by Safikhani et al. [42] identified common design choices, interaction methods, and usability challenges associated with $_{IN}$VRQ interfaces. Their proposed taxonomy categorizes $_{IN}$VRQs into 2D and 3D interfaces, with interaction methods ranging from pointer-based selection to direct hand interactions. While gaze- or gesture-assisted interfaces show potential [36,44], they remain overshadowed by traditional touch or pointer-based UI interaction in practice. The taxonomy notes the predominance of pointer-based interaction in $_{IN}$VRQ interfaces and highlights the difficulty of implementation due to a lack of practice-oriented empirical guidance on how design choices influence user experience [42].

Similarly, Küntzer et al. [30] provide an overview of the presentation and interaction modalities of $_{IN}$VRQ interfaces, showing that world-anchored interfaces with laser pointing or ray casting interactions are commonly most studied and employed. Prior work [4,55] recommends this interface type based on their evaluation results and participant feedback. However, usability issues, such as difficulty with pointing and selection or fatigue from the physical demands of VR interactions, could negatively impact data quality by causing frustration or unintentionally biasing participant responses toward easier interaction choices [4,41]. This is especially relevant for longer sessions or with inexperienced or vulnerable participants [29].

Consideration of usability challenges such as readability [47], interaction difficulty [56], and accessibility [8] is necessary for robust user experience assessments [29]. Consequently, the implementation of $_{IN}$VRQs requires additional expertise, time, and resources from researchers, which inhibits their usage [4].

Recognizing these methodological complexities, several open-source $_{IN}$VRQ tools with 2D interfaces have been developed to aid researchers in conducting questionnaire-based subjective measurements in VEs, including VRate [39], VRQT [13], and ROVER [30]. The VRQT by Feick et al. [13] is a Unity3D-based asset that provides high flexibility, though it also demands technical expertise for integration into study VEs and has third-party plugin dependencies. Although VRQT has not been formally evaluated, it is based on research-backed design guidelines [4, 13] and has been utilized successfully in various studies by other researchers (e.g., [48, 58]). ROVER by Küntzer et al. [30] emphasizes ease of use and accessibility, receiving high usability ratings (>90 UMUX score [14]) from 68 participants across several studies [26].

Fundamentally, VRQT and ROVER adhere to the design guidelines proposed by Alexandrovsky et al. [4], focusing on world-anchored interfaces with laser pointing or ray casting interaction, though their implementations differ. In contrast, the Questionnaire Watch (Q-Watch) proposed by Al Kassm et al. [3] employs a body-anchored interface mounted on a virtual wristwatch, utilizing bi-manual interaction with hand tracking. The Q-Watch was evaluated in an unpublished, exploratory user study (n=12) [2], receiving modest task load ratings (M=2.72(0.38) as mean score on a 7-point Likert scale based on items by Harris et al. [21]) and an overall System Usability Score [7] of 78.75.

These $_{IN}$VRQ tools vary in presentation, interaction, and ergonomics, yet empirical evaluation is limited regarding their usability and impact on the reliability and validity of subjective measurements. Given the documented sensitivity of experiential measures to BIPs and environmental modalities [16, 37, 43], different tool interfaces could impair the reliability and comparability of research findings by introducing usability-related biases.

In an exploratory study ($n = 16$), Safikhani et al. [41] compared a traditional web-based ($_{OUT}$VRQ) baseline against two $_{IN}$VRQ designs (a 2D panel and an interactive 3D object). Their analyses reported no significant differences in SUS and Presence across designs but indicated trade-offs in workload and completion time and a participant preference for the 2D presentation. Importantly, the authors highlight the need for larger user studies to obtain more conclusive evidence.

Building on this call, the present work focuses on 2D $_{IN}$VRQ interfaces that dominate current practice [4, 30, 42] and narrows the methodological question to a practically relevant gap: whether markedly different usability/workload profiles across common 2D $_{IN}$VRQ tool implementations can meaningfully bias primary experiential self-reports when the VR stimulus is held constant.

3 Methods

Most empirical evidence on $_{IN}$VRQ effects has examined combinations of the VR stimulus, the self-report environment, and the questionnaire interface. Studies consistently report $_{IN}$VRQs to yield valid and reliable results comparable to $_{OUT}$VRQ baselines [4, 19, 37, 40, 43].

This leaves a gap in the literature regarding the isolated impact of different $_{IN}$VRQ interface designs on subjective measurements under a consistent stimulus (task and VE). This aligns with limitations and research directions articulated by Safikhani et al. [41], who call for larger user studies to obtain more conclusive evidence on $_{IN}$VRQ design effects. It also aligns with their subsequent taxonomy [42], which motivates practice-oriented comparisons along dominant interface dimensions. Prior work shows the sensitivity of experiential measures to context and task [9, 31]. Accordingly, this study isolates interface-induced differences using a balanced within-subject design that enables direct comparisons across $_{IN}$VRQ interfaces. An $_{OUT}$VRQ baseline was omitted due to prior evidence that $_{IN}$VRQs provide valid and reliable results comparable to $_{OUT}$VRQ baselines [4, 19, 40, 43].

This study evaluates three 2D $_{IN}$VRQ interfaces that represent dominant interface archetypes in current practice [4, 30, 42] and relevant to the research community (e.g., [27–29, 48, 58]): the Q-Watch (body-anchored, near-distance wristwatch, direct touch) [3], the ROVER Station (world-anchored, medium-distance rating station, handheld pointer) [30], and the VRQT Board (world-anchored, far-distance questionnaire board, laser pointer) [13]. Each tool combines multiple design choices. Although this prevents attributing effects to individual attributes, it enables a practice-oriented evaluation of the overall impact of these and similar interfaces. The selected $_{IN}$VRQ tools successfully instantiate high-frequency anchor- and interaction-categories in Safikhani et al.'s [42] taxonomy, thereby enabling a controlled comparison that directly operationalizes the taxonomy's dimensions for decision-making in VR study design. The mixed-methods study involved 44 participants and integrates subjective self-reports, objective interaction metrics, and qualitative feedback (see Fig. 2) to evaluate the different $_{IN}$VRQ interfaces and the following hypotheses:

- **H1** Perceived Usability differs between $_{IN}$VRQ interfaces.
  - **H1a** User characteristics moderate Usability ratings.
- **H2** Task Load differs between $_{IN}$VRQ interfaces.
- **H3** Subjective measurements related to the primary VR experience differ between $_{IN}$VRQ interfaces.
  - **H3a** Presence ratings differ between $_{IN}$VRQ interfaces.
  - **H3b** Flow ratings differ between $_{IN}$VRQ interfaces.

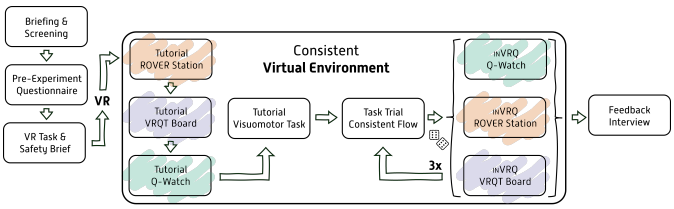

Fig. 2: Flow chart of the study procedure.

3.1 Technical Setup

The experimental setup included a high-performance VR system (NVIDIA RTX 4090) to ensure participant comfort at stable frame rates of 90 Hz. SteamVR was used as OpenVR runtime with four SteamVR 2.0 base stations in a 2.5 × 2.5 m tracking space. The Varjo Aero VR headset (focal distance: 85 cm) was selected for comfort and visual fidelity and was paired with Valve Index controllers enabling capacitive finger tracking. The VE integrating the $_{IN}$VRQ tools and the visuomotor task (see Fig. 1) was developed using the Unity 2021.3.14f1 game engine.

3.2 Preparation of $_{IN}$VRQ Tools

The tools (Q-Watch [3], ROVER [30], VRQT [13]) represent different 2D $_{IN}$VRQ interface designs (Station, Watch, Board, see Fig. 1). According to the taxonomy by Safikhani et al. [42], the Q-Watch exemplifies a body-anchored interface, while ROVER and VRQT exemplify world-anchored interfaces at medium and far distances, respectively.

The Q-Watch utilizes near-distance, direct touch interaction facilitated by hand tracking, whereas ROVER and VRQT use controller-based pointing. ROVER's interface features separation of interaction elements in the "touch zone" and display of text content at medium range for better readability. The Q-Watch implementation was provided by the author for research use. ROVER[1], a standalone application overlaid on the VE, was used in default configuration (version 2408r1). VRQT[2] was used in version 1.4.1.

Q-Watch and VRQT required minor modifications to meet experimental requirements in addition to regular Unity integration efforts. Logging of $_{IN}$VRQ interactions was also implemented for the two integrated tools. Additional tutorial sections and items were included in the Q-Watch and VRQT questionnaire specification files, based on ROVER's default tutorial. For better consistency and comparability, the scripts setting up the radio button interface layouts of VRQT and Q-Watch were modified slightly to accommodate 10-point Likert scales. For Q-Watch that resulted in a 70×30 cm canvas with buttons of 2.5×2.5 cm size attached to the left wrist. To make a selection, the tip of the virtual index finger of the right hand had to intersect with

[1] github.com/kuentzel/ROVER, accessed 01/26
[2] github.com/MartinFk/VRQuestionnaireToolkit, accessed 01/26

the button's collider, immediately counting as a press. The default Valve Index controller's capacitive sensor thresholds for finger curl were used. The original implementation of Q-Watch similarly relied on Quest controllers for interaction.

In line with the far-distance placement in its sample scene, the VRQT's Board, resembling a large floating screen, was positioned 5 m from the participant's default position. This led to an angular size of laser pointer interaction targets between $\approx 1.5^\circ$ and $\approx 1.85^\circ$ depending on the participant's height and movement from the default position. This is above the minimum of $\approx 1^\circ$ recommended by Microsoft for hand-ray/gaze targets [33]. Overall, the modifications to Q-Watch and VRQT were minimal and did not alter the tools' original design and interaction methods.

For replication, the supplemental materials include the Unity build of the VR application and the ROVER configuration files.

3.3 Study Design

The study aimed to identify within-subject differences in Perceived Usability, Task Load, Flow, and Presence ratings assessed at three levels (Watch, Station, Board). An a priori power analysis indicated a required sample size of 42 participants, assuming a moderate effect size of $f = 0.2$ based on prior studies [1, 4, 40].

Standardized scales were administered using the $_{IN}$VRQ tools to assess the primary dependent variables, similar to measurements in prior studies [1, 4, 16, 41]. Perceived Usability, reflecting users' subjective evaluations of a system's effectiveness and efficiency, was evaluated using the 5-point Likert scale of the widely used 10-item System Usability Scale (SUS) [7]. Task Load across multiple dimensions was measured as a mean score based on the first 9 items of the Simulation Task Load Index (SIM-TLX) [21], employing a 10-point Likert scale for various workload dimensions, including Mental Demand, Physical Demand, Temporal Demand, Frustration, and Task Control. Presence was assessed using the Slater-Usoh-Steed Presence Questionnaire (SUS-PQ) [50], calculated as the mean score of 6 items on a 7-point Likert scale. Flow experience was assessed using the 10 main items of the Flow Short Scale (FSS) [12] as a mean score on a 7-point Likert scale. Questionnaires were translated if validated translations were unavailable. Wording was adjusted to fit the context of the evaluation (e.g., "system/task" to "questionnaire tool").

Between interface evaluations, participants engaged in a reaction-based visuomotor task inspired by the classic arcade game "Whack-A-Mole" (see Fig. 1). The task served to standardize participants' activity between different $_{IN}$VRQ interface conditions and to minimize continuity and carry-over effects related to task engagement. The study employed on a single, simple task to isolate the effects of the $_{IN}$VRQ interfaces on subjective measurements in a tightly controlled experimental setting. This enabled a robust comparative analysis with high statistical power to test whether interface friction alone is sufficient to disrupt Flow or Presence under controlled conditions. Tool order was counterbalanced using a Latin square design.

The gamified task adapted to participants' performance through dynamic difficulty adjustment, modulating the challenge level to maintain stable engagement within the Flow corridor [11]. During the task and all $_{IN}$VRQ interactions, participants consistently remained in the same VE. A feasibility study with 9 student participants (4 women, 5 men, $M_{age} = 24.3$ years) confirmed the task elicited a consistent level of engagement, with moderate Presence ratings and low cybersickness. These findings indicated that, during the main experiment using the same task and environment, variability in self-reports would primarily reflect differences between $_{IN}$VRQ interfaces and individual participant characteristics rather than disengagement, discomfort, or environmental changes. Neither the presentation format nor the interaction mechanics of the task resembled any of the evaluated $_{IN}$VRQs, minimizing the risk of bias in interface evaluations due to prior task exposure.

Duration and other objective interaction metrics were recorded depending on the interface design. *Misclicks* were recorded for Station and Board, the latter also recording *Deselects*. *Misclicks* metrics were not recorded for the Watch due to the limited explanatory power offered in the context of Wolf et al.'s [56] observations when pointing with controllers. Semi-structured interviews were conducted to complement quantitative findings, providing qualitative insights into user preferences and issues with each $_{IN}$VRQ interface. Participants ranked the tools based on preference and provided suggestions for improvement. Items and interview guide are included in the supplemental materials.

3.4 Procedure

Following informed consent, participants filled in a questionnaire assessing demographics, Affinity for Technology Interaction (ATI) [49], and VR familiarity based on usage in the past 12 months. Fig. 2 outlines the study procedure. The detailed onboarding took approximately 10 minutes and included explanations of VR hardware, physical safety measures, and an overview of the procedure in VR. After acclimatization within the VE, participants completed interactive tutorials for the three $_{IN}$VRQ interfaces at their own pace, typically within 5 minutes.

Following 30 seconds of practice in the visuomotor task, participants engaged in three 2-minute task trials. After each task trial, participants self-reported their experience using one of the three $_{IN}$VRQ interfaces. They first completed task- and VE-related items (Flow, Presence), followed by gameplay-related filler items to extend interaction time to at least 5 minutes, and finally tool-related items (Usability, Task Load).

On average, participants spent around 30-35 minutes in VR, 20 minutes of which were spent interacting with the $_{IN}$VRQ tools. Cybersickness was monitored by adapting the Fast Motion Sickness Scale (FMS) [24] as a 10-point scale with predefined safety thresholds ($\geq$5: check-in, $\geq$8: termination) to ensure participant well-being.

Participants were asked about their experiences with the different interfaces in a post-experiment interview, usually concluding within 20-30 minutes. Overall, sessions were scheduled in time slots of up to 90 minutes, rarely exceeded by interviews with forthcoming participants.

3.5 Participants

Between November 2024 and January 2025, the study recruited 44 participants in Germany, including students, staff, and individuals from outside the institution, through lecture visits, mailing lists, and word-of-mouth. Each participant received €15 compensation. Prior approval was obtained from the institutional review board ("Ethikkommission FB Informatik") at Trier University of Applied Sciences.

Participants comprised 17 women, 26 men, and 1 diverse individual (self-identified), with ages ranging from 19 to 63 years ($M = 31.4(12.7)$). Most participants were right-handed (91%). Visual impairments such as nearsightedness (50%) or farsightedness (11%) were reported, but all participants could retain their vision correction under the VR headset. The majority were IT students (64%) or IT professionals (4%), while others were from non-IT professions (16%), other fields of study (7%), or unrelated occupations (9%). Affinity for Technology Interaction (ATI) scores ($M = 4.29(0.85)$) indicated a generally high comfort with technology [49]. VR familiarity was limited, with most participants reporting no (54%) or rare (32%) use of VR in the past year. Only a few used VR sometimes (5%), often (7%), or regularly (2%). Detailed characteristics provided in the supplemental materials.

3.6 Data Preparation and Analysis

The consolidated logs yielded a complete dataset ($n = 44$). One participant was excluded due to outlier ratings. Protocol notes indicated rushed self-reports and poor well-being (supported by FMS check-ins and post-experiment confirmation). The remaining participants stayed below the FMS thresholds ($M = 1.67(1.10)$).

Statistical analysis ($n = 43$) was conducted in R (version 4.5.1) [38]. Outcomes were analyzed using linear mixed-effects models (LMMs; `lme4/lmerTest`) with fixed effects of *interface* (Station, Watch, Board) and *order*, covariates (*age*, *gender*, *ATI*, *VR experience*), and random intercepts for participant *id*. Omnibus effects were tested via Type-III ANOVA (Wald F-tests) with Kenward-Roger degrees of freedom. Pairwise differences were estimated using EMMs (`emmeans`; Tukey-adjusted CIs and p-values). Practical significance was assessed with equivalence tests (TOST; 90% CIs) using scale-specific smallest effect sizes of interest (SESOI) bounds. Raw SESOI score bounds ranged around $\approx$5% of the respective scale and were defined based

on the observed standard deviations and results from similar studies (Cohen's d ≈0.2-0.3 as threshold for small effects) [1, 4, 25, 35, 40, 55]. Complementary Bayesian models (`brms`) quantified evidence for *interface* effects via Bayes Factors (`bridgesampling`) and ROPE inclusion of the 90% HDI using the same SESOI bounds.

Interview protocols and audio recordings were transcribed and subjected to a structured qualitative content analysis. Responses were paraphrased and deductively assigned to predefined categories based on theoretical expectations. In several rounds the initial coding was inductively refined and extended in a shared coding matrix to identify recurring patterns. To contextualize the quantitative results, thematic summaries and frequency-based comparisons were used to identify and characterize user experiences across tools.

4 Results

Raw means (M) and standard deviations (SD) for all primary outcomes are presented in Tab. 1. Internal consistencies of standardized scales were satisfactory (Cronbach's $\alpha > 0.7$). Higher SUS and lower SIM-TLX scores indicate better user experience with interfaces. Consistently moderate ratings of Presence and Flow across tools suggest an overall positive user experience with the task and VE. Tab. 2 shows estimated marginal means and omnibus tests for the effect of *interface*, controlling for order and user characteristics. R code, outputs, order-related probes, equivalence tests, robustness checks, and diagnostics are available in the supplemental materials. Diagnostics indicated approximate normality and no problematic multicollinearity for all primary measures. Frequentist residuals and posterior predictive checks showed ceiling (SUS) and floor (SIM-TLX) effects as expected given Likert scale constraints.

4.1 Perceived Usability (H1)

Model fit indices $R^2_m = .37$ and $R^2_c = .42$ indicate fixed effects account for the majority of explained variance. The omnibus test showed a statistically significant ($p < .001$) effect of the $_{IN}$VRQ interface condition on SUS ratings, confirming H1 (*Perceived Usability differs between $_{IN}$VRQ interfaces*). A Bayes factor of BF_{10}=3.73 × 10^7 supports this finding, indicating decisive evidence. H1a (*User characteristics moderate Usability ratings*) was not supported. Analysis found no moderating effects of demographic between-subject factors (age, gender, ATI, familiarity with VR).

The separation between ratings for the Station and the other interfaces (see Fig. 3) was confirmed in pairwise comparisons (see Tab. 5). The Station interface scored significantly higher than both the Watch ($p = .019$) and the Board interface ($p < .001$), with large practical differences >15 SUS points. Watch and Board ratings showed no significant difference ($p = .51$). Using a conservative raw SESOI of ±5 points (5% of scale range), the equivalence test for the contrast of Watch and Board remained inconclusive (see Tab. 5) with only 65% of the Bayesian posterior within the ROPE.

The omnibus test suggests an overall *order* effect ($F(2, 78.73) = 4.96$, $p = .009$), with mean scores decreasing across order positions. However, the significant *interface*×*order* interaction ($F(4, 109.37) = 2.98$, $p = .022$) indicates an interface-specific pattern. Station ratings remained stable and Watch ratings showed only a mild decline. In contrast, Board ratings declined markedly ($85.0 \rightarrow 69.4 \rightarrow 61.0$, $p_{12} = .017$, $p_{13} < .001$), suggesting that the overall order effect is primarily driven by the Board interface.

Table 1: Raw means and Cronbach's alpha for primary outcomes.

Measure	Station M (SD); α_{Cr}	Watch M (SD); α_{Cr}	Board M (SD); α_{Cr}
SUS	92.4 (8.1); .78	74.7 (17.1); .87	71.9 (18.7); .87
SIM-TLX	1.65 (0.65); .70	2.83 (1.39); .83	2.79 (1.59); .84
SUS-PQ	4.65 (1.43); .89	4.70 (1.48); .89	4.50 (1.56); .91
FSS	5.34 (1.02); .89	5.46 (0.95); .88	5.47 (0.80); .81

The omnibus test suggests a general *order* effect of ($F(2, 78.73) = 4.96$, $p = .009$) with mean scores decreasing across order positions overall. However, the significant *interface*×*order* interaction ($F(4, 109.37) = 2.98$, $p = .022$) indicating an interface-specific pattern. Station ratings remained stable and Watch ratings showed only a mild decline. In contrast, Board ratings showed a significant decline ($85.0 \rightarrow 69.4 \rightarrow 61.0$, $p_{12} = .017$, $p_{13} < .001$). Together, this suggests the absence of systematic drift (fatigue, learning) as the overall order effect is primarily driven by the Board interface.

Overall, the results indicate a clear usability advantage for Station over Watch and Board that remains after accounting for participant characteristics and presentation order.

Qualitative analysis of the feedback interviews identified several factors shaping usability perceptions. Participants strongly preferred the Station: 35/43 ranked it first and none ranked it last. Watch and Board split the remaining ranks (second: Watch 17/43, Board 18/43; third: Watch 21/43, Board 23/43). Participants described all interfaces as easy to learn due to simple interactions and familiar layouts. The Station was most often described as easiest to use (n=40), attributed to unobtrusive multimodal feedback (n=30) including visual hints, auditory signals, and haptic vibrations.

Watch feedback was mixed. Some noted the limited feedback beyond a visual cue (n=19), while others felt the direct touch interaction compensated (n=19). The Board received the most negative feedback, particularly regarding unclear or inconsistent feedback that caused confusion, irritation, and unintended deselections (n=34). Participants also reported difficulty with laser-pointer input, often leading to misclicks (n=26). Board logs showed slightly higher misclicks/deselects and shorter durations at later order positions (Tab. 3). Shorter durations may reflect faster responding, while stable error rates are compatible with fatigue or rushed responding, consistent with a speed-accuracy trade-off.

4.2 Task Load (H2)

Model fit indices $R^2_m = .22$ and $R^2_c = .30$ suggest moderate explanatory power of fixed effects over between-subject random effects. Confirming H2 (*Task Load differs between $_{IN}$VRQ interfaces*), the omnibus test showed a statistically significant ($p < .001$) effect of the $_{IN}$VRQ interface condition on SIM-TLX ratings. A Bayes factor of $BF_{10} = 4709$ supports this finding, indicating decisive evidence. Analysis finds no significant moderating effects of demographic between-subject factors on Task Load.

In line with SUS ratings, the data distribution in Fig. 3 suggests clear separation between the SIM-TLX ratings for the Station and the other interfaces, confirmed in pairwise comparisons (see Tab. 5): The Station interface was rated significantly lower than both the Watch ($p = .032$) and the Board interface ($p = .008$), with a meaningful practical advantage (Δ >10% of scale). Watch and Board ratings showed no significant difference ($p = .897$). Using a conservative raw SESOI of ±0.5 points (5% of scale range), the TOST did not support equivalence for Watch versus Board (Tab. 5).

However, the Bayesian ROPE analysis indicated 98.6% of the posterior within the ROPE, consistent with plausible practical equivalence. Watch and Station showed no evidence of linear drift. In contrast, trend analysis indicated that Task Load for Board increased with order position (slope= +0.68, 95% CI [0.20, 1.16], $p = .006$), consistent with the usability findings. A sensitivity model (variance by *interface*) suggests the absence of a systematic fatigue/learning effect and supports the conclusion that Station minimizes task load relative to the other interfaces. Taken together, perceived workload was lowest for Station. Watch and Board were similar on average, although Board workload tended to increase when presented later.

Observations during the study and the feedback interviews revealed distinct ergonomic challenges across the interfaces. Interface anchoring and placement affected readability. Twenty-eight participants (n=28) mentioned the need to extend their Watch arm for better readability, the canvas size leading to blurry text at the edges of their field of view. This led to frequent ergonomic criticism from participants (n=34) for necessitating awkward and tiring arm positions during interactions,

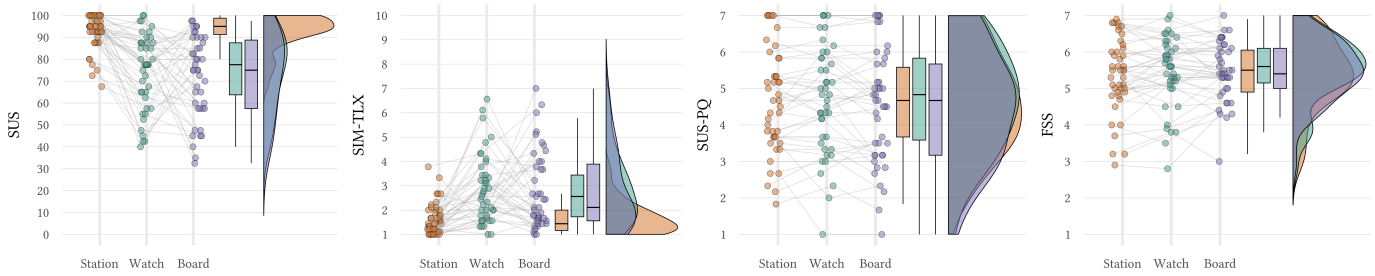

Fig. 3: Raincloud plots showing distribution of SUS (0-100), SIM-TLX (1-10), SUS-PQ (1-7), and FSS (1-7) ratings.

Table 2: Adjusted marginal means (95% CI) by interface, Type-III LMM ANOVA and Bayes factors for the effect of interface.

Measure	EMMs [95% CI] Station	Watch	Board	Type-III (interface) $F(df_1, df_2)$	p	BF_{10} (interface)
SUS (0-100)	91.1 [83.3, 99.0]	75.6 [67.7, 83.5]	69.4 [61.7, 77.1]	27.19 (2, 78.73)	$< .001^{***}$	3.73×10^{7}
SIM-TLX (1-10)	1.66 [0.98, 2.35]	2.92 [2.23, 3.60]	3.13 [2.47, 3.80]	13.27 (2, 77.70)	$< .001^{***}$	4709
SUS-PQ (1-7)	4.53 [4.01, 5.05]	4.66 [4.14, 5.18]	4.64 [4.12, 5.15]	1.97 (2, 73.00)	0.148	0.078
FSS (1-7)	5.35 [4.99, 5.71]	5.42 [5.06, 5.78]	5.53 [5.17, 5.88]	0.55 (2, 78.36)	0.580	0.049

Notes. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; EMMs/CIs from `emmeans` and Type-III tests from `lmerTest` (KR ddf). BF_{10} from (`brms`).

which resulted in discomfort and fatigue (n=38). In contrast, the Station was generally described as having good readability (n=26) and participants noted its ergonomic comfort (n=13) and ease of interaction (n=24). Participants praised the Board for providing a clear overview of questionnaire items (n=17) but criticized having to move their head side-to-side when reading (n=14). Although the Board did not inherently require extended arm movements, participants raised their arms for better accuracy, noting elevated physical effort (n=10).

Additionally, the Board was consistently mentioned by participants as the most difficult to control (n=26) due to its requirement for high pointing accuracy and concentration to mitigate frequent misclicks and unclear feedback, leading to higher mental effort (n=16) and frustration (n=34). Some participants (n=8) noted frustration due to discomfort and fatigue when using the Watch. The Station received only one report of frustration related to the transition between items and was praised by some participants (n=12) as the most time efficient.

Post-hoc comparisons were conducted using pairwise t-tests with Holm-Bonferroni correction to examine specific SIM-TLX subscales based on the interview results (see Tab. 4). As feedback indicated, Mental Demand ratings were significantly higher for the Board interface compared to the Station interface ($p < .001$), and for the Watch interface compared to the Station ($p = .003$).

Confirming feedback regarding ergonomic discomfort and fatigue, Physical Demand ratings were significantly higher for the Watch interface compared to both the Station ($p < .001$) and Board interfaces ($p < .001$). Participants also reported significantly lower Physical Demand for the Station compared to the Board ($p < .001$). Ratings of Temporal Demand did not differ significantly between tool interfaces, despite feedback indicating the Station was perceived as the most time-efficient interface.

In line with interview feedback, Frustration ratings were significantly higher for both the Board and the Watch interface compared to the Station ($p < .001$), reflecting participants' difficulties with pointing at the Board and the unergonomic handling of the Watch.

Ratings of Task Control Demand were significantly lower for the Station interface compared to both the Board ($p < .001$) and Watch ($p < .001$), indicating that participants felt the Station interface afforded easier interaction. Furthermore, participants reported significantly higher Task Control Demand for the Board's laser pointing compared to the Watch's more direct touch interaction ($p < .001$).

The post-hoc tests suggest that, relative to Station, Watch and Board increased Mental Demand and Frustration, Watch increased Physical Demand, and Station reduced Task Control Demand. Board also required higher Task Control than Watch.

Usability and Task Load ratings exhibit a high correlation across all interfaces ($p < .001$), suggesting that factors such as increased Physical Demand from the Watch's unergonomic placement and Frustration from the Board's laser pointer interaction significantly contribute to lower Usability ratings.

Descriptive analysis of interaction metrics supports these findings. A substantial number of *Misclicks* ($M = 78.4$) occurred with the Board, indicating higher interaction effort and reduced efficiency compared to the Station ($M = 0.19$). *Deselections* for the Board averaged 82.5, some of which may have been intentional, while most likely resulted from unclear interaction feedback.

In line with interview results, metrics show participants completed ratings faster using the Station ($M = 365\,(113)$) compared to the Watch ($M = 454\,(167)$) and Board ($M = 442\,(123)$).

An omnibus test (Type III ANOVA with KR ddf, controlling for participant characteristics) suggested overall interface differences

Table 3: Board performance metrics by order (raw means, SD).

Board	Duration [s]	Misclicks	Deselects
Order 1	528 (127)	68.7 (35.3)	82.2 (3.6)
Order 2	412 (94)	81.9 (28.7)	81.5 (4.0)
Order 3	388 (107)	84.4 (33.1)	83.9 (4.3)
ANOVA $F(2,40)$, p	6.53, $.004^{**}$	0.95, .394	1.38, .263

Notes. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 4: Post-hoc t-Tests for Task Load Items

SIM-TLX Subscales (1-10)	Watch -Station ΔM_p	Board -Station ΔM_p	Watch -Board ΔM_p
1 Mental D.	$1.00^{**}_{.003}$	$1.54^{***}_{<.001}$	$-0.54_{.243}$
2 Physical D.	$3.44^{***}_{<.001}$	$0.89^{*}_{.016}$	$2.56^{***}_{<.001}$
3 Temporal D.	$0.30_{.361}$	$0.44_{.331}$	$-0.14_{.634}$
4 Frustration	$1.14^{***}_{<.001}$	$1.67^{***}_{<.001}$	$-0.54_{.338}$
9 Task Control D.	$1.56^{***}_{<.001}$	$2.89^{***}_{<.001}$	$-1.33^{*}_{.030}$

Notes. $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$; Holm-adjusted.

($F(2,78) = 21.1$, $p < .001$), but pairwise Tukey comparisons were not individually significant. Overall, duration decreased sharply across orders ($F(2,78) = 43.7$, $p < .001$, linear downward trend), indicating a strong learning effect. No evidence was found for *interface* × *order*, suggesting a learning curve independent of presentation order.

4.3 Flow and Presence (H3)

Model fit indices for Presence ($R^2_m = .08$, $R^2_c = .94$) and Flow ($R^2_m = .18$, $R^2_c = .80$) suggest that most variance is explained by between-subject differences and that fixed effects contribute little. H3 (*Subjective measurements related to the primary VR experience differ between $_{IN}$VRQ interfaces*) was not supported. The omnibus tests showed no evidence of interface effects on Flow (FSS) or Presence (SUS-PQ) ratings. Bayes factors ($BF_{10} \approx 0.049$ for Presence; $BF_{10} \approx 0.078$ for Flow) provide moderate to strong evidence against models including *interface*. This is further confirmed by the data distribution indicating consistent ratings between interfaces (see Fig. 3).

EMMs show wide, overlapping 95% CIs (see Tab. 2) across interfaces and in pairwise comparisons (see Tab. 5). Analysis finds no evidence for order effects, like fatigue or drift, and no other moderating effects of user characteristics, except for a significant effect of higher ATI associated with slightly higher Flow ratings ($p = .034$, $\beta = 0.32$).

To test whether interface friction disrupted attentional continuity, additional models were fitted including within-interface centered usability (SUS_w) and workload (TLX_w) as predictors. Trial-level deviations in Usability and Task Load were not associated with changes in Flow ($p_{SUS} = .41$; $p_{TLX} = .62$) or Presence ($p_{SUS} = .20$; $p_{TLX} = .15$).

Using conservative raw SESOI bounds (around 5% of scale range) of ± 0.4 (SUS-PQ) and ± 0.3 (FSS), a TOST indicated equivalence between Watch and Board for Presence. EMM-based TOSTs for other pairs were inconclusive due to the larger uncertainty in the fitted models (covariate adjustment, KR ddf) driving 90% CIs across these conservative bounds (see Tab. 5). In contrast, TOSTs on the observed data (tighter CIs) indicate equivalence for all comparisons (see supplemental materials), consistent with Bayesian ROPE tests indicating >97% ROPE coverage of the posterior across all pairs.

This provides convincing evidence for plausible practical equivalence specific to the task-class. Reliability attenuation checks indicate that internal consistency (Cronbach's α; Tab. 1) would need to be very low ($< .36$) to mask a raw effect $\geq$ the SESOI bounds. Given the observed reliabilities (.81–.91; Tab. 1), the absence of differences is unlikely to be explained by scale unreliability.

Overall, this matches the study design's intention: Presence and Flow ratings reflect the consistent task/VE experience rather than interaction with the $_{IN}$VRQ interfaces and interface friction did not disrupt attentional continuity. The small observed differences are of limited practical relevance given pragmatic thresholds.

In line with the quantitative findings, qualitative feedback on the immersion of the interfaces integrated into the VE revealed modest perceived differences. All interfaces were seen as familiar, well-integrated, and suitable for the VE (n_S=30, n_W=25, n_B=21). However, some participants (n=15) described the Watch as the most immersive due to the natural feel of its direct touch interaction and its futuristic yet realistic appearance matching the VE.

5 Discussion

The analysis reveals meaningful differences in Perceived Usability and Task Load across the three investigated $_{IN}$VRQ tools, highlighting the importance of interface design choices. Interestingly, the evaluated tools did not differ in subjective measurements (Presence, Flow) related to the experimental task and VE. This evaluation of the interface attributes combined in the respective tools provides methodological guidance for designing and selecting $_{IN}$VRQs.

5.1 Tool Usability and Task Load

All $_{IN}$VRQ tools examined in this study received good to excellent Usability ratings overall (see [5]), yet the Station interface clearly outperformed both the Watch and Board, as reflected by significantly higher SUS ratings and significantly lower Task Load scores (see Tab. 1, Fig. 3). These high usability outcomes for the Station align with prior evaluations of ROVER that reported very high usability scores (> 90) in different study contexts, including with older adults [26, 29]. Likewise, the Watch's SUS scores remained in a similar range to the exploratory evaluation of Q-Watch reported by Al Kassm [2], suggesting that the observed ergonomic trade-offs are not unique to the present participant pool. Interview feedback supports these findings, with participants clearly ranking the Station first. In contrast, the Watch was consistently criticized for the physically demanding interaction with the wrist-attached interface, while the Board evoked frustrations due to laser pointing difficulties and inconsistent interaction feedback.

Task Load subscale analysis corroborated these themes, revealing elevated Physical Demand for the Watch as well as higher Mental Demand, Task Control Demand, and Frustration for the Board (see Tab. 4). The logged interaction metrics confirm the Station's reliable pointing interaction while revealing selection issues for the Board. Other studies [4,40,41,55] suggest similar challenges with laser pointer interaction. These are partly explained by Wolf et al.'s [56] description of the "Heisenberg Effect of Spatial Interaction": When pressing a button on the VR controller the slight change in the device's rotation can lead to the aim going off-target, particularly for smaller targets.

While interview feedback indicated the onset of physical and mental fatigue from using the Watch and the overall number of items, order effects in the analysis are driven by the Board interface. This suggests the abscence of systematic drift (fatigue or familiarization) and a well-tolerated balance of overall questionnaire duration.

The decline in Board ratings across order positions is likely to reflect an effect of comparative exposure: when experienced after other interfaces, participants more often attributed difficulties (e.g., laser pointer precision, inconsistent feedback) to the design itself. This led to lower usability scores for the Board. This pattern can amplify interface weaknesses: once participants have a benchmark for how usable interaction in VR can be, problematic designs are penalized more strongly. This would suggest that usability issues become more salient with higher VR familiarity. However, analysis found no significant moderation by self-reported VR familiarity levels, general ATI, age, or gender on usability ratings. While this suggests that the observed effects are consistent across user profiles, only few participants in this study reported high familiarity with VR, limiting generalizability.

Watch and Board did not differ significantly in pairwise tests. Equivalence tests were inconclusive, leaving open subtle differences in perceived usability. Interview rankings and raincloud plots indicate heterogeneous preferences, but the main drivers remain unclear (e.g., physically demanding Watch handling versus frustrating Board pointing, Watch immersion versus Board familiarity and readability). Metrics showed a clear difference in questionnaire completion times between interfaces with the Station enabling significantly more time-efficient self-reporting. Though the SIM-TLX comparisons of Temporal Demand did not corroborate this difference, participants did highlight the Station's reliable, simple, and quick interaction compared to Watch and Board. Despite the limited effect on other subjective measurements, the overall results affirm that ergonomic design of $_{IN}$VRQs strongly influences test efficiency, participant comfort, and frustration levels during VR studies.

5.2 Flow and Presence Measurements

Despite the large usability differences, no significant effects were observed on Presence or Flow ratings across interface conditions. Equivalence tests with pragmatic but conservative bounds confirmed that differences across tools were practically negligible. Raw score differences $\leq$5% of scale ranges are unlikely to reflect a meaningful experiential change. Bayesian analysis provides moderate evidence supporting this interpretation (see Tab. 2). No significant effect of order on Presence or Flow ratings was found, suggesting successful counterbalancing and appropriate study duration without systematic fatigue or familiarization.

Mid-to-high Presence or Flow ratings, together with reasonable variance, indicate that the combination of a simple, repeating task and consistent VE resulted in a comparable VR experience across condi-

Table 5: Pairwise comparisons for SUS, TLX, SUS-PQ, and FSS. Estimates are mean differences. TOST and ROPE equivalence tests.

		Pairwise comparisons				Equivalence			
Measure	**Contrast**	**Estimate**	**95% CI**	**90% CI**	p	p_{low}	p_{high}	**TOST**	**ROPE%**
SUS	Station - Watch	15.50	[2.08, 28.92]	[3.78, 27.21]	.019**	.967	<.001	not eq.	0%
(0-100)	Station - Board	21.72	[8.55, 34.90]	[10.22, 33.23]	<.001***	<.001	.998	not eq.	0%
SESOI ±5	Watch - Board	6.23	[-7.22, 19.67]	[-5.52, 17.97]	.517	.025	.585	inconcl.	65%
TLX	Station - Watch	-1.25	[-2.41, -0.09]	[-2.27, -0.24]	.032*	<.001	.936	not eq.	0%
(1-10)	Station - Board	-1.47	[-2.61, -0.33]	[-2.47, -0.47]	.008**	.977	<.001	not eq.	0%
SESOI ±0.5	Watch - Board	-0.22	[-1.38, 0.95]	[-1.23, 0.80]	.897	.283	.073	not eq.	98.6%
SUS-PQ	Station - Watch	-0.13	[-0.61, 0.35]	[-0.54, 0.29]	.791	.005	.089	inconcl.	100%
(1-7)	Station - Board	-0.11	[-0.58, 0.36]	[-0.52, 0.30]	.849	.070	.006	inconcl.	100%
SESOI ±0.4	Watch - Board	0.02	[-0.45, 0.49]	[-0.39, 0.43]	.993	.018	0.030	eq.	99.8%
FSS	Station - Watch	-0.07	[-0.56, 0.42]	[-0.50, 0.36]	.940	.038	.134	inconcl.	100%
(1-7)	Station - Board	-0.18	[-0.66, 0.31]	[-0.60, 0.25]	.658	.276	.011	inconcl.	97%
SESOI ±0.3	Watch - Board	-0.11	[-0.60, 0.38]	[-0.53, 0.32]	.856	.177	.025	inconcl.	100%

Notes. $*p < .05$, $**p < .01$, $***p < .001$. EMMs, CIs, and TOST from `emmeans` (Tukey-adjusted, KR ddf); *ROPE* from `bayestestR`.

tions. The visuomotor task maintained stable engagement levels across trials, providing a standardized experiential context for comparing different interface conditions. Flow experience hinges on a delicate balance between challenge and skill [12] yet appeared robust to differences in Task Control demand and Frustration during self-reporting with the evaluated $_{IN}$VRQ interfaces.

Variability in Presence and Flow ratings was primarily associated with individual participant characteristics, with Flow showing comparatively lower variance due to the task structure, whereas Presence reflected a more subjective appraisal of the virtual environment. Consistent with prior work, Flow and Presence are known to be sensitive to task context. In contrast to more immersive and extensive VR scenarios, the present task imposed minimal cognitive and emotional stakes. In scenarios, where attentional continuity and emotional engagement are more critical, interface friction may exert stronger disruptive effects on Flow and Presence.

Qualitative feedback suggested that all interfaces were perceived as reasonably well integrated into the virtual environment, with some participants describing the Watch interface as particularly immersive. However, these subjective impressions did not translate into systematic differences in ratings. This supports the notion that visual or interaction congruence of $_{IN}$VRQs with the task and environment has a small or negligible effect on measurement validity [19,41].

If the $_{IN}$VRQ interfaces introduced BIPs, it is plausible that these occurred in a comparable manner across conditions, as breaks in presence during in-VR self-reporting are likely driven by increased awareness of the experimental situation itself rather than by specific interface characteristics [37]. The results suggest that key experiential self-reports are likely to be robust against BIPs and variations in $_{IN}$VRQ design, which aligns with prior findings highlighting minimal differences across interface variations and $_{IN}$VRQ / $_{OUT}$VRQ conditions [4,19,40,41,43].

Overall, the findings offer methodological reassurance for employing $_{IN}$VRQs in controlled behavioral studies. Given adequate usability of an $_{IN}$VRQ interface, sensitive experiential ratings like presence and flow are likely to remain reliable even under ergonomic trade-offs. However, ethical considerations regarding participant comfort (fatigue, frustration) should influence $_{IN}$VRQ design and careful interface evaluation remains essential when extending such tools to scenarios with higher requirements for attentional continuity and engagement.

5.3 $_{IN}$VRQ Interface Design

Usability and ergonomics of a given $_{IN}$VRQ interface are likely dependent on the context of the task and VE. This includes scenarios such as multi-user VR and whether the user is moving around and interacting with objects in varied ways. The present evaluation should be interpreted as an assessment of three concrete tool instances as deployed by researchers "out of the box," rather than as a test of each interface archetype in an ergonomically optimized configuration. While other creative design solutions, such as body-anchored [3,40], intradiegetic [41], or gaze-based interfaces [36] have emerged, laser pointer interaction remains most common in VR software (industry and academia). The cause for this preference is likely found in its good balance between usability, familiarity, and immersion [4,30,41].

Based on this study's results, guidelines for $_{IN}$VRQ interfaces are provided, consistent with prior recommendations [4,42] and meant to complement universal usability principles.

When presenting $_{IN}$VRQs, readability should be a major concern for validity, test efficiency, and participant comfort [47]. The Q-Watch interface affords very direct, natural touch interaction. However, the bi-manual handling of the Watch requires a higher level of control as it is physically demanding to position and keep the content display steady for better readability. In contrast, the placement of the Board at a far distance beyond the focal distance of the HMD (see vergence-accommodation conflict [6]) facilitates good readability but renders the laser pointer interaction more difficult. In both cases, participants commented on text not being the right size or being blurry at the edges of their field of view, requiring side-to-side head movements.

Additionally, the low focal distance of the Varjo Aero (85 cm, 1.5-2 m is more common) might have mitigated readability issues on the Watch. Whether the interface uses direct touch or pointing at range, designs should not only consider angular size but also relative size and distance between interaction elements, as if applying Fitts' Law to spatial interactions [53]. Overall, the findings highlight ergonomic trade-offs and the need for thorough design considerations when integrating $_{IN}$VRQ tools like VRQT or Q-Watch into a study VE.

The design of the ROVER Station [30] addresses these issues by decoupling the control panel in the interaction zone from the content display The design allows ROVER's Station to accommodate users' individual ergonomic needs. Font values are based on various recommendations from literature (see [47]). Distance and height of the control panel are independently adjustable [30]. Additionally, button size dynamically scales with the questionnaire scale range while also implementing vertically extending collider zones on the buttons to address the "Heisenberg Effect of Spatial Interaction" (see [56]).

The study evaluated the $_{IN}$VRQ tools largely in their default configurations, with only minor modifications required for the experiment. This reflects how teams with limited resources may deploy existing tools without extensive ergonomic refinement. Deployability is a practical constraint: ROVER operates as an overlay alongside the study application, reducing integration effort. However, this approach depends on SteamVR (e.g., via Oculus Air Link or Steam Link), which can limit seamless integration into an experiment's VE. In contrast, VRQT and Q-Watch integrate directly into Unity scenes and are easier to customize, but they require more technical expertise. Several us-

ability issues observed in this study could be mitigated with additional implementation effort.

Quantitative and qualitative analysis provide convincing evidence that the Watch's default body-anchored interface and the Board's long-distance pointer interaction are the isolated design aspects most strongly impacting usability and task load. These trade-offs are not intrinsic consequences of the underlying interface archetypes [42]. Instead, they reflect suboptimal default configurations (Watch with 70×30 cm canvas, 2.5×2.5 cm buttons; Board at 5 m distance, button angular size between $\approx 1.5^\circ$ and $\approx 1.85^\circ$). Practical factors such as display scale, default wrist placement, and the lack of a convenient posture for "resting" the arm can be addressed through targeted ergonomic refinements of Watch-like interfaces (e.g., scaling, alternative anchoring points, or optional world-anchoring for longer questionnaires).

Inconsistent angular target sizes, small tolerance margins, and inconsistent feedback all amplify the "Heisenberg Effect of Spatial Interaction" and increase mis-selections in Board-like interfaces. An improved implementation of the VRQT Board could enforce fixed angular size for text and buttons, apply hover-based dwell or "magnetic" stabilization, and provide unambiguous multimodal confirmation of selection and de-selection. While the feedback of the Q-Watch and VRQT's Board was perceived as weak or inconsistent, participants noted ROVER's multimodal feedback was subtle and not intrusive. Combined with the forgiving interaction design, this was likely the main contributing factor to its higher perceived usability Individual adjustability of feedback density and strength, such as vibration intensity or volume of audio cues, should be provided on top of standard accessibility features and ergonomic considerations [8, 57].

Based on their results and user preference, prior studies recommend intradiegetic [52], world-anchored [40], familiar 2D-layout [41] interfaces with laser pointer interaction [4] within the VE of the task situation [40]. While $_{OUT}$VRQs remain a valid alternative, $_{IN}$VRQs are more efficient and reduce participant burden by minimizing interruptions and headset fittings in study designs using repeated sampling and in situ measurement. ROVER and VRQT provide researchers with valid, publicly available, well-documented $_{IN}$VRQ solutions based on recommendations from the literature, reducing the effort required for VR study setups, particularly for non-technical teams [13, 30].

ROVER is specifically recommended for research scenarios prioritizing plug-and-play integration, especially for teams with less technical expertise or implementation resources. In contrast, VRQT is recommended when flexible integration within the VE, custom appearance, or usage in standalone VR (e.g., Meta Quest, Pico) is needed. VRQT has proven to be a reliable tool across many studies [13,58], and ROVER appears particularly usable for VR beginners and vulnerable participants based on participant feedback, ratings, and prior studies [26, 29, 30].

Given resources and expertise for implementation, custom $_{IN}$VRQ solutions are highly encouraged, depending on the context of the study. Despite potential ergonomic challenges, body-anchored $_{IN}$VRQ interfaces like the Q-Watch could be preferred when study designs call for frequent, short self-reports while "on-the-move" or when immersive integration into the study VE is paramount.

5.4 Limitations and Further Work

Several limitations should be considered when interpreting these findings. The sample was drawn largely from an academic setting and included many IT-affiliated participants, with relatively high ATI scores. This may limit generalizability to broader or less experienced populations. However, sample size and composition are mostly comparable to related studies yet feature a slightlty more representative age range and lower VR familiarity among participants (compare overviews in [30, 42]). Additionally, the Station's high usability ratings observed in this study are consistent with prior ROVER evaluations in more complex interaction contexts, including with older adults of lower technology affinity [26, 29]. This suggests that the Station's usability advantage is unlikely to be driven solely by the simple task-design and present participant characteristics.

The study employed only Likert-type radio button response formats. More complex formats (e.g., sliders, ranking, or text input) may interact differently with interface design, potentially increasing cognitive or physical demands. Additionally, input modalities such as ray casting or gesture-based interaction may impose greater motor demands. This could disproportionately affect users with limited mobility or motor impairments and potentially result in different usability profiles [8, 29]. Future research should examine $_{IN}$VRQ interfaces from an inclusive or ability-based design perspective, while recruiting more diverse participant samples to capture a broader range of user needs and capabilities.

The findings are based on a tightly controlled experimental design employing a single, simple visuomotor task and a static VE. This design choice strengthens internal validity and statistical power but limits generalizability to more complex VR scenarios. The repetitive task may have reduced sensitivity to subtle interface-induced disruptions of Presence and Flow. Different scenarios, especially those involving narrative progression, emotional engagement, or high stakes, may amplify the impact of interface friction on attentional continuity.

While the tool-level evaluation was deliberately constrained to the present task-class, further work should systematically vary interaction modalities and question formats to isolate the impact of specific design features. Evaluations should also span diverse VR contexts (e.g., immersive storytelling, rehearsal or training simulations, therapeutic applications, and multi-user environments) to identify when reduced usability or increased interface complexity begins to compromise measurement validity.

6 Conclusion

This study systematically compared user experience across three $_{IN}$VRQ interfaces under a controlled task and VE. Interface design meaningfully influenced perceived usability and comfort during in-VR self-reporting. Despite significant differences in Usability and Task Load across the three evaluated interfaces, Presence and Flow ratings were not meaningfully affected by interface variations.

Under a consistent task and VE, variability in these experiential measures was primarily attributable to individual differences rather than interface effects. Taken together, the findings indicate that, under low-risk, non-narrative task conditions, Flow and Presence ratings are robust to $_{IN}$VRQ interface differences once basic usability requirements are met. However, more immersive or high-stakes VR scenarios may increase sensitivity to interface friction, warranting careful evaluation when extending $_{IN}$VRQs to such contexts.

All evaluated $_{IN}$VRQ interfaces achieved acceptable usability and supported reliable self-reports, but participants showed a clear preference for the open-source ROVER Station [30]. Reported reasons included forgiving interactions, high readability, and adjustable ergonomics. Employing $_{IN}$VRQ tools can enhance test efficiency and participant comfort, particularly in study designs with repeated sampling and for VR novices.

Overall, this work provides practice-oriented guidance for selecting and implementing $_{IN}$VRQ tools, reducing barriers for non-technical research teams and supporting comparability and reproducibility in VR research.

Supplemental Materials

Supplemental materials, which include data, R code, additional analysis, scale items, interview guide, and VR build files, are available at osf.io/528yk (CC BY 4.0).

Acknowledgments

This work was funded in part by the Ministry for Science and Health of Rhineland-Palatinate as part of the research training group (Forschungskolleg) "Immersive Extended Reality for Physical Activity and Health" (XR-PATH).

AI systems were used for proofreading, editing, and R code.

References

[1] P. Acevedo, A. L. Jimenez, A. J. Magana, B. Benes, and C. Mousas. An Exploration of the Effects of in-VR Assessment Format on User Performance and Experience. In S. Hasegawa, N. Sakata, and V. Sundstedt, eds., *ICAT-EGVE 2024 - International Conference on Artificial Reality*

and Telexistence and Eurographics Symposium on Virtual Environments, pp. 1–10. The Eurographics Association, Tsukuba, Japan, 2024. doi: 10.2312/egve.20241370 2, 4, 5

[2] N. M. Al Kassm. The virtual reality questionnaire watch: Exploring novel methods integrated with google forms. Master's thesis, Carleton University, Ottawa, ON, Canada, 2022. doi: 10.22215/etd/2022-15151 3, 7

[3] N. M. Al Kassm, J. Henderson, and R. J. Teather. Administering vr questionnaires generated in google forms. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*, SUI '24, article no. 32, 2 pages. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3677386.3688888 1, 2, 3, 8

[4] D. Alexandrovsky, S. Putze, M. Bonfert, S. Höffner, P. Michelmann, D. Wenig, R. Malaka, and J. D. Smeddinck. Examining design choices of questionnaires in vr user studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 21 pages, pp. 1–21. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376260 1, 2, 3, 4, 5, 7, 8, 9

[5] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Stud.*, 4:114–123, 04 2009. 7

[6] A. U. Batmaz, M. D. Barrera Machuca, J. Sun, and W. Stuerzlinger. The Effect of the Vergence-Accommodation Conflict on Virtual Hand Pointing in Immersive Displays. In S. Barbosa, ed., *CHI Conference on Human Factors in Computing Systems*, ACM Digital Library, pp. 1–15. Association for Computing Machinery, New York,NY,United States, 2022. doi: 10.1145/3491102.3502067 8

[7] J. Brooke. *"SUS-A quick and dirty usability scale." Usability evaluation in industry*, pp. 189–194. Taylor & Francis, London, UK, 1996. 3, 4

[8] J. A. Brown, A. T. Dinh, and C. Oh. Safety and ethical considerations when designing a virtual reality study with older adult participants. In Q. Gao and J. Zhou, eds., *Human Aspects of IT for the Aged Population. Design, Interaction and Technology Acceptance*, pp. 12–26. Springer International Publishing, Cham, 2022. 2, 9

[9] T. J. Cahill and J. J. Cummings. Effects of congruity on the state of user presence in virtual environments: Results from a breaching experiment. *Frontiers in Virtual Reality*, Volume 4 - 2023, 2023. doi: 10.3389/frvir.2023.1048812 2, 3

[10] J. F. P. Cheiran, D. R. Bandeira, and M. S. Pimenta. Measuring the key components of the user experience in immersive virtual reality environments. *Frontiers in Virtual Reality*, Volume 6 - 2025:1–15, 2025. doi: 10.3389/frvir.2025.1585614 2

[11] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, NY, March 1991. 4

[12] S. Engeser and F. Rheinberg. Flow, performance and moderators of challenge-skill balance. *Motivation and emotion*, 32:158–172, 2008. 2, 4, 8

[13] M. Feick, N. Kleer, A. Tang, and A. Krüger. The virtual reality questionnaire toolkit. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, 2 pages, p. 68–69. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3379350.3416188 1, 2, 3, 9

[14] K. Finstad. The usability metric for user experience. *Interacting with Computers*, 22(5):323–327, 2010. Modelling user experience - An agenda for research and practice. doi: 10.1016/j.intcom.2010.04.004 3

[15] M. Gottsacker, N. Norouzi, K. Kim, G. Bruder, and G. Welch. Diegetic representations for seamless cross-reality interruptions. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 310–319, 2021. doi: 10.1109/ISMAR52148.2021.00047 2

[16] S. Graf and V. Schwind. Inconsistencies of presence questionnaires in virtual reality. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, VRST '20, article no. 60, 3 pages. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3385956.3422105 2, 3, 4

[17] J. P. Gründling, N. Feld, D. Zielasko, and B. Weyers. Correlations of flow, usability, workload, and presence with task performance in a spatially distributed memory task. In *Virtual Reality and Mixed Reality: 20th EuroXR International Conference, EuroXR 2023, Rotterdam, The Netherlands, November 29 – December 1, 2023, Proceedings*, 13 pages, p. 153–165. Springer-Verlag, Berlin, Heidelberg, 2023. doi: 10.1007/978-3-031-48495-7_10 2

[18] J. P. Gründling and B. Weyers. Immersive analytics: The influence of flow, sense of agency, and presence on performance and satisfaction. *Proc. ACM Hum.-Comput. Interact.*, 8(EICS), article no. 254, 27 pages, June 2024. doi: 10.1145/3661144 2

[19] J. P. Gründling, D. Zeiler, and B. Weyers. Answering with bow and arrow: Questionnaires and vr blend without distorting the outcome. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 683–692. IEEE, New York, USA, 2022. doi: 10.1109/VR51125.2022.00089 2, 3, 8

[20] S. Guertin-Lahoud, C. K. Coursaris, S. Sénécal, and P.-M. Léger. User experience evaluation in shared interactive virtual reality. *Cyberpsychology, Behavior, and Social Networking*, 26(4):263–272, April 2023. doi: 10.1089/cyber.2022.0261 2

[21] D. Harris, M. Wilson, and S. Vine. Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). *Virtual Reality*, 24(4):557–566, 2020. doi: 10.1007/s10055-019-00422-9 3, 4

[22] D. Hepperle, T. Dienlin, and M. Wölfel. Reducing the human factor in virtual reality research to increase reproducibility and replicability. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 100–105. IEEE, New York, USA, 2021. doi: 10.1109/ISMAR-Adjunct54149.2021.00030 1

[23] D. Hepperle and M. Wölfel. Similarities and differences between immersive virtual reality, real world, and computer screens: A systematic scoping review in human behavior studies. *Multimodal Technologies and Interaction*, 7(6):56, 2023. doi: 10.3390/mti7060056 1

[24] B. Keshavarz and H. Hecht. Validating an efficient method to quantify motion sickness. *Human factors*, 53(4):415–426, 2011. doi: 10.1177/0018720811403736 4

[25] T. Kojić, M. Vergari, S. Knuth, M. Warsinke, S. Möller, and J.-N. Voigt-Antons. Influence of gameplay duration, hand tracking, and controller based control methods on ux in vr. In *Proceedings of the 16th International Workshop on Immersive Mixed and Virtual Environment Systems*, MMVE '24, 7 pages, p. 22–28. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3652212.3652222 5

[26] L. Küntzer and G. Rock. Beyond the headset: Towards a practical integrative framework for measuring user experience in immersive virtual reality environments. Available at SSRN, 2024. SSRN: https://ssrn.com/abstract=5011256, http://dx.doi.org/10.2139/ssrn.5011256. doi: 10.2139/ssrn.5011256 2, 3, 7, 9

[27] L. Küntzer and G. Rock. XR-CISE: Towards Promoting Physical Activity with Inclusive Virtual Reality Exergaming. In *Transdisciplinary Engineering for Social Change, Proceedings of the 31st ISTE International Conference on Transdisciplinary Engineering*, pp. 1–10. IOS Press, London, United Kigdom, 2024. 3

[28] L. Küntzer, M. Scherer, T. Mentler, and G. Rock. Dynamic difficulty adjustment in virtual reality exergaming to regulate exertion levels via heart rate monitoring. In *30th ACM Symposium on Virtual Reality Software and Technology*, VRST '24, article no. 82, pp. 1–2. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3641825.3689504 3

[29] L. Küntzer, S. Schwab, H. Spaderna, and G. Rock. Measuring User Experience of Older Adults during Virtual Reality Exergaming. In *Proceedings of the 16th International Conference on Quality of Multimedia Experience (QoMEX 2024)*, pp. 1–7. IEEE, Karlshamn, Sweden, 2024. 2, 3, 7, 9

[30] L. Küntzer, S. Schwab, H. Spaderna, and G. Rock. ROVER: A Standalone Overlay Tool for Questionnaires in Virtual Reality. In *EICS '24 Companion: Companion Proceedings of the 2024 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pp. 1–8. ACM, Cagliari, Italy, 2024. 1, 2, 3, 8, 9

[31] S. J. Lackey, J. N. Salcedo, J. Szalma, and P. Hancock. The stress and workload of virtual reality training: the effects of presence, immersion and flow. *Ergonomics*, 59(8):1060–1072, 2016. doi: 10.1080/00140139.2015.1122234 2, 3

[32] W. Liu, H. Hu, C. Zhou, Y. Bian, and J. Liu. Exploring the role of distraction in weak flow–performance link based on vr searching tasks. *Applied Sciences*, 11(13), 2021. doi: 10.3390/app11135799 2

[33] Microsoft. Interactable object design in mixed reality. https://learn.microsoft.com/en-us/windows/mixed-reality/design/interactable-object, n.d. Accessed: 2025-09-11. 4

[34] H. Moosbrugger and A. Kelava. *Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien)*, pp. 7–26. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-20072-4_2 2

[35] L. T. D. Paolis and V. D. Luca. The effects of touchless interaction on usability and sense of presence in a virtual environment. *Virtual Reality*, 26(4):1551–1571, 2022. doi: 10.1007/s10055-022-00647-1 5

[36] K. Pfeuffer, L. Mecke, S. Delgado Rodriguez, M. Hassib, H. Maier, and

F. Alt. Empirical evaluation of gaze-enhanced menus in virtual reality. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, VRST '20, article no. 20, 11 pages. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3385956.3418962 2, 8

[37] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka. Breaking the experience: Effects of questionnaires in vr user studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 15 pages, p. 1–15. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376144 1, 2, 3, 8

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. 4

[39] G. Regal, R. Schatz, J. Schrammel, and S. Suette. Vrate: A unity3d asset for integrating subjective assessment questionnaires in virtual environments. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, Cagliari, Italy, 2018. doi: 10.1109/QoMEX.2018.8463296 3

[40] G. Regal, J.-N. Voigt-Antons, S. Schmidt, J. Schrammel, T. Kojic, M. Tscheligi, and S. Möller. Questionnaires embedded in virtual environments: reliability and positioning of rating scales in virtual environments. *Quality and User Experience*, 4(1):1–13, 2019. doi: 10.1007/s41233-019-0029-1 1, 2, 3, 4, 5, 7, 8, 9

[41] S. Safikhani, M. Holly, A. Kainz, and J. Pirker. The influence of in-vr questionnaire design on the user experience. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, article no. 12, 8 pages. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3489849.3489884 1, 2, 3, 4, 7, 8, 9

[42] S. Safikhani, L. Nacke, and J. Pirker. A literature review and taxonomy of in-vr questionnaire user interfaces. In J. M. Krüger, D. Pedrosa, D. Beck, M.-L. Bourguet, A. Dengel, R. Ghannam, A. Miller, A. Peña-Rios, and J. Richter, eds., *Immersive Learning Research Network*, pp. 95–111. Springer Nature Switzerland, Cham, 2025. doi: 10.1007/978-3-031-80475-5_7 1, 2, 3, 8, 9

[43] V. Schwind, P. Knierim, N. Haas, and N. Henze. Using presence questionnaires in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 12 pages, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300590 1, 2, 3, 8

[44] L. Sidenmark and H. Gellersen. Eye&head: Synergetic eye and head movement for gaze pointing and selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, 14 pages, p. 1161–1174. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3332165.3347921 2

[45] M. Slater. How colorful was your day? why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 13:484–493, 2004. 2

[46] M. Slater and M. V. Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:1–47, 2016. doi: 10.3389/frobt.2016.00074 1

[47] T. Kojić, D. Ali, R. Greinacher, S. Möller, and J. -N. Voigt-Antons. User Experience of Reading in Virtual Reality — Finding Values for Text Distance, Size and Contrast. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, Athlone, Ireland, 2020. doi: 10.1109/QoMEX48832.2020.9123091 2, 8

[48] R. Tamaki and T. Nakajima. Shoot down drones with your answer, an integration of a questionnaire into a vr experience. In *Proceedings of the 2021 ACM Symposium on Spatial User Interaction*, SUI '21, article no. 24, 2 pages. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3485279.3488282 2, 3

[49] C. A. Thomas Franke and D. Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2019. doi: 10.1080/10447318.2018.1456150 4

[50] M. Usoh, E. Catena, S. Arman, and M. Slater. Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments*, 9(5):497–503, 10 2000. doi: 10.1162/105474600566989 2, 4

[51] S. Vlahovic, M. Suznjevic, and L. Skorin-Kapov. A survey of challenges and methods for quality of experience assessment of interactive VR applications. *Journal on Multimodal User Interfaces*, 16(3):257–291, 9 2022. doi: 10.1007/s12193-022-00388-0 1

[52] N. Wagener, M. Stamer, J. Schöning, and J. Tümler. Investigating effects and user preferences of extra- and intradiegetic virtual reality questionnaires. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, VRST '20, article no. 23, 11 pages. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3385956.3418972 1, 2, 9

[53] U. Wagner, M. N. Lystbæk, P. Manakhov, J. E. S. Grønbæk, K. Pfeuffer, and H. Gellersen. A Fitts' Law Study of Gaze-Hand Alignment for Selection in 3D User Interfaces. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson, eds., *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, New York, NY, USA, 2023. doi: 10.1145/3544548.3581423 8

[54] P. Wang and J. N. Bailenson. Virtual reality as a research tool. In T. Reimer, L. van Swol, and A. Florack, eds., *The Routledge Handbook of Communication and Social Cognition*, pp. 1–25. Routledge/Taylor and Francis, New York, NY, 2024. In Press. doi: 10.2139/ssrn.4805041 1

[55] X. Wei and Y. Li. Evaluating user performance, workload, and presence of virtual reality questionnaires using joystick and raycasting selection techniques. In *Proceedings of the 2023 7th International Conference on Virtual and Augmented Reality Simulations*, ICVARS '23, 6 pages, p. 29–34. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3603421.3603426 1, 2, 5, 7

[56] D. Wolf, J. Gugenheimer, M. Combosch, and E. Rukzio. Understanding the heisenberg effect of spatial interaction: A selection induced error for spatially tracked input devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 10 pages, p. 1–10. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376876 2, 4, 7, 8

[57] World Wide Web Consortium, Accessible Platform Architectures Working Group. XR Accessibility User Requirements, W3C Working Group Note xaur-20210825, 2021. 9

[58] A. Zenner, K. Ullmann, and A. Krüger. Combining dynamic passive haptics and haptic retargeting for enhanced haptic feedback in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2627–2637, 2021. doi: 10.1109/TVCG.2021.3067777 3, 9

[59] A. Zheleva, L. De Marez, D. Talsma, and K. Bombeke. Intersecting realms: a cross-disciplinary examination of VR quality of experience research. *Virtual Reality*, 28(3):135, 7 2024. doi: 10.1007/s10055-024-01031-x 1