



Usability and UX Evaluation in a Mixed Reality Puzzle Game Using Questionnaires

Thiago Prado de Campos   [Federal Technological University of Paraná | contact@thiagotpc.com]

Saul Delabrida  [Federal University of Ouro Preto | saul@sdelabrida.com]

Natasha Malveira Costa Valentim  [Federal University of Paraná | natasha@inf.ufpr.br]

 Federal University of Technology of Paraná, Av. João Miguel Caram 731, Pioneiros, Londrina, PR, 86036-700, Brazil.

Received: 02 June 2025 • Accepted: 24 July 2025 • Published: 10 August 2025

Abstract: *Background:* Mixed Reality (MR) technologies present unique usability and user experience (UX) challenges, particularly in gesture-based interaction and spatial perception. Puzzle games in MR environments rely heavily on intuitive control and immersive feedback to maintain user engagement. *Purpose:* This study aims to evaluate the usability and UX of the Cubism puzzle game in an MR context, comparing different standardized instruments to capture user perceptions and identify specific interaction issues. *Methods:* The evaluation was conducted during a usability and UX workshop, where participants used Meta Quest 2 and Meta Quest 3 headsets. Participants were divided into two groups: Group A completed the UUXE-ToH v4 questionnaire, while Group B answered the USE, Slater-Usuh-Steed (SUS), and UEQ questionnaires. In addition to quantitative assessments, participants reported issues, which were classified and analyzed thematically. *Results:* Cubism was positively evaluated in terms of learnability, memorability, and satisfaction. However, notable challenges were identified in object manipulation, gesture accuracy, and error recovery. Immersion and presence scores were comparatively lower, particularly in SUS, suggesting limited perceptual integration of virtual objects into the real environment. Interaction precision and environmental coherence were also affected by hardware limitations, especially on devices with monochrome passthrough. *Conclusion:* The findings highlight specific usability and UX limitations in MR puzzle games and illustrate how different instruments (applied to distinct groupings of participants) captured complementary aspects of the experience. Although no direct statistical comparison was performed, indirect analysis suggests varying diagnostic contributions across instruments. These findings suggest that multidimensional evaluation frameworks (either through a single comprehensive instrument or a combination of focused tools) can provide richer insights for guiding design improvements and identifying usability and UX issues in MR applications.

Keywords: Usability, User Experience, Puzzle Game, Mixed Reality, Questionnaire Comparison, Evaluation Methods

1 Introduction

Puzzle games have long been a popular form of entertainment, challenging players with problem-solving and strategic thinking. Advances in technology have transformed them into digital experiences, incorporating interactivity and immersion. Mixed Reality (MR) puzzle games further enhance engagement by leveraging spatial computing and hand tracking, seamlessly blending virtual and physical environments. Cubism [Bouwel, 2024] is an example of an MR puzzle game that allows users to manipulate geometric virtual blocks to build tridimensional shapes through direct hand interactions.

Evaluating usability and user experience (UX) in MR puzzle games is essential to understand player interactions and identify challenges that affect engagement and performance. Effective usability enables intuitive interaction, reducing cognitive load and frustration, while positive UX enhances motivation and immersion. In addition, puzzle games promote cognitive engagement and learning [Ritterfeld *et al.*, 2009]. User feedback is vital to improve software quality and identify missing features [Pagano and Bruegge, 2013]. Research in this field advances interaction techniques, player satisfaction, and best practices in immersive game design.

Assessing usability and UX in MR games requires analyzing key dimensions that shape player experience, includ-

ing usability factors such as effectiveness, efficiency, and learnability [Nielsen, 1993]. Playability is also crucial, covering functional aspects (e.g., precision, control responsiveness) and experiential factors (e.g., immersion, enjoyment) [Desurvire and Wiberg, 2009]. In MR environments, immersion and presence are fundamental to how users perceive and interact with virtual objects, directly impacting engagement and realism [Slater, 2009]. Comfort and cybersickness must also be addressed, as motion discomfort can hinder usability and negatively affect UX [LaViola, 2000]. Evaluating these dimensions improves game design, improving accessibility, inclusion, and player retention [Johnson *et al.*, 2016].

Several studies investigate methods for evaluating usability and UX in serious games and MR environments [Swan and Gabbard, 2005; Dünser *et al.*, 2008; Dünser and Billingham, 2011; Bai and Blackwell, 2012; Yáñez-Gómez *et al.*, 2017; Borges *et al.*, 2020; Veriscimo *et al.*, 2020; Campos *et al.*, 2023, 2025a; Frata Furlan Peres *et al.*, 2024]. Common evaluation methods include user testing, expert inspection, and user feedback [Ivory and Hearst, 2001; Roto *et al.*, 2009]. The choice of method depends on study objectives and available resources, with many studies combining approaches for a more comprehensive analysis [Pranoto *et al.*, 2017; Rhiu *et al.*, 2020; Zhang *et al.*, 2020].

This study uses questionnaires to collect user feedback,

capturing subjective perceptions of the experience of the player. Questionnaires are widely used in UX evaluations, employing Likert scales, semantic differentials, and open responses [Hartson and Pyla, 2012]. However, responses can be influenced by emotional factors, recency bias, and novelty effects, while some users may struggle with specific terminology [Sharples *et al.*, 2008; Merino *et al.*, 2020]. Despite these limitations, standardized questionnaires allow study comparisons, improving reliability and validity.

Various questionnaires assess usability and UX in interactive systems, covering general aspects and specific dimensions relevant to MR. The Usefulness, Satisfaction, and Ease of Use (USE) questionnaire [Lund, 2001] evaluates usability based on perceived usefulness, satisfaction, and ease of use. Presence and immersion are commonly measured with the Slater-Usch-Steed (SUS) Inventory [Slater *et al.*, 1994], while the broader aspects of UX are captured through the User Experience Questionnaire (UEQ) [Laugwitz *et al.*, 2008]. For hand interaction in MR, the Usability and User Experience in Touchable Holography (UUXE-ToH) questionnaire [Campos *et al.*, 2024a; Prado De Campos *et al.*, 2024; Campos *et al.*, 2025b] provides a comprehensive evaluation, including effectiveness, efficiency, learnability, comfort, immersion, and presence.

Considering this scenario, a study was performed to evaluate the usability and UX of the Cubism game using different questionnaires to identify interaction issues, areas for improvement, and to compare the effectiveness of various evaluation methods. Cubism was selected not only because it is one of the most popular and highest-rated puzzle games on the Meta Quest Store, compatible with the devices available in our research lab, but also due to its support for hand tracking and touch-based gestures without physical controllers. These interaction techniques remain relatively novel for many users and present unique usability and UX challenges in MR environments, such as gesture recognition accuracy, hand occlusion, and interaction feedback. As such, Cubism provides a representative case for assessing user experiences in mid-air gesture-based interaction scenarios.

This study builds upon previous work by the authors that evaluated Cubism using an earlier version of the UUXE-ToH questionnaire [Campos *et al.*, 2024b], and now extends the investigation by integrating multiple validated instruments to achieve a more complete and multidimensional understanding of usability and UX in MR puzzle games. The study was conducted during a workshop on usability and UX evaluation at the Brazilian Symposium on Human Factors in Computing Systems (IHC 2024). The participants played Cubism using Meta Quest 2 and Meta Quest 3 headsets. After the experience, the participants assessed the game using the UUXE-ToH, USE, SUS, and UEQ questionnaires.

The results identified areas for improvement in Cubism, such as the need to enhance object textures to improve perception of presence. Furthermore, the findings provided information on how different questionnaire structures impact usability and UX assessments, contributing to future research on Human-Computer Interaction (HCI) evaluation methods for MR games.

The remainder of this paper is structured as follows. Section 2 reviews related work on usability and UX evalua-

tion. Section 3 outlines the study planning, including design, procedures, instrumentation, and data analysis. Section 4 presents the results for each questionnaire, UUXE-ToH, USE, SUS, and UEQ. Section 5 discusses the findings, while Section 6 examines the limitations of the study. Finally, Section 7 provides conclusions and future research directions.

2 Related Work

Evaluating usability and UX in interactive systems involves various well-established instruments, each designed to assess different aspects of user interaction. Borges *et al.* [2020] present a comprehensive analysis of 58 instruments used to evaluate Player Experience (PX) in digital games, highlighting the diversity of components assessed, the predominance of scales and questionnaires, and the challenges related to terminological inconsistencies and cultural adaptation of instruments. Their study also introduces an interactive catalog to support researchers and practitioners in selecting appropriate tools according to the evaluation context. This theoretical contribution provides a relevant foundation for the present work, which applies and compares different validated instruments in an MR game scenario, helping to address the gap identified by Borges *et al.* [2020] regarding the empirical and comparative use of PX evaluation tools in specific interaction contexts.

In this study, four primary questionnaires are used: the Usability and User Experience in Touchable Holography (UUXE-ToH) questionnaire [Campos *et al.*, 2024a; Prado De Campos *et al.*, 2024], the Usefulness, Satisfaction and Ease of Use (USE) questionnaire [Lund, 2001], the Slater-Usch-Steed (SUS) presence questionnaire [Slater *et al.*, 1994], and the User Experience Questionnaire (UEQ) [Laugwitz *et al.*, 2008]. These instruments were selected because of their applicability to MR environments and their ability to measure relevant usability and UX dimensions.

The **UUXE-ToH** questionnaire was designed to assess usability and UX in MR applications where users interact with holograms using their hands. It covers multiple dimensions, including effectiveness, efficiency, learnability, comfort, memorability, immersion, presence, pleasure, interest, and absorption (flow). In its current version (v4), UUXE-ToH consists of 56 items using a 7-point Likert scale and semantic differential. The first part evaluates specific aspects of the solution, while the second assesses the overall experience through a semantic differential scale and four open questions for spontaneous feedback.

The **USE** questionnaire, originally developed to assess the usability of software, hardware, and services, consists of 30 items in four dimensions: Usefulness, Ease of Use, Ease of Learning, and Satisfaction. Although traditionally used in conventional usability testing, its adaptability allows assessment of practical usability aspects in MR applications. Fast-Berglund *et al.* (2018) suggested that incorporating additional dimensions, such as aesthetics and enjoyment, could enhance its applicability to entertainment-based experiences.

The **SUS** presence questionnaire is widely used to measure the user's sense of presence in virtual environments using 6 items of Likert scale. The presence is a fundamental

factor in MR applications that influences user engagement and quality of interaction. The questionnaire evaluates how strongly users feel immersed in the virtual environment, capturing their perception of seamless integration between virtual and physical elements. Studies highlight its significance in MR gaming and training applications, where presence is essential to deliver realistic and engaging experiences.

The UEQ assesses user experience using a semantic differential scale, presenting 26 items as pairs of opposite adjectives (e.g., “complicated” vs. “easy”). It covers six dimensions: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. This questionnaire provides a balanced assessment of both pragmatic and hedonic UX aspects, making it particularly useful in MR applications where engagement, enjoyment, and perceived usability play key roles.

A previous study [Campos *et al.*, 2024b] evaluated the same Cubism game using an earlier version of the UUXE-ToH questionnaire (v2) with seven participants. That work focused exclusively on a single instrument and emphasized playability-related aspects in a more controlled setting. While it identified relevant usability and UX issues, its scope was limited by sample size and questionnaire version. The present study goes beyond that work by incorporating the updated UUXE-ToH v4 [Campos *et al.*, 2025b] and three additional instruments (USE, SUS, and UEQ), enabling a broader, multi-dimensional assessment. It also explores how different evaluation tools converge or diverge in capturing key constructs such as immersion, presence, and satisfaction, in a more socially varied context, a public workshop with multiple headset models.

Lee *et al.* [2023] evaluated the usability and UX of a virtual reality (VR) puzzle game using the System Usability Scale (SUS) and the UEQ. The study examined UX aspects such as usability and hedonic qualities, providing information on the overall perception of the user. The findings identified areas for improvement, particularly in the intuitiveness of the user interface.

Although Lee *et al.* [2023]’s study provides valuable insights, it differs from the present research in key aspects. First, our study employs a broader set of evaluation instruments, incorporating the UUXE-ToH, USE, SUS and UEQ for a more comprehensive assessment. Second, while Lee *et al.* focused on usability and general UX aspects, our research also evaluates immersion and presence, critical factors in MR applications but not addressed in their study. By integrating these additional dimensions, our research provides a more in-depth understanding of user interaction and engagement in the puzzle game.

3 User Study

This section describes the study design, participant characteristics, the evaluation instruments used in the study, and data analysis.

3.1 Study Design and Procedures

The study happened during a workshop on usability and UX evaluation in augmented and virtual reality (AR/VR) applications at the Brazilian Symposium on Human Factors in Computing Systems (IHC 2024). The study followed ethical guidelines and was approved by the Research Ethics Committee of the Federal University of Paraná (UFPR), registered in Plataforma Brasil by Certificate of Presentation for Ethical Consideration (CAAE) number 77369524.6.0000.0102 (approval n° 7.111.664, issued on September 30, 2024).

The study involved 14 participants, who were invited to try to evaluate the Cubism puzzle game (Figure 1) using Meta Quest 2 and Meta Quest 3 headsets. The study was conducted as follows. Participants were informed about the study objectives, risks, and benefits and were invited to participate voluntarily. Those who agreed signed an Informed Consent Form (ICF) and completed a characterization form.

Each participant played Cubism for 10 minutes, interacting with the game using hand tracking in an MR environment. This duration was considered sufficient based on information provided by the game’s developer, indicating that the first five puzzles take an average of 1.3 minutes each to complete. Thus, within the allocated time, most participants could complete at least the first three puzzles, allowing them to experience the game’s mechanics and interactions. Since the puzzles progressively increase in complexity, solving two to three challenges was enough for participants to engage with the core mechanics and perform all the primary movements required in the game.

After playing, the participants evaluated the game using the proposed questionnaires. The participants were randomly divided into two groups: A and B. This division aimed to prevent participant fatigue by avoiding the burden of answering all available questionnaires in the study. Additionally, the distribution ensured that each participant responded to a reasonably balanced number of items, resulting in a similar overall completion time between groups. Furthermore, the questionnaires assigned to Group B were selected to cover complementary dimensions, as there was minimal overlap between them.

Group A: They used the UUXE-ToH questionnaire in its current version (v4)¹. Because the study took place as an activity during a workshop, to shorten the time required for completion, only Part 1 (with 56 items) of the questionnaire was used, omitting the section containing open-ended questions.

Group B: They used a combination of three questionnaires: USE, Slater-Usoh-Steed (SUS), and UEQ². This combination was chosen due to their established validity and extensive use in usability and UX research. The USE questionnaire evaluates usability through structured constructs, the SUS questionnaire measures presence and immersion, and the UEQ captures a wider range of UX attributes. Together, these tools encompass a total of 62 items in multiple dimensions, allowing for a comparative analysis with the UUXE-ToH v4.

The participants completed their respective evaluation

¹UUXE-ToH v4: <https://doi.org/10.6084/m9.figshare.28143752>

²USE+SUS+UEQ: <https://doi.org/10.6084/m9.figshare.28437455>



Figure 1. Scenes taken from a promotional trailer demonstrating the game Cubism [Bouwel, 2024].

questionnaires on paper and, after, they conducted an inspection-based evaluation of the game, using the questionnaire items as a guide to identify and report issues. Because each participant completed only one set of questionnaires, no one responded to all questionnaires. Thus, the comparisons between the instruments throughout the study are based on indirect analysis of aggregated responses and inspection outcomes, not on statistical tests within the subject.

3.2 Data Processing and Analysis

The collected data were organized into two distinct datasets: (i) Dataset A, containing responses to the UUXE-ToH questionnaire; and (ii) Dataset B, containing responses to the USE, SUS, and UEQ questionnaires. Additionally, participants reported gameplay issues in a spreadsheet, which was consolidated into a single list, with problems classified as unique (reported by only one participant) or duplicated (reported by multiple participants). These issues were later grouped by topic to support qualitative analysis and identify areas for improvement.

Given the independent nature of Groups A and B, comprising different participants without pairing, the data do not allow for direct correlation analysis between instruments. Correlations require measurements from the same observational units, which is not the case here. Moreover, the small sample size ($n = 7$ per group) limits the statistical power of inferential tests. To address these constraints, a multifaceted analytical strategy was adopted, emphasizing: (1) detailed descriptive statistics (e.g., medians, modes, interquartile ranges) for each item and dimension; (2) graphical visualizations, including comparative bar charts and heatmaps of responses; and (3) indirect comparisons of grouped constructs (e.g., *Satisfaction*, *Usability*), based on central tendency measures. This approach enables a rigorous exploration of patterns without overextending the data's inferential capacity.

All data processing was conducted in the R environment. The datasets were checked for structure and integrity, ensuring all variables were correctly represented as ordinal values. Descriptive statistics (mean, median, standard deviation, and interquartile range) were calculated for each item. For the UEQ, items with inverted polarity were appropriately adjusted, and the official spreadsheet provided by the instrument developers was used to support construct-level aggregation and comparison.

Initially, the datasets were loaded into the R environment.

The structure and integrity of the data were verified to ensure that all variables were correctly represented as ordinal values. The descriptive statistics for each variable (item) were then calculated, including the mean, median, standard deviation, and interquartile range, providing an overall understanding of the response distributions. The inverted pairs in the UEQ questionnaire also had their scores inverted for calculation purposes.

Finally, heatmaps of participant responses were generated for each evaluation instrument to support visual inspection of patterns. These were accompanied by summary tables of median scores per construct. Together, the statistical and visual analyses provided complementary understandings into participant evaluations and system performance across instruments.

3.3 Participants

Fourteen people participated in the study, five men and nine women; three were between 18 and 20 years old, nine between 21 and 30, and two between 41 and 50 years old. Regarding education, one doctorate, one doctoral student, one master's student, and 11 undergraduates participated. The participants indicated that they could give opinions on Usability (13 of them), UX (12), Virtual Reality (4), and AR/MR (4). Of this group, the majority had experience in planning or conducting Usability and UX evaluation processes in the classroom (11 people) and/or in the industry and market (4 people). Of these, 11 participants have already planned or conducted up to 4 evaluations, two between 5 and 15 and one, more than 15 evaluation processes.

Regarding knowledge of AR/VR/MR concepts, one participant identified himself as a novice, with the workshop being their first exposure to the topic; seven considered themselves at a basic level, having partial knowledge of the terms; two at an intermediate level, understanding the terms and their applications; and four at an advanced level, stating that these technologies are part of their daily activities. Regarding the use of AR/VR/MR applications with headsets, three participants had never used them, nine used them rarely, one monthly, and one weekly.

4 Results

Quantitative results are presented in the heat map of individual responses and in the median distribution of the evaluated constructs.

4.1 UUXE-ToH Results

The evaluation of the Cubism game using the UUXE-ToH v4 questionnaire indicated a predominantly positive experience, with high medians in most constructs (Figure 2). Learnability, Memorability, Usefulness, Interest, Absorption, Satisfaction, Pleasure & Fun, and Emotions reached the max median of 7, reflecting that participants considered the game easy to learn, enjoyable, interesting, engaging, and pleasurable. In addition, others dimensions reached high median (6): Effectiveness, Comfort, Controllability and Operability, Immersion, Value, Creativity & Novelty, and Beautiful & Aesthetic.

However, some aspects showed a lower rating: Error Prevention and Recovery (median 4.5), Trustworthiness (5), Presence (5) and Efficiency (5.5). It suggests challenges in the perception of integration of virtual elements into the real environment and in the accuracy of interaction through hands.

The heatmap (Figure 3) allows one to see the items that received the lower ratings and the dimensions with more different opinions between participants. For example, the item S45 (“The holograms appeared to be a natural part of the real world, as if they were truly present around me”) had a mode of 2 and a median of 3. It indicates that the participants did not feel that the objects were part of the physical environment, maybe due to the monochromatic passthrough of Meta Quest 2 and the lack of realistic textures on the blocks.

In Efficiency, item S5 (“This holographic solution increases my productivity in performing this type of activity”) received some of the lowest ratings, which is expected since Cubism is a game, not a productivity tool. Similarly, in Error Prevention and Recovery, item S18 (In the cases where I made errors while using it, the holographic solution helped me resolve them”) received lower scores, suggesting the game lacks explicit error recovery or user assistance mechanisms.

In Immersion (median 6), the lowest scores were recorded in items S22, S23 and S25. In item S22, which assesses whether the holograms appeared to obey the laws of physics, participants indicated that the objects did not react as if they were real, which may have impacted the feeling of realism. Item S23, which deals with the accurate perception of objects amidst multiple elements in the field of view, also received lower ratings, suggesting that participants did not feel that the system facilitated interaction by highlighting the objects of interest. In addition, item S25, related to the accuracy of interaction and the occurrence of “passing through” the holograms, had reduced scores, indicating that the detection of gestures may not have been good enough to avoid this effect, which may compromise the feeling of control and realism in the experience.

4.2 USE, SUS and UEQ Results

Data analysis using the USE questionnaire revealed an overall positive perception, with high medians for Satisfaction (7), Ease of Learning (7), Ease of Use (6), and Usefulness (6) (Figure 4). Similarly, the evaluation through the UEQ indicated a largely positive reception, with high medians in most constructs: Attractiveness and Perspicuity (7), Dependability (6.5), Efficiency and Stimulation (6), and Novelty (5). However, the SUS questionnaire, which represents the dimension of Immersion and Presence, obtained the lowest median (3).

A notable aspect of the data was the presence of the option “not applicable” in several items of the Usefulness dimension, especially in item Use7 (“It meets my needs”), marked by three participants (Figure 5). This led to missing data for these observations, indicating that some users did not see the game as meeting a specific need. This pattern suggests that the perception of usefulness varied, with some considering it meaningful and others viewing it as purely recreational.

The Use6 item (“It saves me time when I use it”) in Usefulness (USE) showed the greatest imbalance, with a mode of 1 and a median of 3.5. In Ease of Use (USE), items EoU6 (“It saves me time when I use it”) and EoU8 (“I don’t notice any inconsistencies as I use it”) also had divergent scores, with modes of 1 and 3 and medians of 4 and 3, respectively. In Satisfaction (USE), Sat6 (“I feel I need to have it”) showed the greatest divergence, with both mode and median at 4.

For Immersion and Presence (SUS), the common mode was 1 for most items, except SUS3 (“The game pieces seemed more like images I saw or real objects in the environment”) and SUS4, which had modes of 3 and 4, respectively. The common median ranged from 2 to 3, except for SUS1, with a median of 5, and SUS5 (“I remember the game pieces as if they were real objects, similar to others I interacted with today”), with a median of 4.

The Attractiveness construct (UEQ), which evaluates a pleasant and satisfying experience, received highly positive responses, with a median of 7 for all items except one (Unpleasant - Pleasant), which had a mode of 6 and a single rating of 3 (Figure 6). Similarly, Stimulation (UEQ), which measures engagement and motivation, had medians of 6 and 7 across all items, reinforcing that the game maintained participants’ interest. Dependability (UEQ), assessing predictability and reliability, showed medians between 5 and 7, indicating that most participants felt the game responded stably to interactions. However, the Unpredictable - Predictable pair was rated 1 by one participant and 4 by two others.

Regarding Perspicuity (UEQ), which evaluates ease of learning, had positive results with medians between 6 and 7, although one participant rated the Complicated - Easy pair as 2 and two participants rated the Difficult / Easy to Learn pair as 3. Efficiency (UEQ) had a median of 6 and 7 for three items, while one item (Slow-Fast pair) had a median and mode of 4, indicating that some users found interactions slower than expected, possibly due to gesture recognition limitations. Novelty (UEQ), which assesses innovation and creativity, had the most varied responses. Two items achieved the highest medians (6 and 7), while two others had medians of 5 and 4. This suggests that while the game pro-

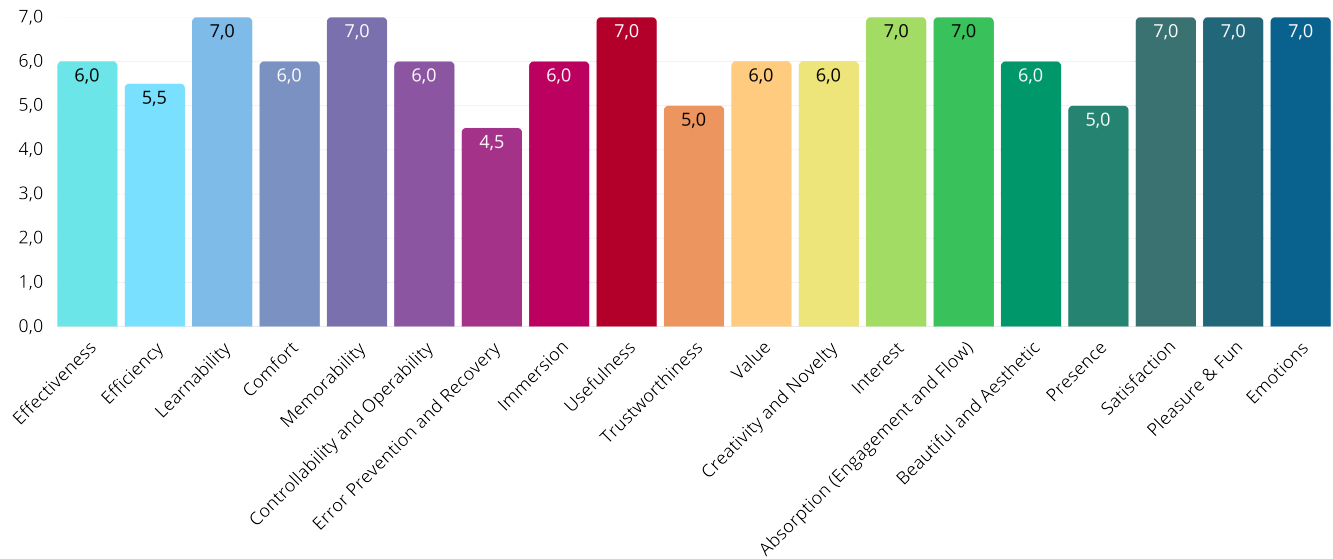


Figure 2. Medians for UUXE-ToH v4 dimensions

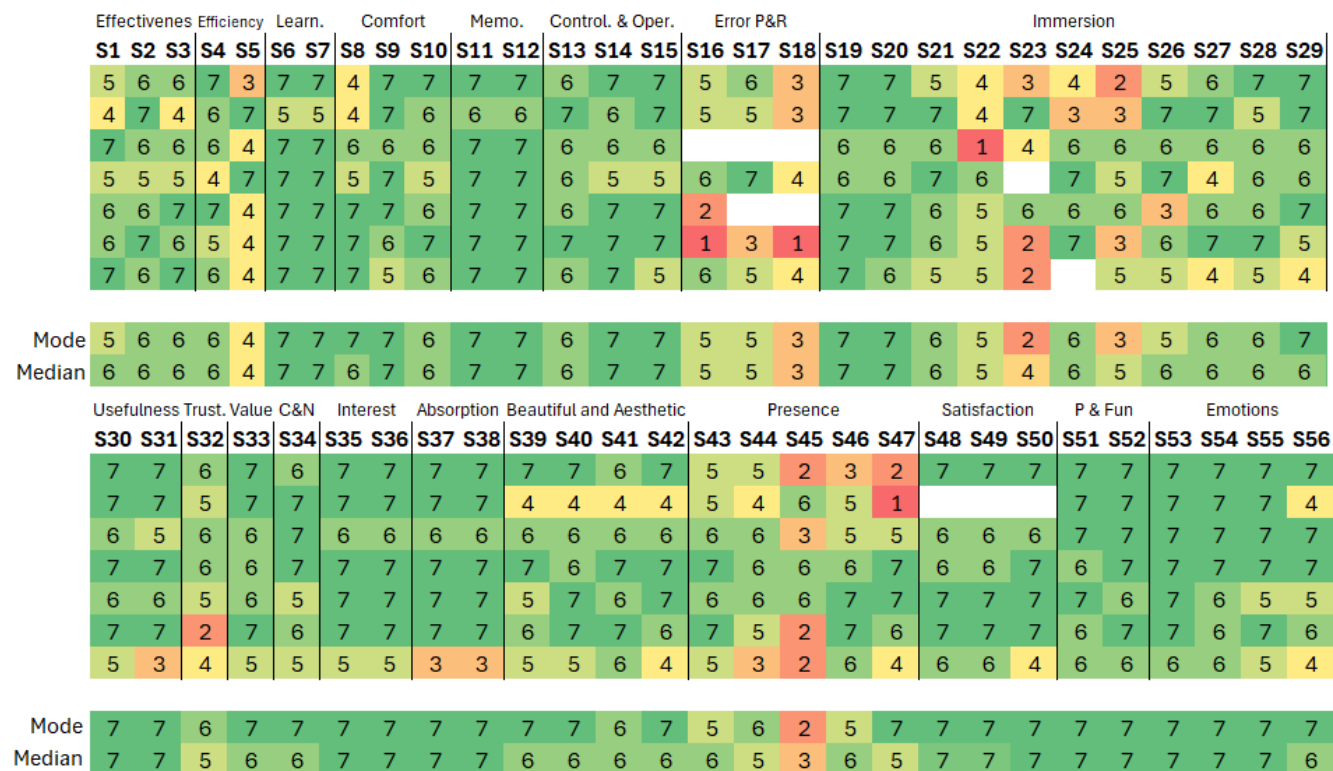


Figure 3. Heatmap for UUXE-ToH responses

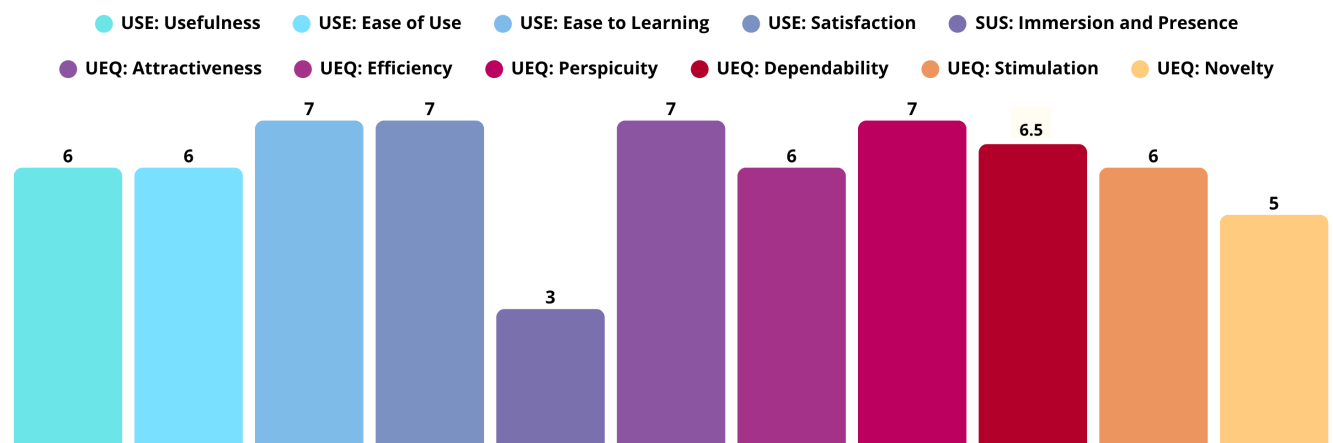


Figure 4. Bar chart for medians of USE, SUS and UEQ dimensions

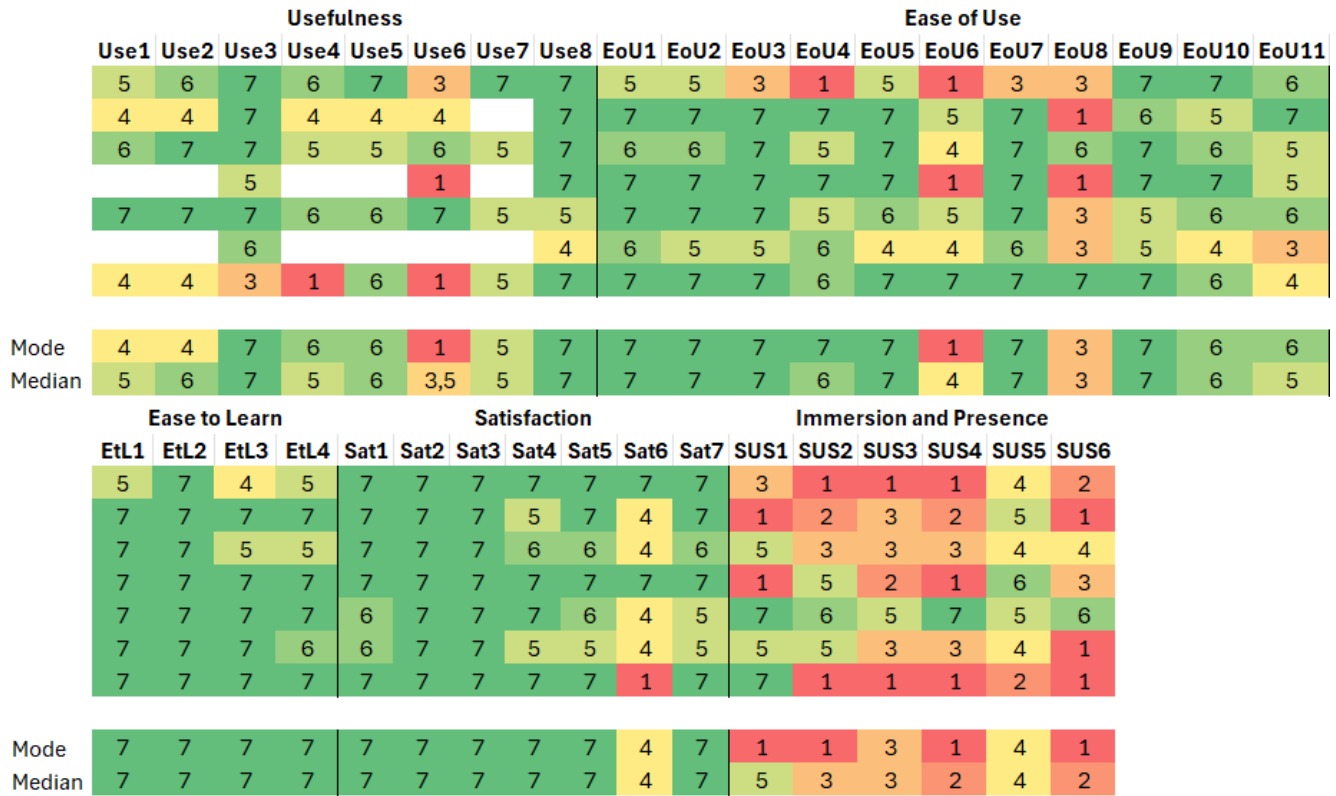


Figure 5. Heatmap for responses on USE and SUS questionnaires

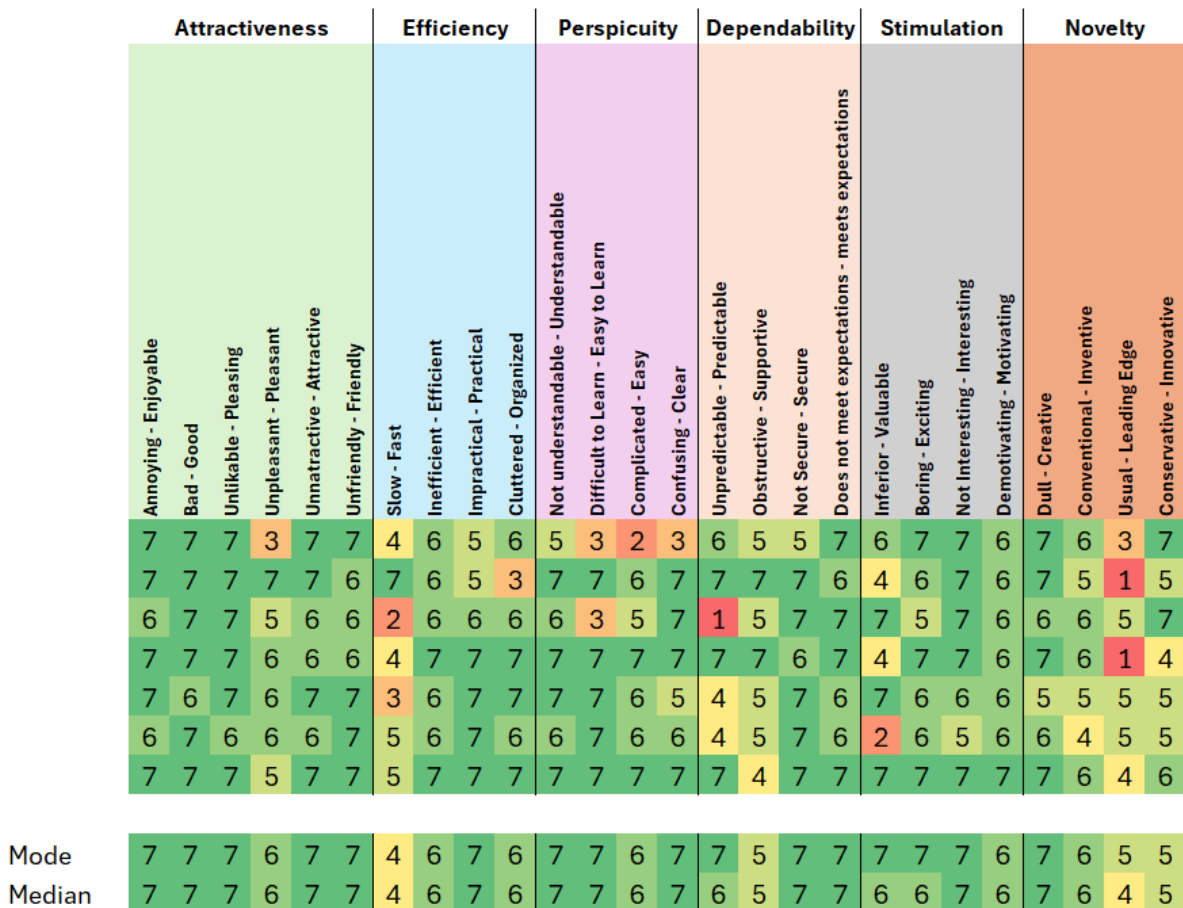


Figure 6. Heatmap for UEQ responses

vided a well-rated experience, it was not perceived as particularly innovative compared to other augmented reality or puzzle games.

The analysis performed using the UEQ Data Analysis Tool reinforces the findings of the previous evaluation, but presents a more structured quantitative view with means and variances of the constructs (Figure 7 a). Overall, the results indicate that the experience with Cubism was predominantly positive, with Attractiveness presenting the highest mean value (2.50) and Novelty the lowest (1.21) - in a scale from -3 to +3, which suggests that the game was considered enjoyable, but not necessarily innovative. The greatest variability was presented in the Perspicuity construct.

The results for pragmatic and hedonic quality provide a broader view of the UX (Figure 7 b). Pragmatic quality (1.87), which includes Perspicuity, Efficiency, and Dependability, was rated more positively than hedonic quality (1.64), which includes Stimulation and Novelty. This suggests that participants perceived the game as well-structured and functional, but lacking in innovative or significantly differentiating elements in terms of experience. Attractiveness (2.50), which represents the user's judgment regarding the product as a whole, was positive, presenting the highest average and lowest variability among the constructs.

4.3 Issues Lists

A total of 43 issues were identified in the game (Table 1). When associated with the dimensions assessed by the questionnaires, such as the UUXE-ToH, most issues were related to controllability and operability (15), immersion (10), error prevention and recovery (6), and effectiveness (5).

Some of these issues stem from the headset's capabilities rather than the game's functionality. For instance, issue 37 highlights a complete separation between virtual objects and the real world when using Meta Quest 2, due to its low-resolution, black-and-white passthrough view, which emphasizes virtual content rather than blending it with reality. Similarly, issue 30 reports distortions in the real-world image near the hand, a known challenge in video-see-through devices, particularly in hand-tracking applications. These limitations are inherent to the hardware and operating system and beyond the game developer's control. However, recognizing them helps define the boundaries of the game's experience and expectations for mixed reality improvements.

The remaining issues are related to the game itself and can be addressed by developers. For example, issue 3 highlights the difficulty in noticing the menu button, which uses a standard hamburger icon but has a translucent background and is smaller than other interface elements, making it less visible. A possible solution is to expand the button and improve its contrast. Issue 23, where users struggled to find the correct block placement, could be mitigated by adding subtle cues, such as a glow when a piece is correctly positioned.

Beyond reporting the issues, participants were instructed to indicate which questionnaire items led them to identify each problem during the inspection. This process revealed important differences in the way each instrument supported issue detection.

In Group A, which used the 56-item UUXE-ToH (covering 19 dimensions), participants identified 31 issues, 19 of which were explicitly linked to 17 different questionnaire items. Although 12 issues lacked specific item references, the overall breadth of detection suggests that UUXE-ToH supported a wide range of observations.

In Group B, which used the USE, SUS and UEQ questionnaires, the USE questionnaire was the most effective in supporting inspection-based issue identification (13 distinct issues were related to 19 of its 30 items). SUS was referenced for three issues, two of which also overlapped with USE, and UEQ was referenced for three issues too, two of which also overlapped with USE.

Interestingly, no issues were related to dimensions such as Comfort, Memorability, Usefulness, Value, Creativity and Novelty, Satisfaction, or Emotions dimensions of UUXE-ToH. This finding aligns with the issue detection via UEQ, which also targets hedonic and emotional aspects, but had no issues related to items of Attractiveness, Dependability, Stimulation or Novelty dimensions. Possibly, the short time for trying the game and inspecting for issues limited reflections about affective and long-term issues.

This information provides initial evidence on how the structure and dimensional focus of each questionnaire can affect its contribution to usability and UX diagnosis. In the next section, a deeper discussion of these differences and their implications follows.

5 Discussion

This section is divided into three parts. Section 5.1 compares the results of the questionnaires. Section 5.2 discusses how the reported issues impact usability and UX dimensions. Section 5.3 classifies these issues by implementation feasibility and scope to support prioritization.

5.1 Comparing Results

Comparing usability and UX questionnaire results is challenging, even when the constructs share the same name, as the definitions, scope, and evaluated facets often differ. Some questionnaires provide detailed constructs, while others take a more general approach, making direct comparisons difficult. Despite this, analyzing construct medians can offer insights into participants' perceptions of the game. However, differences in definitions, scope, and measurement methods require careful interpretation of similarities and discrepancies.

It is important to note that the comparison between instruments in this section is based on responses from different groups of participants. Since each group completed a different set of questionnaires, results are not directly comparable at the statistical level. The analysis presented here aims to explore qualitative patterns and differences in instrument sensitivity and diagnostic coverage.

Effectiveness, Efficiency and Ease of Use: In UUXE-ToH v4, the Effectiveness construct had a median of 6, indicating that users achieved their objectives satisfactorily. This aligns with the median of 6 for Usefulness in the USE

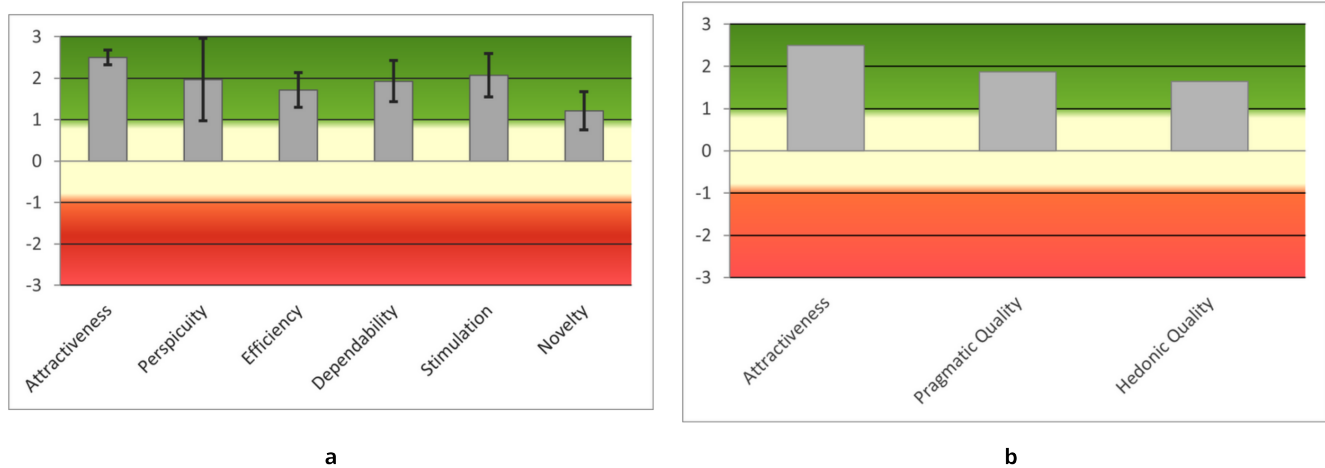


Figure 7. a) Constructs means; b) Means of pragmatic and hedonics dimensions

Id	Description	Topic
1	Difficulty in selecting due to distance	Learnability
2	Lack of initial instructions on game objectives	Learnability
3	The access button was difficult to perceive	Beauty and Aesthetics
4	The solution seems unaware of misaligned pieces	Trustworthiness
5	The game accepted misaligned pieces	Trustworthiness
6	Objects flew too close to the user	Comfort
7	Found the pieces too small	Controllability
8	Sometimes the game moved everything instead of a single piece	Controllability
9	Delay in game response when trying to pick up objects	Controllability
10	Difficult to use the rotation axis of pieces	Controllability
11	Difficulty in understanding depth	Controllability
12	Difficulty in understanding holographic hands	Controllability
13	Difficulty in changing perspective view	Controllability
14	Difficulty in picking up pieces behind the mold	Controllability
15	Difficulty in executing the pick-and-release motion	Controllability
16	Difficulty in selecting the correct object when overlapping	Controllability
17	Difficulty in interacting with pieces in the corner	Controllability
18	Unpredictability of the return location of a thrown object	Controllability
19	Game did not allow rotation on all axes	Controllability
20	Pieces go too far away	Controllability
21	Perception of having to use only one hand	Controllability
22	Difficulty in perceiving the piece's dimension	Effectiveness
23	Difficulty in finding the correct fit	Effectiveness
24	Failures when interacting with both hands	Effectiveness
25	Failures in some attempts to rotate the object	Effectiveness
26	Failures in some selection attempts	Effectiveness
27	The real environment caused distractions	Immersion
28	Lack of noticeable feedback when selecting pieces correctly	Immersion
29	Lack of realism	Immersion
30	Distorted real-world image near the left hand	Immersion
31	Game did not focus only on pieces within the field of view	Immersion
32	Objects respond only to specific gestures	Immersion
33	Pieces did not seem to have weight	Immersion
34	Perceived overlap and breaking of physical outlines	Immersion
35	Touches did not correspond to object dimensions	Immersion
36	Passing through objects	Immersion
37	Total separation between objects and the real world (Quest 2)	Presence
38	The holographic solution led me to movement errors	Error Handling
39	The solution does not help to correct errors	Error Handling
40	Accidental triggering of commands	Error Handling
41	Confusion when picking up objects	Error Handling
42	Lack of instructions to resume the game after accessing the system menu	Error Handling
43	Game did not indicate when the user made a mistake	Error Handling

Table 1. List of issues identified on Cubism

questionnaire, which also assesses perceived effectiveness. However, the Efficiency construct in UUXE-ToH v4 had a slightly lower median of 5.5, suggesting that while tasks were completed, users perceived them as less efficient. This perception is reinforced by item S5 (“This holographic solution increases my productivity”), which had a median of 4, indicating that there was no significant perceived productivity gain. In UEQ, the Efficiency construct also had a median of 6, similar to UUXE-ToH v4. Meanwhile, in USE, Ease of Use had a median of 6, suggesting the solution was generally considered easy to use. However, item EoU6 (“Using it is effortless”) had a lower median of 4, revealing that some users still found it somewhat demanding.

Learnability and Memorability: The constructs Learnability and Ease of Learning had a median of 7 in both UUXE-ToH v4 and USE, indicating that users found the solution intuitive and easy to learn. This is reinforced by the high Perspicuity score (UEQ, median 7), which measures clarity and ease of learning. Furthermore, Memorability in UUXE-ToH v4 also had a median of 7, suggesting that users felt that they would remember how to use the solution even after a long period without interaction.

Immersion and Presence: Among the evaluated dimensions, the most significant discrepancy across instruments was observed in constructs related to Immersion and Presence. While UUXE-ToH v4 recorded a median of 6 for Immersion and 5 for Presence, the SUS questionnaire, which specifically evaluates these aspects, had an overall median of only 3.

In UUXE-ToH, for Immersion, the worst rates values were in sentences S23 (“The holographic solution accurately perceived and highlighted the specific object that I was looking at, facilitating my interaction even when several objects were in my field of vision”, median 4 and mode 2), and S25 (“My interaction with the holograms was accurate, without my hand or other body parts passing through the contact surface of the objects”, median 5 and mode 3). These values were decisive for the Immersion median not to exceed 6, even with another 8 of 11 items presenting medians of 6 or 7. For Presence in UUXE-ToH, the sentence S45 (“The holograms appeared to be a natural part of the real world, as if they were truly present around me”, median 3 and mode 2) presented the worst evaluation, and the other sentences of the aspect had medians between 5 and 6, bringing the final evaluation of this aspect to 5.

In the SUS, 4 of the 6 items had a median of 2 or 3, and the other two items had a median of 4 and 5, contributing to an overall median of 3. The most positive evaluation in the SUS was item SUS1 (“I had the feeling that the game pieces were part of the environment around me”, median 5). Despite a median 5 in this item, the ratings showed extreme opinions among users, with higher frequencies for ratings 1 and 7. On the other hand, SUS2 (“There were moments during the experience when the game pieces seemed completely real to me”) and SUS4 (“During the experience I had a stronger feeling that the pieces were a real part of the environment”) items presented a median of 2 and a mode of 1. These items deal with the perception that the virtual elements are so integrated into the real world, as if they really existed there.

Note that these SUS items evaluations resemble the evalua-

tions of item S45 of UUXE-ToH. The presence dimension in SUS is greatly influenced by the perception of the element being in reality, covering at least half of its evaluation items. UUXE-ToH addresses other factors in the presence dimension, such as the similarity of the interaction with the virtual objects in relation to their real counterparts, the coherent behavior of the virtual objects in response to the user’s actions. It also addresses the perception that the sound emitted by the solution is also integrated with the environment. Therefore, it is not limited to visual elements. This difference in the assessment of the Presence dimension justifies the difference in general scores for Presence between these two instruments.

These values differences can be attributed to the distinct theoretical foundations and methodological scopes of the two instruments. While both aim to assess the user’s sense of “being there,” UUXE-ToH conceptualizes Immersion and Presence as separate constructs: Immersion pertains to technical and interactional aspects of the system (such as tracking stability, latency, and feedback) while Presence refers to the user’s perceptual and psychological sense of realism and spatial integration. In contrast, SUS focuses exclusively on Presence, adopting a holistic and memory-based perspective grounded in VR research, where the evaluation is centered on the intensity and authenticity of the experience as remembered by the user.

Moreover, UUXE-ToH follows a diagnostic and system-oriented structure, providing subscales that capture specific technical characteristics and their contribution to the user experience. This design allows users to recognize and reward aspects like interaction accuracy and system responsiveness, even in the absence of a fully convincing illusion of reality. SUS, on the other hand, consolidates Presence into a single subjective judgment, which may be more sensitive to factors such as limited session time, minimal physical movement, or hardware constraints like passthrough quality and hand-tracking inconsistencies.

In this context, the higher Immersion scores observed in UUXE-ToH may reflect participants’ positive evaluations of system performance and interaction design, whereas the lower SUS Presence scores capture a less compelling illusion of realism and spatial blending. These findings highlight the importance of selecting evaluation tools aligned with the specific goals and constraints of MR experiences, especially when distinguishing between technical immersion and psychological presence.

Satisfaction and Emotional Experience: Satisfaction showed high results in different questionnaires, with a median of 7 in both UUXE-ToH v4 and USE. Similarly, the Attractiveness construct in UEQ, which measures a related perception, also had a median of 7. However, item Sat6 in USE (“I feel I need to have it”) had a lower median of 4, indicating that despite high satisfaction, users did not perceive the solution as essential. The Pleasure and Fun construct in UUXE-ToH v4 also had a median of 7, suggesting that participants found the experience enjoyable.

Stimulation and Engagement: UUXE-ToH and UEQ assess factors related to user motivation and engagement with the product. In this context, the Stimulation construct in the UEQ, which measures how engaging and motivating the product is, had a median of 6. In UUXE-ToH v4, the Interest

and Absorption (Flow and Engagement) constructs also had medians of 7, indicating a positive alignment in evaluating these aspects.

Trustworthiness and Control: The perception of the trustworthiness of the solution varied between questionnaires. In UUXE-ToH v4, Trustworthiness had a median of 5, while Dependability in UEQ was slightly higher at 6.5. This difference may stem from item S32 in UUXE-ToH v4 (“I trust that the solution correctly performs the actions I determine”), which had a median of 5 but a minimum of 2, indicating variations in user confidence. In USE, Ease of Use included trust-related items, such as EoU8 (“I don’t notice any inconsistencies as I use it”), which had a low median of 3. This suggests that perceived inconsistencies in the interface or system behavior may have impacted user trust.

In Conclusion: Differences in construct formulation and item evaluation reinforce the challenges of comparing median values across questionnaires. However, some general trends emerge:

- 1) The game was perceived as effective and useful, with high medians for Effectiveness and Usefulness, though perceived efficiency and productivity were slightly lower.
- 2) Learning and memorability received positive ratings, suggesting the solution is intuitive and easy to remember.
- 3) Perceptions of immersion and presence varied significantly, with UUXE-ToH v4 assigning higher medians than SUS, which had notably lower scores.
- 4) Users reported high satisfaction and enjoyment, but trust in the system and absence of inconsistencies were not consistently well-rated.

These findings highlight the impact of questionnaire choice on evaluation results and emphasize the need to consider construct definitions and scope when interpreting data. These also underscore the limitations of applying instruments originally developed for VR, such as the SUS, in MR settings. The unidimensional nature of SUS may fail to capture nuances relevant to spatial and perceptual coherence between real and virtual content. In contrast, UUXE-ToH’s separation of Immersion and Presence, grounded in MR-specific interaction factors, yielded more consistent and diagnostic results. This reinforces the importance of context-appropriate instrument design or adaptation.

5.2 Usability and UX Issues by Dimension: Impact and Implications

The identified issues in Cubism reflect distinct weaknesses across several usability and UX dimensions. Rather than reiterating each issue, this section focuses on how the reported problems affect user perception, task performance, and overall experience, highlighting the dimensions most impacted and the importance of addressing these aspects to improve interaction quality in MR environments.

Controllability and Operability was the dimension most affected, with issues reported by nearly all participants. Difficulties in selecting, rotating, and positioning blocks directly compromise the user’s sense of control, one of the pillars of effective interaction in gesture-based systems. When users fail to predict or execute simple actions, their confidence in the system declines, increasing frustration and task abandonment risk. Enhancing input recognition, reducing gesture am-

biguity, and ensuring predictable behavior are essential to reinforce users’ agency and autonomy in MR environments.

Effectiveness was impacted primarily by problems in visual and spatial perception. Participants struggled to accurately perceive the size and fit of the puzzle pieces, which hindered goal completion. These issues affect both the accuracy and completeness of tasks (core attributes of effectiveness), especially when feedback on correct or incorrect placements is insufficient or absent. Without appropriate guidance or affordances, users may not realize their mistakes or understand how to proceed, disrupting problem-solving flow.

Error Prevention and Recovery emerged as a critical weakness. Users frequently triggered unintended actions or failed to recover from interaction errors. Although trial-and-error is intrinsic to puzzle games, the absence of undo features or recovery cues intensified frustration. From a UX perspective, this compromises trust and perceived robustness. Systems that allow users to recover gracefully from mistakes not only enhance usability but also encourage exploration and learning through experimentation.

Immersion was undermined by multiple factors: low feedback fidelity, limited gesture responsiveness, and a perceived disconnect between virtual objects and the real environment. Unrealistic behaviors—such as passing through holograms or the lack of weight and collision—broke the illusion of presence and reduced sensory engagement. In MR applications, immersion depends not only on visual realism but also on interaction realism. When these fail, users remain aware of the artificiality of the experience, impairing emotional engagement.

Presence, as captured by both UUXE-ToH and SUS, received some of the lowest ratings. The feeling that virtual elements were truly part of the real world was weak, particularly when using devices with monochrome passthrough or when virtual objects lacked physical integration cues. This reflects a psychological disconnection that affects the sense of realism and spatial coherence—key aspects for MR experiences to be perceived as natural.

Trustworthiness and Comfort were also negatively affected. Issues like the acceptance of misaligned pieces or unrecognized misplacements reduced users’ confidence in the system’s internal logic. When interaction rules appear inconsistent, users question the system’s reliability. Furthermore, some participants felt physically uncomfortable due to object proximity, especially when content was initialized without recalibrating head position. Such discomfort can reduce session duration and increase motion fatigue.

Learnability issues, though less frequent, highlighted the absence of clear initial guidance. Lack of onboarding compromised users’ understanding of objectives and interaction possibilities, particularly for less experienced users. This gap undermines the discovery of functionalities and delays engagement, stressing the importance of contextual instructions and progressive learning aids.

Finally, issues related to **Beauty and Aesthetics** affected the perception of quality and professionalism. A poorly visible menu button, for instance, may seem trivial, but signals low design attention, negatively influencing first impressions and user satisfaction.

In summary, the reported issues not only impaired task per-

formance and interaction fluency, but also disrupted users' emotional connection to the game. While some challenges stem from hardware limitations, many affectable dimensions, such as controllability, effectiveness, feedback, and immersion, are critical to the overall UX and can be significantly improved through design refinements and enhanced interaction strategies.

5.3 Classification of Issues by Feasibility and Implementation Scope

While the previous section examined the identified issues in terms of UX and usability dimensions, it is also useful to categorize them based on the feasibility of implementation and the scope of responsibility for addressing them. This perspective helps differentiate which improvements are directly actionable by the game developers, which require substantial development effort, and which are constrained by hardware or system-level limitations.

Based on this classification, the issues can be grouped into three categories: (1) those that can be addressed with minor adjustments (e.g., UI elements, feedback mechanisms); (2) those that require significant changes to the game logic or interaction model; and (3) those that result from the limitations of the underlying hardware or operating system and are therefore outside the direct control of the developer.

5.3.1 Low-effort improvements (developer-resolvable): Issues 1, 2, 3, 6, 7, 13, 21, 23, 28, 42, and 43

These issues can be addressed through simple modifications in interface design, onboarding, or feedback mechanisms.

Issue 1 refers to difficulties selecting pieces at a distance, possibly caused by slight descalibration during headset handover. Implementing a brief position recalibration or ensuring proper headset positioning before starting could mitigate this. Issue 2 and 13 both reflect gaps in initial orientation (regarding game objectives and spatial interaction) that can be improved with interactive tutorials. Regarding Issue 13, specifically, since the participants were seated and not instructed to stand or move physically around the virtual content, their perspective was restricted to head and torso motion. This suggests a need for an initial orientation or tutorial that clarifies physical mobility options and encourages full-body exploration when appropriate.

Issue 3 involves low visibility of the menu button, solvable with enhanced contrast or dynamic highlighting. Issue 6 concerns discomfort from objects appearing too close; adjusting spawn distance can resolve this. Issue 7 relates to piece size, addressable by allowing scaling options. Issue 21 notes the perception of using only one hand, likely due to ambiguous onboarding, clarifying this possibility would help. Issues 23 and 28 reflect the need for visual or auditory cues when placing or selecting objects. Issue 42 involves lack of guidance after resuming the game from the system menu. Adding contextual messages would solve this. Finally, issue 43 addresses the absence of feedback when a mistake occurs; subtle cues or animations could inform the user without interrupting gameplay.

5.3.2 Moderate/high-effort improvements (developer-resolvable): Issues 4, 5, 8, 9, 10, 14, 15, 16, 17, 18, 19, 20, 22, 24, 25, 26, 29, 33, 36, 38, and 41

These issues involve deeper changes to the game's interaction model, manipulation logic, or feedback mechanisms. Their resolution requires structural design adjustments and extensive testing, particularly to ensure consistency with hand tracking and gesture-based control.

Issue 4 and 5 relate to trustworthiness: the game sometimes allows misaligned pieces to be placed or fails to recognize small misalignments. Addressing this requires improving the precision of the alignment detection algorithm and possibly adding real-time feedback (such as snapping, cues, or color changes) to guide correct placement.

Issue 8 involves unintentional manipulation of the entire puzzle mold when users attempt to grab a piece near its boundary. This occurs because the rotation anchors for the mold are activated by gestures that closely resemble the ones used for object selection. A solution would involve adding a toggle mechanism to lock or unlock the mold's movement, combined with clearer visual indicators to distinguish between manipulating the mold and individual pieces.

Issues 9 and 26 refer to interaction delays and selection failures. These could result from hardware input lag or insufficient gesture sensitivity. Solutions may include increasing gesture tolerance, filtering unstable input, and providing stronger visual feedback to confirm successful selection.

Issues 10, 19, and 25 involve difficulties rotating objects using natural wrist motions. A promising alternative is to implement a visual interface that appears when a user holds an object, offering axis-based rotation controls (e.g., X, Y, Z) that can be triggered with the opposite hand. This dual-mode control, that combines natural gestures and graphical rotation aids, would improve precision, reduce fatigue, and enhance accessibility.

Issues 14, 15, 16, and 17 all reflect problems with piece manipulation in constrained or visually ambiguous contexts. Difficulty in picking up pieces behind the mold (14), executing pick-and-release motions (15), selecting among overlapping objects (16), and interacting with pieces in corners (17) may be improved by implementing occlusion-aware highlighting, proximity-based selection prioritization, or cycling mechanisms to navigate overlapping elements.

Issue 18 reports unpredictable return points after throwing a piece away. Though the object eventually returns to reach range, its return position appears arbitrary. Refining this logic so objects return to a consistent, predictable location (such as a defined drop zone) would improve controllability.

Issue 20 reflects frustration with pieces that travel too far when released. Developers could implement soft movement boundaries or magnetic zones to retain objects within comfortable reach.

Issue 22 concerns difficulty in perceiving the dimensions of pieces. Enhanced depth cues (such as shading, occlusion, or environmental lighting) may help users better interpret three-dimensional volume.

Issue 24 reports problems when interacting with both hands simultaneously, likely due to limitations in the hand

tracking or gesture disambiguation system. Improved recognition models, or interaction modes explicitly designed for bimanual manipulation, could resolve these inconsistencies.

Issue 29 concerns lack of realism, possibly tied to minimalist visual design; while not technically demanding, it requires redesigning textures and physics to enhance perceived authenticity. Issues 33 and 36 relate to immersion-breaking factors: the lack of physical properties such as weight (33) and the ability of hands or pieces to pass through each other (36). These issues require improvements to the game's physics engine and collision models to simulate realistic constraints and contact feedback, even in the absence of haptic devices.

Issue 38 describes movement errors caused by the game interface itself, potentially due to unexpected behavior of objects or misleading visual cues. Addressing this may involve refining visual affordances and interaction precision to reduce unintentional actions.

Finally, issue 39 reveals a lack of support for error recovery; implementing undo mechanisms or confirmation prompts could reduce user frustration. Issue 41, which concerns confusion during object pickup, is an example of a problem that could be avoided or resolved by the user if undo mechanisms or similar features were available. This issue also highlights the need for clearer interaction feedback and selection clarity.

5.3.3 Hardware, system or human limitations (not developer-resolvable): Issues 30, 31, 32, 34, 35, 37, and 40

These issues arise from constraints imposed by the hardware (such as passthrough resolution, gesture recognition precision, or absence of haptic feedback) or by the system software, and are therefore outside the game developer's direct control.

Issue 11 concerns difficulty understanding depth, which may result from user-specific visual conditions (e.g., myopia), not software. Issue 12 involves confusion with holographic hand representations. These are system-rendered overlays not controlled by the application.

Issue 27 points to real-world distractions caused by the high-resolution passthrough of Quest 3, where the unblurred background competes for user attention.

Issue 30 highlights visual distortions near the user's left hand, which are common in video-see-through systems where the rendering pipeline and hand tracking struggle to align real and virtual content consistently.

Issue 31 concerns the absence of visual prioritization for objects based on the user's gaze direction. Ideally, MR systems could rely on eye tracking to dynamically adjust focus and relevance, but such features are unavailable in Meta Quest 2 and 3. As a result, content outside the central field of view can become visually distracting or cognitively demanding to manage.

Issue 32 notes that objects respond only to specific gestures, which limits interaction flexibility. This behavior is shaped not only by the game's design but also by the capabilities and constraints of the headset's gesture recognition

system. Expanding gesture vocabularies often requires support at the system level to ensure reliability and consistency.

Issue 34 adds to this by describing how hands or objects appear to break physical boundaries or intersect unrealistically, again pointing to limitations in spatial mapping and physics simulation without depth-aware collision handling or haptic reinforcement.

Issue 35 highlights a mismatch between hand positions and perceived object boundaries, where touches or interactions do not align with the physical dimensions of virtual items. This reflects limitations in spatial calibration and the absence of tactile feedback, which are common challenges in systems that rely solely on mid-air gestures and visual feedback.

Issue 37 refers to the lack of integration between virtual objects and the real environment when using Meta Quest 2. The device's low-resolution black-and-white passthrough limits spatial blending and reduces the perceived realism of mixed reality scenes.

Finally, issue 40 involves accidental triggering of system-level commands (such as invoking the Meta Quest menu) via unintended gestures. Since these gestures are interpreted by the headset's operating system and not the application, they cannot be disabled or filtered by the game itself. This highlights the importance of preparing users in advance by providing orientation or onboarding content that clarifies the functioning of system gestures and how to avoid interrupting gameplay unintentionally.

5.3.4 Concluding Remarks

The classification of issues by feasibility and implementation scope reveals how the effort required to improve specific UX dimensions in Cubism varies significantly depending on their nature. Dimensions such as *Learnability*, *Controllability*, *Error Handling*, and *Beauty and Aesthetics* include several issues that can be addressed with relatively low effort. These typically involve enhancements to interface clarity, initial onboarding, feedback mechanisms, and object placement logic, improvements that are feasible within the current design without requiring major architectural changes.

In contrast, dimensions like *Effectiveness*, *Trustworthiness*, *Comfort*, and *Immersion* often demand more substantial revisions to the interaction model, gesture handling, or physical simulation. For example, increasing the realism of virtual objects, improving the recognition of two-handed gestures, or refining rotation mechanics requires careful redesign and testing, especially to ensure consistency in gesture-based control across varied contexts.

Certain challenges, however, extend beyond what can be addressed at the software level. Issues related to *Presence* and some aspects of *Immersion* (such as visual blending with the real world or reliable gesture tracking) depend on hardware constraints like passthrough resolution, lack of eye tracking, and limitations of the hand-tracking system. Additionally, the difficulty in interpreting depth (Issue 11), while partially related to visual rendering, may also reflect user-specific perceptual limitations such as myopia or spatial cognition, aligning with broader concerns of accessibility. These aspects highlight the importance of designing inclusive MR

experiences that accommodate a wider range of human capabilities.

Overall, while many usability and UX issues in Cubism are within reach of practical and incremental improvements, others will only be meaningfully resolved through advances in MR hardware and greater support for adaptive, accessible interaction paradigms.

6 Limitations

This study provides valuable findings into the usability and UX evaluation of the Cubism puzzle game in an MR environment. However, several limitations must be acknowledged.

First, the small sample size (14 participants) limits the generalizability of the findings and restricts the scope of statistical analysis. This limitation did not result from methodological flaws but from practical constraints during data collection. The study was designed to be conducted within a four-hour workshop, with limited hardware availability and voluntary participation. Only three Meta Quest 2 and one Meta Quest 3 devices were available, and participation depended on event attendance and individual willingness to take part. These constraints also precluded selective recruitment or balancing of participant characteristics, leading to a sample mostly composed of students and researchers with some background in HCI.

In this context, the authors deliberately chose not to expand data collection to other environments or time periods. In previous studies conducted by the same team, larger samples were achieved through distributed sessions over several weeks, reaching up to 260 participants. However, this specific study was intended as a focused evaluation within a live workshop setting. It is also important to recognize the inherent difficulty in recruiting participants with sufficient expertise in usability and UX, especially during time-constrained technical events, due to their professional commitments. For future studies in similar settings, we recommend strategies such as pre-registration, time-slot allocation, or coordination with academic courses to ensure broader participation without compromising methodological integrity.

Second, the use of different headsets (Quest 2 and Quest 3) may have introduced variability in user experience. Hardware differences (such as passthrough resolution and hand-tracking precision) can affect perceived immersion and interaction quality. While the inclusion of both devices enabled broader participation, it also makes it more difficult to attribute usability issues to either the game design or device capabilities.

Third, participants engaged with the game for only 10 minutes. This short interaction window allowed exploration of core mechanics but limited the evaluation of long-term engagement, fatigue, and task progression. Although some UUXE-ToH items address comfort and tiredness, a more detailed assessment of these aspects would require extended or repeated sessions.

Four, while the choice of questionnaires is based on relevance and prior validation, inherently influences the results. Each instrument emphasizes different aspects of the experience, making direct comparisons difficult and potentially

limiting the completeness of the evaluation. Furthermore, only the first part of the UUXE-ToH v4 questionnaire was used, excluding open-ended questions that could have provided richer qualitative feedback.

Finally, the workshop setting, despite offering a controlled environment, may have influenced the behavior of the participants due to time constraints, social factors, or external distractions. In addition, participants were divided into two groups (A and B) to balance questionnaire completion time and prevent cognitive overload. Although this facilitated a manageable evaluation, it also meant that no single participant assessed the game across all selected instruments, limiting direct cross-questionnaire comparisons.

Future studies should consider conducting the evaluation outside event-based settings, recruiting a larger and more diverse sample, allowing longer interaction periods, and controlling for device variability. Including the full version of the UUXE-ToH questionnaire would also enable richer mixed-methods analysis.

7 Conclusion and Future Work

This study examined the usability and UX of the Cubism puzzle game in an MR environment using multiple evaluation instruments: UUXE-ToH v4, USE, SUS, and UEQ. The findings highlight strengths while identifying areas for improvement in interaction control, immersion, presence, and error prevention.

The results indicate that Cubism offers an engaging and intuitive experience, particularly in terms of learnability, memorability, and satisfaction. However, lower ratings in immersion, presence, and controllability reveal challenges in integrating virtual objects with the real world, especially with hand-tracking interactions. Usability issues such as object selection difficulties, gesture inconsistencies, and lack of error feedback were also noted.

Additionally, the classification of identified issues by implementation scope provided practical insights into which aspects can be improved with minimal effort, such as UI visibility or feedback cues. These aspects depend on more substantial design adjustments or external factors, such as hardware capabilities. This categorization supports prioritization for development. The findings also suggest that certain interaction challenges may stem from human perceptual limitations (e.g., depth perception), highlighting the importance of accessibility-aware design in MR.

The study also highlights the benefits of combining standardized usability and UX questionnaires with domain-specific instruments like UUXE-ToH. Comparative analysis of questionnaires underscores how variations in construct definitions and measurement approaches affect result interpretation, emphasizing the importance of selecting appropriate evaluation tools for MR applications. Although these comparisons were based on data collected from different groups and thus do not allow inferential statistical conclusions, the patterns observed offer valuable insights into how different instruments contribute to usability and UX evaluations in MR contexts.

In particular, the study sheds light on the diagnostic utility

of multidimensional evaluation technologies. During the inspection process, participants linked identified issues to specific questionnaire items, offering insight into which tools better supported usability and UX problem detection. Group A, using only the multidimensional UUXE-ToH (covering 19 dimensions), identified 31 issues. Group B, using a combination of USE (4 dimensions), SUS (one dimension), and UEQ (6 dimensions), identified 15 issues, with USE alone supporting most of the detection, while SUS and UEQ contributed marginally and redundantly. These patterns suggest that instruments designed to cover multiple dimensions (either through a comprehensive questionnaire or a well-balanced combination of questionnaires) may better support diagnostic evaluation in MR contexts. Nonetheless, further studies with larger and more controlled samples are needed to confirm these findings and explore the impact of instrument structure on evaluation outcomes.

The complementary nature of USE, SUS, and UEQ was evident, with each questionnaire emphasizing different aspects of usability and UX. USE assessed usability, ease of use, and satisfaction, while UEQ covered broader UX dimensions such as attractiveness, stimulation, and dependability. SUS evaluated immersion and presence. UUXE-ToH v4 emerged as a comprehensive alternative, covering effectiveness, efficiency, learnability, comfort, trustworthiness, and absorption. Compared to SUS, UUXE-ToH provided a more detailed breakdown of immersion and presence, incorporating both technical and perceptual elements.

Future research should expand the participant pool for a more diverse analysis of user experiences across skill levels. Longer play sessions could provide information on long-term engagement, fatigue, and usability at advanced game levels. Evaluating interface elements (e.g., menus, settings, tutorials) would enhance usability assessment, while exploring alternative interaction techniques (e.g., haptic feedback, adaptive gestures, voice commands) may improve accuracy and ease of use. Investigating hardware differences (e.g., Meta Quest, Apple Vision, MS HoloLens, Magic Leap) could clarify their impact on presence, immersion, and interaction precision. Finally, comparing evaluation methodologies, such as heuristic and expert-based assessments, could provide a more comprehensive understanding of usability and UX in MR environments.

Declarations

Funding

This research was partially funded by the Coordination for the Improvement of Higher Education Personnel (CAPES) — Program of Academic Excellence (PROEX), FAPEMIG (APQ-00890-23 and APQ-03665-22) and CNPq (306101/2021-1).

Authors' Contributions

TC and NV contributed to the conception of this study. TC also was responsible for data curation, formal analysis, investigation, resources, visualization, and writing the original draft of this manuscript. SD and NV also participated in the investigation and

were responsible for funding acquisition, supervision, and reviewing and editing this manuscript.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Availability of data and materials

The datasets generated and/or analysed during the current study are available online ³.

References

- Bai, Z. and Blackwell, A. F. (2012). Analytic review of usability evaluation in ISMAR. *Interacting with Computers*, 24(6):450–460. DOI: <https://doi.org/10.1016/j.intcom.2012.07.004>.
- Borges, J. B., Juy, C. L., Matos, I. S. d. A., Silveira, P. V. A., and Darin, T. d. G. R. (2020). Player Experience Evaluation: a Brief Panorama of Instruments and Research Opportunities. *Journal on Interactive Systems*, 11(1):74–91. DOI: <https://doi.org/10.5753/jis.2020.765>.
- Bouwel, T. V. (2024). Cubism Presskit. Retrieved from <https://www.cubism-vr.com/presskit/index.html> on July 28, 2025.
- Campos, T., Castello, M., Damasceno, E., and Valentim, N. (2025a). An Updated Systematic Mapping Study on Usability and User Experience Evaluation of Touchable Holographic Solutions. *Journal on Interactive Systems*, 16(1):172–198. DOI: <https://doi.org/10.5753/jis.2025.4694>.
- Campos, T., Damasceno, E., and Valentim, N. (2024a). Usability and User Experience Questionnaire Evaluation and Evolution for Touchable Holography. In *Proceedings of the 26th International Conference on Enterprise Information Systems*, pages 449–460, Angers, France. SCITEPRESS - Science and Technology Publications. DOI: <https://doi.org/10.5220/0012564100003690>.
- Campos, T., Delabrida, S., Damasceno, E., and Valentim, N. (2025b). Evaluating Performance and Acceptance of the UUXE-ToH Questionnaire for Touchable Holographic Solutions. pages 641–648.
- Campos, T. P. d., Damasceno, E. F., and Valentim, N. M. C. (2023). Usability and User Experience Evaluation of Touchable Holographic solutions: A Systematic Mapping Study. In *IHC '23: Proceedings of the 22st Brazilian Symposium on Human Factors in Computing Systems*, IHC '23, pages 1–13, Maceio, Brazil. ACM. DOI: <https://doi.org/10.1145/3638067.3638071>.
- Campos, T. P. D., Hounsell, M. D. S., Damasceno, E. F., and Valentim, N. M. C. (2024b). Evaluating Usability, User Experience, and Playability of a Puzzle Game in Mixed Reality. In *Symposium on Virtual and Augmented Reality*, pages 61–70, Manaus Brazil. ACM. DOI: <https://doi.org/10.1145/3691573.3691584>.

³Data and Scripts: <https://doi.org/10.6084/m9.figshare.28942391.v1>

- Desurvire, H. and Wiberg, C. (2009). Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games: The Next Iteration. In Ozok, A. A. and Zaphiris, P., editors, *Online Communities and Social Computing*, pages 557–566, Berlin, Heidelberg. Springer. DOI: https://doi.org/10.1007/978-3-642-02774-1_60.
- Dünser, A. and Billinghurst, M. (2011). Evaluating Augmented Reality Systems. In Furht, B., editor, *Handbook of Augmented Reality*, pages 289–307. Springer New York, New York, NY. DOI: https://doi.org/10.1007/978-1-4614-0064-6_13.
- Dünser, A., Grasset, R., and Billinghurst, M. (2008). A survey of evaluation techniques used in augmented reality studies. In *ACM SIGGRAPH ASIA 2008 courses on - SIGGRAPH Asia '08*, pages 1–27, Singapore. ACM Press. DOI: <https://doi.org/10.1145/1508044.1508049>.
- Fast-Berglund, A., Gong, L., and Li, D. (2018). Testing and validating Extended Reality (xR) technologies in manufacturing. *Procedia Manufacturing*, 25:31–38. DOI: <https://doi.org/10.1016/j.promfg.2018.06.054>.
- Frata Furlan Peres, F., Nunes, F., Teixeira, J. M., Maurício, C. R. M., Conceição, K. P., and Yoshida, L. (2024). Methods for Evaluating Immersive 3D Virtual Environments: a Systematic Literature Review. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality*, SVR '24, pages 140–151, New York, NY, USA. Association for Computing Machinery. DOI: <https://doi.org/10.1145/3691573.3691595>.
- Hartson, R. and Pyla, P. (2012). *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Morgan Kaufmann, 1st edition.
- Ivory, M. Y. and Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516. DOI: <https://doi.org/10.1145/503112.503114>.
- Johnson, D., Deterding, S., Kuhn, K.-A., Staneva, A., Stoyanov, S., and Hides, L. (2016). Gamification for health and wellbeing: A systematic review of the literature. *Internet Interventions*, 6:89–106. DOI: <https://doi.org/10.1016/j.invent.2016.10.002>.
- Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In Holzinger, A., editor, *HCI and Usability for Education and Work*, Lecture Notes in Computer Science, pages 63–76, Graz, Austria. Springer. DOI: https://doi.org/10.1007/978-3-540-89350-9_6.
- LaViola, J. J. (2000). A discussion of cybersickness in virtual environments. *SIGCHI Bull.*, 32(1):47–56. DOI: <https://doi.org/10.1145/333329.333344>.
- Lee, B. G., Tang, H., and Wen, X. (2023). Exploring the Fusion of Mixed Reality and Digital Game-Based Learning: The Case of Puzzle Box Games for Education. In *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 1–8. DOI: <https://doi.org/10.1109/TALE56641.2023.10398389>.
- Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability interface*, 8(2):3–6.
- Merino, L., Schwarzl, M., Kraus, M., Sedlmair, M., Schmalstieg, D., and Weiskopf, D. (2020). Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009-2019). In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 438–451, Porto de Galinhas, Brazil. IEEE. DOI: <https://doi.org/10.1109/ISMAR50242.2020.00069>.
- Nielsen, J. (1993). *Usability engineering*. Academic Press, Boston.
- Pagano, D. and Bruegge, B. (2013). User involvement in software evolution practice: A case study. In *2013 35th International Conference on Software Engineering (ICSE)*, pages 953–962. ISSN: 1558-1225. DOI: <https://doi.org/10.1109/ICSE.2013.6606645>.
- Prado De Campos, T., Damasceno, E. F., and Valentim, N. M. C. (2024). Evaluating Usability and UX in Touchable Holographic Solutions: A Validation Study of the UUXE-ToH Questionnaire. *International Journal of Human-Computer Interaction*, pages 1–21. DOI: <https://doi.org/10.1080/10447318.2024.2400755>.
- Pranoto, H., Tho, C., Warnars, H. L. H. S., Abdurachman, E., Gaol, F. L., and Soewito, B. (2017). Usability testing method in augmented reality application. In *2017 International Conference on Information Management and Technology (ICIMTech)*, pages 181–186. DOI: <https://doi.org/10.1109/ICIMTech.2017.8273534>.
- Rhiu, I., Kim, Y. M., Kim, W., and Yun, M. H. (2020). The evaluation of user experience of a human walking and a driving simulation in the virtual reality. *International Journal of Industrial Ergonomics*, 79:103002. DOI: <https://doi.org/10.1016/j.ergon.2020.103002>.
- Ritterfeld, U., Cody, M., and Vorderer, P., editors (2009). *Serious Games: Mechanisms and Effects*. Routledge, New York. DOI: <https://doi.org/10.4324/9780203891650>.
- Roto, V., Obrist, M., and Väänänen-vainio mattila, K. (2009). User Experience Evaluation Methods in Academic and Industrial Contexts. In *Proceedings of the Workshop UXEM'09*, volume II, page 4 p, Uppsala, Sweden. Springer.
- Sharples, S., Cobb, S., Moody, A., and Wilson, J. R. (2008). Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems. *Displays*, 29(2):58–69. DOI: <https://doi.org/10.1016/j.displa.2007.09.005>.
- Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557. DOI: <https://doi.org/10.1098/rstb.2009.0138>.
- Slater, M., Usoh, M., and Steed, A. (1994). Depth of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 3(2):130–144. DOI: <https://doi.org/10.1162/pres.1994.3.2.130>.
- Swan, J. and Gabbard, J. L. (2005). Survey of User-Based Experimentation in Augmented Reality. In *Proceedings 1st International Conference on Virtual Reality*, pages 1–9, Las Vegas, Nevada, USA. Mira Digital Publishing.
- Veriscimo, E. D. S., Bernardes Junior, J. L., and Di-giampietri, L. A. (2020). Evaluating User Experience in 3D Interaction: a Systematic Review. In *XVI Brazilian Symposium on Information Systems*, pages

- 1–8, São Bernardo do Campo Brazil. ACM. DOI: <https://doi.org/10.1145/3411564.3411640>.
- Yáñez-Gómez, R., Cascado-Caballero, D., and Sevillano, J.-L. (2017). Academic methods for usability evaluation of serious games: a systematic review. *Multimedia Tools and Applications*, 76(4):5755–5784. DOI: <https://doi.org/10.1007/s11042-016-3845-9>.
- Zhang, T., Booth, R., Jean-Louis, R., Chan, R., Yeung, A., Gratzer, D., and Strudwick, G. (2020). A Primer on Usability Assessment Approaches for Health-Related Applications of Virtual Reality. *JMIR Serious Games*, 8(4):e18153. DOI: <https://doi.org/10.2196/18153>.