Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 17 No. 4 (2025) 59-69
jtec.utem.edu.my
DOI: https://doi.org/10.54554/jtec.2025.17.04.008

Towards a Framework for Multi-Metric Evaluation of AR Experience (AR UX): Presence, Embodiment, and Task Load

Kathy Wee Mi Lang, Mohd Adili Norasikin*
Pervasive Computing & Educational Technology (PET),
Center for Advanced Computing Technology (C-ACT),
Fakulti Teknologi Maklumat Dan Komunikasi (FTMK),
Universiti Teknikal Malaysia Melaka (UTeM), 76100, Durian Tunggal, Melaka, Malaysia.

Article Info

Article history:
Received Sep 23rd, 2025
Revised Oct 28th, 2025
Accepted Dec 15th, 2025
Published Dec 24th, 2025

Index Terms:
Augmented Reality (AR)
User Experience (UX)
Presence and Embodiment
Task Load Assessment

Abstract

This pilot study examines evaluation metrics for the experiential quality of augmented reality (AR), which plays a critical role in shaping the effectivenss of decision-making. Three established instruments were employed: the Igroup Presence Questionnaire (IPQ) to assess immersion, the Virtual Embodiment Questionnaire (VEQ) to capture ownership and control, and the NASA Raw Task Load Index (NASA-RTLX) to measure cognitive and physical demands. Findings revealed high reliability for the IPQ and VEQ subscales, indicating strong immersion and embodiment, while the RTLX reflected greater variability in workload perceptions. A Venn diagram of participant responses further highlighted how presence, embodiment, and workload intersect to shape overall AR experiences. These multidimensional metrics provide valuable insights into how AR applications can support confident, efficient, and user-centered decision-making in design contexts. The findings demonstrate that presence, embodiment, and task load collectively influence users' confidence in decision-making and performance when interacting with AR systems. These insights contribute to developing AR applications that better support real-world design and decision-making processes.

*Corresponding Author: adili@utem.edu.my

I. INTRODUCTION

Traditional methods of visualizing interior design options, such as printed catalogs, sample boards, or two-dimensional digital renderings, often make it difficult for users to anticipate the outcomes of their choices. These limitations can create mismatches between expectation and reality, leading to poor decisions and post-installation regret. The challenge is particularly evident in do-it-yourself (DIY) scenarios, where non-professional users must rely on their own judgment without expert guidance.

Augmented Reality (AR) offers a promising solution by allowing users to preview design options directly within their real environments using mobile devices. With the rapid growth of AR-enabled smartphones and tablets, such applications have become increasingly accessible, creating new opportunities for enhancing decision-making in interior customization tasks.

Despite the growing potential of AR, research on evaluating the quality of the AR user experience (AR UX) remains limited, particularly when assessments rely on minimal or simplified metrics. Since the effectiveness of AR in supporting user decision-making depends greatly on how users perceive and interact with the virtual environment, this study employs three well-established instruments to examine key experiential dimensions: presence, embodiment, and task load. These three metrics provide a multidimensional perspective on the usability and experiential quality of AR applications. While they do not measure decision-making directly, they capture the conditions that shape decision quality. Higher presence can strengthen trust in virtual previews, stronger embodiment can increase confidence in interactions, and lower task load can free cognitive resources for evaluating design options rather than managing interface complexities.

Presence, measured using the Igroup Presence Questionnaire (IPQ), reflects the user's sense of immersion and realism within the virtual environment. Embodiment, assessed through the Virtual Embodiment Questionnaire (VEQ), captures the sense of body ownership and control over virtual elements, which is critical for natural interaction. Task load, measured with the NASA Raw Task Load Index (NASA-RTLX), evaluates the cognitive and physical demands experienced during AR tasks.

Building on these considerations, this study addresses the following research questions:

RQ1: How do users perceive presence, embodiment, and task load when interacting with an AR-based design visualization task?

RQ2: What relationships exist between these metrics in influencing user decision-making?

RQ3: How can these findings inform a multi-metric evaluation framework for AR user experience (AR UX)?

This study uses a mobile AR application for wall tile selection as a practical and relatable testbed to examine metrics for evaluating AR UX. Tiles provide a suitable case because they represent a common yet consequential choice in home customization, where mistakes are costly and difficult to reverse. Within this context, a pilot study with N = 20 participants demonstrates the feasibility and reliability of applying presence, embodiment, and task load as a unified evaluation strategy. The findings highlight the potential of multi-metric assessment to capture the experiential quality of AR beyond task performance, offering a foundation for future research and informing the design of AR systems that support more confident, efficient, and user-centered decision-making.

II. LITERATURE REVIEW

This review situates the study within prior research on AR usability and user experience, emphasizing how presence, embodiment, and task load have been individually examined but rarely integrated. It highlights the limited attention to these metrics in mobile AR for interior customization, highlighting the need for a multi-metric approach. By drawing on this literature, the review establishes the gap that the present pilot study addresses.

While presence captures the user's sense of "being there," embodiment reflects how naturally users identify with or control virtual elements. Task load, in contrast, measures the cognitive and physical effort involved. In decision-making contexts, these dimensions are interrelated: high presence and embodiment can enhance spatial understanding, while excessive task load can hinder effective choices. Thus, evaluating AR UX through these combined metrics offers a more comprehensive understanding of users' decision-making performance.

A. Augmented Reality in Interior Design

AR has emerged as a transformative tool in interior design, enabling users to place virtual furnishings, textures, and layouts directly into their physical environments using mobile devices. Recent works demonstrate the growing popularity of AR for spatial visualization, especially in e-commerce and DIY home décor [1–3]. For example, Aparicio et al. [1] used ARCore to display furniture, allow users to extract furniture from images, reconstruct it in 3D, and view it in-room via AR. Similarly, Revathy et al. [4] highlighted AR's potential in visualizing furniture in real interiors, noting improvements in user decision-making and engagement.

Despite such advances, systematic evaluation of AR interfaces in interior applications remains limited. Most literature focuses on technological implementation, such as environment scanning and object rendering, rather than user experience. Merino et al. [5] conducted a systematic review of mixed and augmented reality evaluations and showed a split between technology-centric studies and human-centric studies, emphasizing the need for more user experience-focused research in real-world contexts.

B. AR Evaluation in Design and Education

AR's applications in design and architecture education offers insights into experience evaluation. Chang et al. [6] used mobile AR to teach interior layout to design students, evaluating its effectiveness through learning motivation models (ARCS), and found AR enhanced learning interest, confidence, and satisfaction. Hajirasouli and Banihashemi [7] conducted a systematic review of AR in architecture and construction education, examining pedagogical philosophies, techno-educational aspects, and content domains; they identified several gaps in how AR is embedded into architecture and construction curricula, especially in areas related to skills development and real-world contextualization.

Broader reviews in architectural visualization indicate that AR enhances spatial understanding but often lacks robust user-experience metrics or standardized evaluation frameworks across many studies [8–10].

Interestingly, though not interior-design-specific, studies provide relevant insights into mobile AR evaluation, particularly in user engagement and informal learning contexts. For example, Baker et al. examined mobile AR elements designed to enhance engagement among hearing-impaired museum visitors, highlighting critical usability and engagement factors such as perceived control and satisfaction [11]. Likewise, Pendit et al. [12] proposed and validated an instrument for assessing "enjoyable informal learning" using a mobile AR prototype at heritage sites, demonstrating robust methods for capturing user experience (UX) dimensions in real-world settings.

C. Evaluating AR Experience (AR UX): Presence, Embodiment, and Load

The quality of an AR experience is not solely defined by technological functionality but also by how users perceive, control, and manage the system. Presence, the sense of immersion or "being there", is critical. A systematic review of immersive environments found that higher presence is associated with greater engagement and, in many contexts, improved performance, positioning it as central to immersive interface research [13–15].

Embodiment, defined as feeling ownership and control over a virtual body, also plays a key role in user interaction, particularly in Virtual Reality (VR). Mejia-Puig and Chandrasekera [16] showed that while visually rich virtual body representations increase presence, they can also introduce additional cognitive load during design tasks.

Task load, as measured by NASA, encompasses mental and physical effort, urgency, frustration, and related stressors experienced during task performance. Although the NASA RTLX has been widely adopted in HCI research [17, 18], its application within AR for interior design remains limited. Considered alongside presence and embodiment, task load offers a complementary perspective for assessing user experience. However, empirical studies that integrate these three dimensions within design-oriented AR contexts, such as interior customization, are still limited.

D. Gap in AR Evaluation Methodologies

Although AR is increasingly leveraged in interior design, current literature lacks rigorous, multi-dimensional evaluation frameworks that focus on user experience quality rather than system performance. Systematic studies tend to focus on display and tracking technologies with limited attention to metrics that directly influence how users perceive and interact with design previews [5].

The coupling of presence (IPQ) [19], embodiment (VEQ) [20], and task load (RTLX) [21] can offer a more robust evaluation framework. While each metric is validated within HCI and VR research, their combined application in an AR-based interior design context remained unexplored, creating a clear opportunity for this study.

Although the IPQ, VEQ, RTLX have each been widely applied to evaluate presence, embodiment, and workload in AR/VR studies [19, 20, 22–26], we found no prior work that synthesizes these three validated instruments into a single Venn-diagram illustrating participant overlap across AR-experience dimensions. We therefore propose the Venn diagram-based multi-metric evaluation synthesis presented in Figure 7 to characterize combined patterns of presence, embodiment, and task load in our pilot study. This approach offers a potential framework for evaluating UX in AR, addressing a gap that requires further attention as summarized in Table 1.

Table 1
Summary of AR Evaluation Literature: Key Findings and Gaps

Area	Key Findings	Research Gap
AR Evaluation Reviews [5, 10]	Reviews show diverse evaluation methods for AR/MR and UX. Most emphasize system/technical metrics; fragmented use of UX instruments.	No standardized multi-metric evaluation framework. Presence, embodiment, and workload rarely integrated.
AR in Education (Interior Design, Architecture, Construction) [6, 7]	AR enhances learning, visualization, and engagement. Mobile AR shown effective for motivation and spatial understanding.	Focus on learning outcomes, not UX quality. Little evidence on embodiment, presence, or task load.
AR for Architectural Design / BIM Integration [8, 9]	AR/VR supports design review, visualization accuracy, and BIM-based workflows. Strong technical benefits identified.	Evaluation centered on functionality/technical integration. UX dimensions remain underexplored

This review supports the argument that presence, embodiment, and task load offer valuable, underexplored perspectives for evaluating mobile AR in interior customization. This pilot study addresses this gap by applying these metrics as a foundation for systematic AR evaluation, an important step toward advancing user-centered AR design.

III. SYSTEM DESIGN: THE FLEXITILES APPLICATION

Building on the literature reviewed, this pilot study addresses the identified gap by operationalizing the evaluation of mobile AR interior customization through the lenses of presence, embodiment, and task load. The proposed application, named FlexiTiles (as shown in Figure 1), was developed as a testbed to systematically apply these metrics in a practical and relatable scenario focused on wall tile selection.

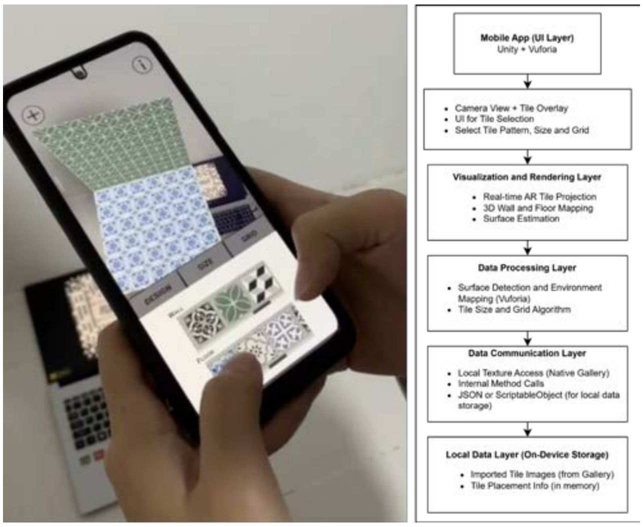
DESIGN
SIZE
GRID
Mobile App (UI Layer)
Unity + Vuforia
Camera View + Tile Overlay
UI for Tile Selection
Select Tile Pattern, Size and Grid
Visualization and Rendering Layer
Real-time AR Tile Projection
3D Wall and Floor Mapping
Surface Estimation
Data Processing Layer
Surface Detection and Environment Mapping (Vuforia)
Tile Size and Grid Algorithm
Data Communication Layer
Local Texture Access (Native Gallery)
Internal Method Calls
JSON or ScriptableObject (for local data storage)
Local Data Layer (On-Device Storage)
Imported Tile Images (from Gallery)
Tile Placement Info (in memory)

Figure 1. FlexiTiles User Interface and Architecture

The AR prototype was developed using Unity 6000.0.42f1 with the Vuforia Engine 11.1.3 for real-time tracking and projection. Cross-platform support was enabled through ARCore (v1.36.0) and ARKit (via Vuforia Engine 11.1.3), ensuring compatibility across Android and iOS devices. The fiducial marker used in the study measured 1.0 meter in width, providing a stable reference for surface detection and virtual tile alignment.

User testing was conducted on a Redmi Note 10S smartphone, representing a typical mid-range consumer device. The phone was powered by a MediaTek Helio G95 (MT6785V/CD) chipset, featuring an octa-core CPU (2×Cortex-A76 at 2.05 GHz and 6×Cortex-A55 at 2.00 GHz), and a Mali-G76 GPU optimized for mobile graphics. The device runs Android 13 (MIUI 14-V140) with OpenGL ES 3.2 support. This configuration was selected to reflect realistic performance conditions for mobile AR applications.

While tile customization is used as a case study, the broader contribution lies in demonstrating how AR systems can be designed and evaluated to capture critical user experience dimensions that influence decision confidence and usability.

The system architecture, shown in Figure 1, is organized into five layers that collectively support interactive and user-centered AR tile visualization for interior customization. Each layer contributes not only to technical efficiency but also to measurable user experience outcomes.

1) Mobile Application Layer (UI Layer)

Built on Unity with Vuforia integration, this layer provides the primary interaction point for users. Through the live camera feed, users can directly overlay virtual tiles onto their physical environment. Intuitive controls for selecting design patterns, adjusting tile size, and applying grid alignment reduce the cognitive effort typically associated with imagining spatial outcomes. These natural interactions enhance user spatial presence and engagement.

2) Visualization and Rendering Layer

This layer enables the projection of tile textures in real time, aligned with walls and floors in the physical environment. Through surface estimation and 3D mapping, virtual tiles appear contextually situated, which is essential for user trust and perceptual realism. The realistic rendering and alignment support the perception of being “inside” the augmented environment, reinforcing spatial presence and the feeling of ownership and agency over virtual objects, while reducing perceptual effort minimizing task load.

3) *Data Processing Layer*

Interpreting camera and sensor inputs, this layer detects surfaces and maintains spatial consistency during tile placement. Tile-size calibration and grid algorithms enhance precision. By ensuring accurate and consistent interactions, this layer minimizes errors and frustration, thereby reducing cognitive and temporal load as measured by TLX scores.

4) *Data Communication Layer*

Managing access to user-imported tile images and internal system data, this layer ensures that design selections and adjustments are immediately reflected in the AR environment. Immediate feedback and low-latency interactions improve responsiveness, supporting both spatial presence and user embodiment, while also streamlining the design process and reducing task load.

5) *Local Data Layer (On-Device Storage)*

This layer provides lightweight storage for tile textures and placement information, allowing the system to operate locally without reliance on network connectivity. By avoiding network delays, the layer maintains fluid interactions, reinforcing continuous presence and user control and minimizing unnecessary cognitive and temporal load.

By structuring the system around these user-centered considerations, the architecture optimizes both technical performance and interaction quality, directly supporting presence, embodiment, and low cognitive load, key evaluation metrics in this study.

Figure 2 presents the complete workflow of the AR wall tile customization process, from marker scanning and surface detection to final visualization. The figure illustrates how FlexiTiles operates not only as a functional customization tool but also as a research platform for systematically evaluating user experience in AR.

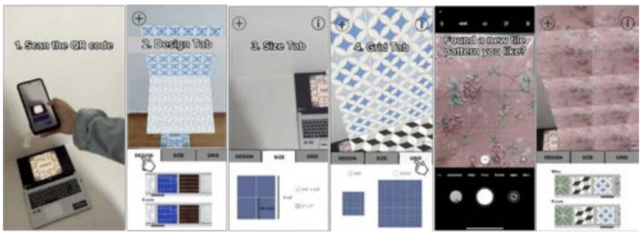

Figure 2. FlexiTiles: Steps to Use

The process begins with marker scanning, which establishes a reliable spatial anchor for the AR session. Once a surface is recognized via the smartphone camera, the system projects a grid overlay on both horizontal and vertical planes to support real-time interaction with tile assets. The grid can be dynamically adjusted in size and tile dimensions, offering flexibility for exploring different layout configurations. Users may also import custom images as textures, enabling them to preview personalized or branded tile designs directly in their physical environments. These features enhance immersion and embodied interaction, aligning with experiential dimensions emphasized in prior AR research.

Interaction is based on intuitive mobile gestures such as tap, swipe, and pinch, which simplify task execution while preserving functionality. This design approach improves accessibility and directly influences perceived workload, an important factor captured by task load assessments. By integrating marker scanning, dynamic grid calibration, and simple yet powerful gestures, FlexiTiles balances immersion with usability, providing a controlled context in which presence, embodiment, and task demand can be meaningfully examined.

IV. METHOD

A. *Participants*

A total of N = 20 participants (8 males, 12 females; age range 18–38 years, *M* = 24.0, *SD* = 6.4 years) were recruited for this pilot study. Participants were university students with basic familiarity with smartphones but no prior experience using AR-based interior customization tools. Recruitment was voluntary, and informed consent was obtained from all participants prior to the study. No monetary compensation was provided; however, participants were permitted to use the application freely during the test session. Figure 3 shows the test session involving participant interaction with the application.

Figure 3. Participants engaging with the AR application

B. *Task Design*

The experimental task was designed around the core functionality of the FlexiTiles application as illustrated in Figure 2. Each participant was asked to perform a wall tile customization activity using the AR interface on an ARCore- or ARKit-compatible smartphone. Specifically, participants were instructed to:

(1) detect a wall surface with a QR code, (2) pick the design in the library, (3) adjust grid size, (4) select tile dimensions, (5) finalize a preferred layout for preview (6) optionally replace tiles with a custom texture The task was intended to simulate a typical DIY scenario in which a user explores design alternatives before making a purchase decision. The duration of each session was approximately 15–20 minutes.

C. *Procedure*

Participants were welcomed individually and given a brief introduction to the purpose of the study. They were then provided with a demonstration of the application’s basic features. Following the demonstration, each participant completed the customization task independently. Upon completing the task, participants filled out three post-experience questionnaires (IPQ, VEQ, and RTLX) delivered on paper. The entire procedure for each participant lasted approximately 30–40 minutes.

D. *Instruments*

Three established user experience metrics were employed in this study: (1) IPQ [19], which assesses participants’ sense of presence and immersion within the augmented reality environment, with higher scores reflecting stronger spatial immersion; (2) VEQ [20], which measures perceived control, body ownership, and agency during tile placement, with higher scores reflecting stronger embodiment experiences; and (3) NASA-RTLX [21], which evaluates perceived task demands, including mental and physical effort and frustration. Lower NASA-RTLX scores indicate that participants experienced the task as less demanding and easier to perform. Table 2 summarizes the items for each metric: four items for IPQ, four for VEQ, and six for NASA-RTLX.

The IPQ was selected due to its established validity in measuring spatial and involvement aspects of presence in immersive environments, including AR. The VEQ was chosen for its sensitivity to capturing body ownership and agency in mixed-reality contexts. The NASA-RTLX was employed because it offers a reliable, multidimensional measure of mental and physical workload, which is essential in understanding user strain during interactive AR tasks. Collectively, these instruments provide a comprehensive coverage of experiential and cognitive dimensions relevant to AR interaction.

E. Analysis

The analysis proceeded in three stages. First, internal consistency of each instrument was assessed using Cronbach's alpha (α). Higher α values indicate stronger consistency among items. Second, measurement precision was evaluated using the standard error of measurement (SEM). Together, Cronbach's α and relative standardized SEM provided complementary perspectives on the reliability of the instruments. Finally, descriptive statistics (mean and standard deviation) were computed for each item and subscale, providing an overview of participant responses. This analysis was based on N = 20 participants in a pilot study, providing preliminary estimates of reliability that form an initial basis for evaluating the validity of these instruments, while acknowledging that the small sample size may limit the stability of the coefficients.

Table 2
Metrics and Intrument Items Summary of AR Evaluation

Instrument	Description
Igroup Presence Questionnaire (IPQ)	a) Somehow I felt that the virtual world surrounded me. b) I felt present in the virtual space. c) I still paid attention to the real environment. d) How real did the virtual world seem to you?
Virtual Embodiment Questionnaire (VEQ) with ownership component	a) The movements of the virtual body felt like they were my movements. b) I felt like I was controlling the movements of the virtual body. c) I felt like I was causing the movements of the virtual body. d) The movements of the virtual body were in sync with my own movements.
A Hart and Staveland's Raw NASA Task Load Index (RTLX)	a) How mentally demanding was the task? b) How physically demanding was the task? c) How hurried or rushed was the pace of the task? d) How successful were you in accomplishing what you were asked to do? e) How hard did you have to work to accomplish your level of performance? f) How insecure, discouraged, irritated, stressed, and annoyed were you?

1) Consistency (Cronbach's Alpha) Analysis

To assess the internal consistency of responses for each metric in this pilot study, Cronbach's alpha was calculated, ranging from 0 to 1, with higher values indicating better consistency. Cronbach's alpha (α) was calculated using Equation 1.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_{total}^2}\right) \tag{1}$$

where: k = Number of items
σ_i^2 = Variance of item *i*
σ_{total}^2 = Variance of the total score

2) Standard Error of the Mean (SEM) Analysis

The relative standard error of measurement, $\mathrm{SEM_{Relative}}(\%)$ represents the standard error of measurement as a percentage of the variability in the data, where smaller values indicate greater measurement precision and thus stronger reliability. The $\mathrm{SEM_{Relative}}(\%)$ is computed using Equation (2) – (4). For interpretation, values below 10% were considered high reliability, 10–20% good, 20–30% moderate, and above 30% low reliability.

The $\mathrm{SEM_{Overall}^{Relative}}(\%)$ is calculated using the Equation (5) – (8). The value expresses the overall measurement error as a percentage of the overall mean score. This allows comparison of reliability across different instruments or domains.

$$SEM_{Relative}(\%) = \frac{SEM}{\bar{x}} \times 100 \tag{2}$$

$$SEM = \frac{\sigma}{\sqrt{n}} \tag{3}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{4}$$

where: $\bar{x}$ = Mean
n = The sample size

$$SEM_{Overall}^{Relative}(\%) = \frac{SEM_{Overall}}{\bar{x}_{Overall}} \tag{5}$$

$$SEM_{Overall} = \frac{\sigma_{pooled}}{\sqrt{N_{total}}} \tag{6}$$

$$\sigma_{pooled} = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)\,\sigma_i^2}{\sum_{i=1}^{k}(n_i - 1)}} \tag{7}$$

$$N_{total} = \sum_{i}^{k} n_i \tag{8}$$

where: n_i = Sample size of a group
σ_i = Standard deviation of group *i*
K = Total number of groups
i = Index of the group ranging from *1* to *k*
σ_i^2 = Group variance

Since the IPQ and VEQ scores ranged from 1 to 7, while the NASA-RTLX scores ranged from 1 to 100, the RTLX values were rescaled to a 1–7 scale prior to computing the

$\mathrm{SEM}_{\mathrm{Overall}}^{\mathrm{Relative}}(\%)$. The rescaling was performed using Equation (9).

$$x' = \frac{(\,x\, -\, a\,)}{(\,b\, -\, a\,)} \times (\,d\, -\, c\,) + c \qquad (9)$$

where: *[a, b]* = Original scale
[c, d] = New scale

3) *Descriptive Analysis*

Descriptive analysis was conducted using means, standard deviations (SD), and standard errors (SEM) to summarize responses. A radar chart profiled presence, embodiment, and task load jointly, while an exploratory Venn diagram mapped their co-occurrence, providing a multidimensional view of AR user experience.

The mean indicated the central tendency, whereas the SD reflected the variability or consistency of responses. For both 1–7 scale instruments (e.g., IPQ, VEQ) and the 1–100 scale instrument (NASA-RTLX, rescaled proportionally), the following interpretive bands for response consistency were applied:

- Strong: SD ≤ 1.0 (for 1–7 scale, ≈16.7% of range) or SD ≤ 16 (for 1–100 scale, ≈16% of range)
- Moderate: SD between 1.1–1.5 (for 1–7 scale, ≈18–25% of range) or SD between 17–25 (for 1–100 scale, ≈17–25% of range)
- Weak: SD 1.6 (for 1–7 scale, 26.7% or more of range) or SD 27 (for 1–100 scale, 27% or more of range)

This proportional scaling approach facilitates comparison across instruments with different scoring ranges and serves as a heuristic rather than a strict statistical criterion

Prior work suggests that for Likert-type scales, SD values typically cluster around 20–25% of the total range, with values exceeding 25% reflecting greater dispersion of responses [27]. Moreover, research in psychology emphasizes that SD values naturally increase with scale range, underscoring the importance of scale-adjusted thresholds when comparing instruments with different ranges [28]. Hence, these interpretive bands are adopted as practical guidelines for assessing response consistency.

V. RESULT AND ANALYSIS

A. *Cronbach's Alpha*

Cronbach's α (0–1) was used to check the consistency of each metric, with higher values indicating better reliability. The IPQ scale yielded an α of 0.64, which indicates moderate consistency. While this value suggests that the items are reasonably consistent, it also reflects that presence measurement may be influenced by variability in participant responses, a result that is not uncommon in small pilot studies.

In contrast, the VEQ scale demonstrated an α of 0.91, indicating excellent reliability and strong internal consistency across its items. Similarly, the NASA-RTLX scale achieved an α of 0.89, which also reflects high reliability. These results suggest that both embodiment and task load were measured with high precision and stability, while presence was measured with acceptable consistency. However, the presence metric may require refinement or additional items for stronger internal consistency in future studies. Overall, these results suggest that the metrics used are sufficiently consistent to support preliminary evaluation of presence, embodiment, and task load in this AR UX study

B. *Reliability Analysis (Standard Error of the Mean, SEM)*

The reliability was further assessed using the relative standard error of measurement, $\mathrm{SEM}_{\mathrm{Relative}}(\%)$, where lower values indicate higher precision and stronger reliability. Table 3 presents these reliability values. All IPQ subscales demonstrated high reliability, with $\mathrm{SEM}_{\mathrm{Relative}}(\%)$ values ranging from 4.35% (Attention to Real Environment) to 7.56% (Realism). This indicates that participants' responses on presence-related items were measured with stable precision.

The slightly higher value for Realism likely reflects the limitations of the Vuforia rendering engine, which provides stable tracking but limited photorealistic lighting and shading. Future studies could employ advanced rendering platforms such as Unreal Engine or Unity's HDRP to enhance perceptual realism and achieve a more accurate assessment of this dimension.

Similarly, the VEQ subscales consistently showed high reliability, with $\mathrm{SEM}_{\mathrm{Relative}}(\%)$ values between 3.52% (Synchrony) and 4.13% (Agency). These results suggest that the embodiment measures captured participants' perceptions with strong measurement precision.

In contrast, the NASA-RTLX subscales yielded $\mathrm{SEM}_{\mathrm{Relative}}(\%)$ values between 14.16% (Effort) and 17.76% (Frustration), which correspond to moderate-to-good reliability. Although less precise than the IPQ and VEQ results, these values remain within an acceptable range, particularly considering the broader 1–100 response scale used by the NASA-RTLX. Importantly, when rescaled to the 1–7 range for overall comparison, the NASA- RTLX produced an $\mathrm{SEM}_{\mathrm{Overall}}^{\mathrm{Relative}}(\%)$ of 1.61%, which falls within the high reliability range.

The slightly reliability observed in the RTLX results may plausibly be linked to the marker-based tracking workflow used (Vuforia). Prior work shows that tracking modality significantly affects perceived system behavior and user effort: vision/marker versus markerless and hybrid approaches differ in stability and user demands [29, 30]. Moreover, mobile-AR interaction studies note that markerless systems are often preferred as they reduce the need for extensive scanning and alignment by the user. These forced scanning/realignment actions (marker repositioning, repeated scans, waiting for reacquisition) likely raise temporal demand, effort, and frustration.

The results indicate that the IPQ and VEQ provided highly reliable measures of presence and embodiment, while the NASA-RTLX demonstrated good reliability across its subscales, with the rescaled overall score further supporting its robustness. The convergence in findings across the three instruments provides additional assurance of the reliability of the measurement strategy employed in this pilot study.

Table 3

Reliability analysis of subscales across IPQ, VEQ, and NASA-RTLX. IPQ and VEQ Response ranges: 1–7 (1 = Fully Disagree, 7 = Fully Agree) ; NASA-RTLX Response ranges: 1–100 (1 = Very Low, 100 = Very High)

Instrument	*Subscale*	$\bar{x}$	*SD*	*SEM*	*Rel. SEM*	*Rel.*
IPQ (Presence)	Presence	*5.25*	*1.16*	*0.26*	*4.96*	*High*
	Involvement	*5.15*	*1.14*	*0.25*	*4.94*	*High*
	Realism	*4.30*	*1.45*	*0.33*	*7.56*	*High*
	Real env. attention	*5.75*	*1.12*	*0.25*	*4.35*	*High*
VEQ (Embodiment)	*Ownership*	*5.20*	*0.89*	*0.20*	*3.85*	*High*
	Agency	*5.30*	*0.98*	*0.22*	*4.13*	*High*
	Causation	*5.60*	*0.99*	*0.22*	*3.97*	*High*
	Synchrony	*5.60*	*0.88*	*0.20*	*3.52*	*High*
RTLX (Task Load)	*Mental*	*36.6*	*26.6*	*5.94*	*16.2*	*Good*
	Physical	*30.7*	*22.7*	*5.08*	*16.5*	*Good*
	Temporal	*40.1*	*25,8*	*5.77*	*14.4*	*Good*
	Performance	*33.7*	*23.0*	*5.13*	*15.3*	*Good*
	Effort	*32.2*	*20.4*	*4.56*	*14.2*	*Good*
	Frustration	*31.2*	*24.8*	*5.54*	*17.8*	*Good*
Overall (NASA-RTLX scaled to 1-7 scores)		*4.67*	*1.26*	*0.08*	*1.61*	*High*

C. *Descriptive Analysis*

The following analysis examines presence, embodiment, and task load through descriptive statistics, multidimensional profiling, and exploratory mapping to establish their collective value in evaluating AR user experience.

1) Descriptive Statistics for AR UX Metrics

The descriptive statistics of the IPQ, VEQ, and NASA-RTLX subscales are summarized in Table 3 and illustrated in Figure 4. Interpretation of response consistency was based on the SD thresholds defined in the Method section, where lower SD values indicate stronger consistency (strong: ≤16% of range, moderate: 17–25% of range, and weak: ≥26% of range).

For the IPQ subscales, mean scores ranged from 4.30 to 5.75, with SD values between 1.12 and 1.45, corresponding to moderate response consistency. Spatial Presence (M = 5.25, SD = 1.16), Involvement (M = 5.15, SD = 1.14), and Attention to Real Environment (M = 5.75, SD = 1.12) reflected moderate consistency, while Realism (M = 4.30, SD = 1.45) showed slightly weaker but still acceptable consistency.

For the VEQ subscales, mean values ranged from 5.20 (Ownership) to 5.60 (Change and Synchrony), with SDs between 0.88 and 0.99. These values fall within the threshold for strong consistency, indicating that participants' embodiment-related responses were tightly clustered around the mean, reflecting stable perceptions.

For the NASA-RTLX subscales, mean scores ranged from 30.7 (Physical Demand) to 40.11 (Temporal Demand), with SDs between 20.4 and 26.6. These values indicate weak response consistency on the 1–100 scale, suggesting broader dispersion in workload ratings across participants. However, when scaled to the 1–7 format, the overall NASA-RTLX score yielded a mean of 4.67 (SD = 1.26), reflecting moderate consistency.

SEM values reinforced these interpretations: IPQ subscales (0.25–0.33) and VEQ subscales (0.20– 0.22) showed minimal error, highlighting the stability of group-level estimates of presence and embodiment. In contrast, NASA-RTLX produced larger *SEM* values (4.6–5.9), consistent with the multidimensional and subjective nature of workload assessment [29].

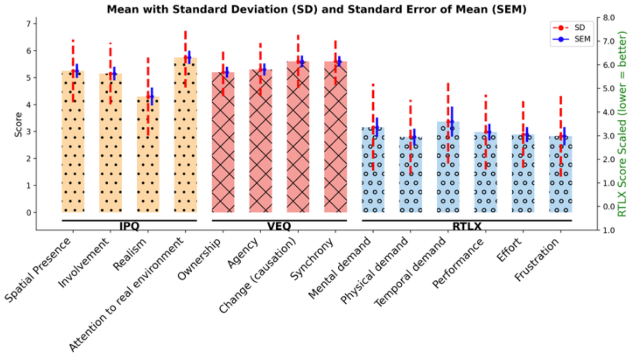

Figure 4. Graph of Mean, SEM and SD for IPQ, VEQ and RTLX

Figure 4 presents the mean scores with standard deviation (SD) and standard error of the mean (SEM) for the IPQ, VEQ, and RTLX questionnaires. The IPQ and VEQ subscales, displayed on the left of the graph, were scored on a scale from 1 to 7, with higher scores indicating a more positive experience. As shown, participants generally reported high scores across all subscales, suggesting a strong sense of presence and embodiment. Synchrony had the highest mean score among all subscales, while Realism had the lowest, likely due to the moderate rendering capabilities of the engine. The standard deviations, represented by red dashed error bars, indicate the variability in participant responses. The standard errors of the mean (SEM), represented by blue error bars, are small for both IPQ and VEQ, suggesting that the sample means are reliable estimates of the population means.

The NASA-RTLX subscales, shown on the right of the graph, were scored on a scale from 1 to 7, where lower scores indicate a better experience (less workload). The mean scores for all RTLX subscales were relatively low, suggesting that participants experienced a manageable workload. The small SEM values (blue error bars) further support that the sample means are reliable.

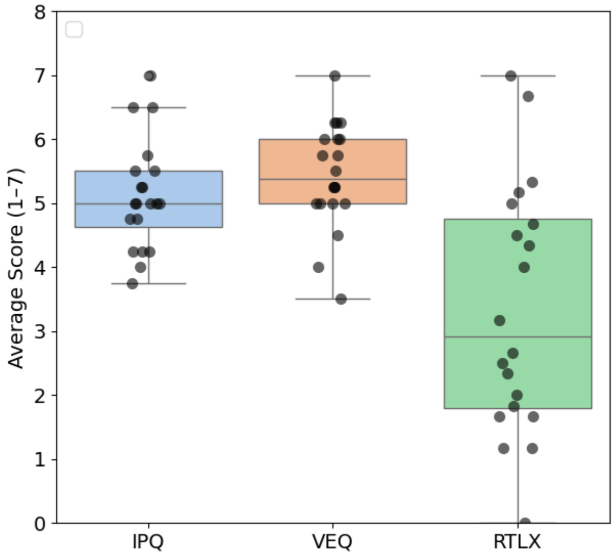

Figure 5. Distribution of Mean Scores Across IPQ, VEQ and RTLX

Figure 5 shows the distribution of mean scores across IPQ, VEQ, and RTLX, providing insights into consistency across the three dimensions. The IPQ exhibited moderate consistency, likely because participants differed in how they experienced spatial presence and realism, common in AR tasks where immersion varies by environment and tracking stability. The VEQ showed strong consistency, indicating more uniform embodiment experiences, while the RTLX showed weaker but interpretable variation, reflecting differences in perceived task effort.

These descriptive results highlight distinct but complementary patterns across the three metrics. IPQ demonstrated moderate consistency, VEQ showed strong consistency, and RTLX reflected weaker but interpretable consistency at the group level. Together, these outcomes confirm that each instrument captures a meaningful dimension of the AR user experience with sufficient reliability for research use. More importantly, the integration of presence, embodiment, and task load provides a balanced view of how users engage, immerse, and cope with AR tasks. These findings support the suitability of these measures as foundational components in ongoing work toward a framework for multi-metric evaluation of AR UX.

2) *Multidimensional Profile of AR UX Metrics*

Figure 6 presents radar charts comparing the mean scores and variability (SD) across the IPQ, VEQ, and NASA-RTLX subscales, providing a multidimensional visualization of the AR UX profile.

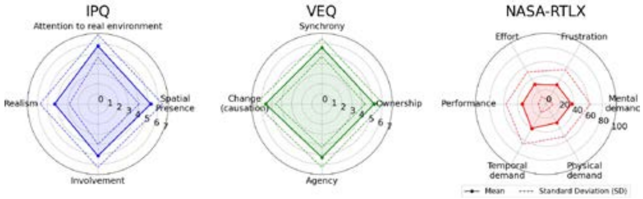

Figure 6. Radar Chart for IPQ, VEQ and RTLX (Mean ± SD)

For the IPQ, the subscales demonstrated consistently high mean ratings, with the highest score observed for Attention to Real Environment ($\bar{x}$ = 5.75) and the lowest for Realism ($\bar{x}$ = 4.3). The standard deviation bands indicate moderate dispersion, particularly for Realism (SD = 1.45). Overall, the clustering around the upper range of the scale reflects stable and strong presence experiences.

In the VEQ, mean values were uniformly high, ranging from 5.2 (Ownership) to 5.6 (Change and Synchrony). Variability across all four subscales was minimal (SD ≈ 0.88–0.99), highlighting strong consistency in embodiment-related judgments. The radar plot shows a nearly symmetric polygon, underscoring balanced and consistently positive evaluations across embodiment dimensions.

The NASA-RTLX subscales yielded moderate mean scores between 30.7 (Physical Demand) and 40.1 (Temporal Demand) on the 1–100 scale (equivalent to ≈ 2.8–3.4 on the 1–7 rescaled metric). Although these means indicate moderate workload levels, the wider standard deviation bands (20.4–26.6) reflect the complexity of workload as a construct that spans multiple demand dimensions.

The radar chart illustrates a multidimensional profile of AR UX that integrates presence, embodiment, and task load into a single view. This visualization goes beyond reporting isolated scores; it highlights how different experiential dimensions can be jointly considered when evaluating AR applications.

By making visible the balance between immersive qualities (IPQ, VEQ) and task demands (NASA-RTLX), the radar chart provides a practical tool for identifying strengths and areas for refinement in AR design. In this way, the profile directly supports informed decision-making for application improvement and contributes to ongoing work toward a framework for multi-metric evaluation of AR UX.

This radar chart exemplifies how presence, embodiment, and task load can be jointly profiled to evaluate AR applications. This multidimensional perspective enables developers and researchers to identify strengths and weaknesses across experiential dimensions, supporting more informed design decisions. As such, it provides a practical step toward a framework for multi-metric evaluation of AR UX, ensuring that AR applications can be refined to deliver more effective and user-centered experiences.

3) *Exploratory Venn Mapping of AR UX Metrics*

The Venn diagram in Figure 7 illustrates the distribution of participants across the three key AR experience metrics: IPQ, VEQ, and NASA-RTLX. To visualize the relationship between presence, embodiment, and workload, normalized IPQ, VEQ, and NASA-RTLX scores were plotted in a three-set Venn diagram. Scores were scaled to a 0–1 range, with thresholds of ≥ 0.6 for presence and embodiment, and ≤ 0.4 for workload. These cutoffs, corresponding to approximately ± 0.25 SD from the mean, distinguished high and low experiential states while minimizing mid-range ambiguity. By highlighting the upper and lower 40 % of the range, the visualization clearly differentiates experiential levels and demonstrates a multi-metric approach to a practical framework for AR experience evaluation. enhancing the clarity and interpretability of results.

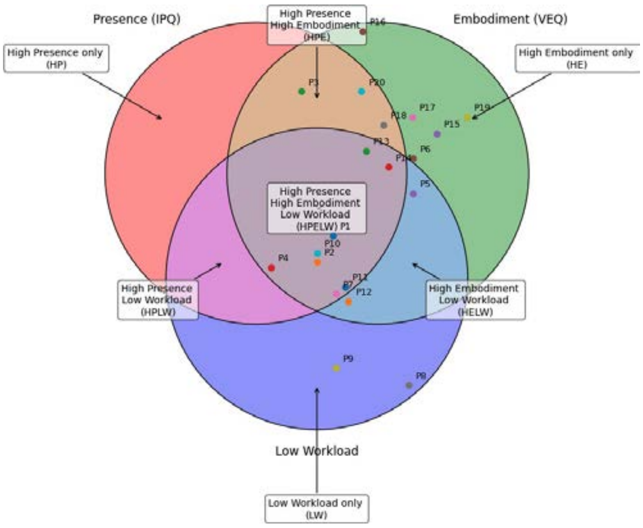
Presence (IPQ)
Embodiment (VEQ)
High Presence High Embodiment (HPE)
High Presence only (HP)
High Embodiment only (HE)
P16
P3
P20
P18
P17
P19
P15
P13
P6
P14
P5
High Presence High Embodiment Low Workload (HPELW) P1
P10
P2
P4
P11
P7
P12
High Presence Low Workload (HPLW)
High Embodiment Low Workload (HELW)
P9
P8
Low Workload
Low Workload only (LW)

Figure 7. Venn Diagram Map of Participants Across AR UX Metrics

The central overlap represents high presence, high embodiment, and low workload (HPELW). It consists of the majority of respondents. This group reflects the ideal user experience, where participants were immersed in the AR environment, embodied their interaction with the system, and perceived the task demands as manageable.

Some participants fell into dual-overlap regions. For example, those in the High Presence and Embodiment (HPE) cluster, but without Low Workload, were deeply immersed and embodied but found the system cognitively or physically demanding. Similarly, participants in the High Embodiment with Low Workload (HELW) region reported strong embodiment and manageable demands, but their sense of spatial presence was not as pronounced.

A smaller number of respondents were located in single-dimension clusters, such as the High Embodiment only (HE) or Low Workload only (LW) regions. These cases suggest that some participants experienced embodiment without immersion, or perceived reduced task demands but without a strong sense of presence.

Overall, this mapping provides a holistic picture of participant experiences, complementing the statistical analysis by visually depicting how presence, embodiment, and workload co-occurred at the individual level. It emphasizes the importance of adopting a multi-metric evaluation approach when assessing AR usability and user experience.

VI. DISCUSSION

The results from the three instruments highlight how different experiential dimensions contribute to the overall quality of the AR application as a decision-support tool for tile selection. Specifically, Presence (IPQ) assessed immersion and confidence in visual previews; Embodiment (VEQ) captured ownership and synchrony that underpin decision agency; and Workload (NASA-RTLX) reflected cognitive and physical demands influencing efficiency. Together, these metrics provide a multidimensional view of user experience, offering insights into how AR systems can be refined to better support decision-making in interior design contexts.

Participants reported consistently high presence and embodiment, confirming that the system enabled immersive and embodied interaction. In contrast, workload ratings showed greater variability, suggesting that while most found the interface manageable, some experienced higher task demands that may hinder efficiency.

The Venn diagram synthesis (Figure 7) complements these results by mapping how the three constructs co-occurred. Most participants fell into the intersection of high presence, high embodiment, and low workload (HPELW), representing the ideal experiential state for AR-based decision support. Non-ideal clusters such as HPE (immersed and embodied but with high workload) or HELW (embodied with manageable workload but low presence), highlight specific areas for design refinement. For example, simplifying interaction flows can reduce workload, while improving visual fidelity can enhance spatial immersion.

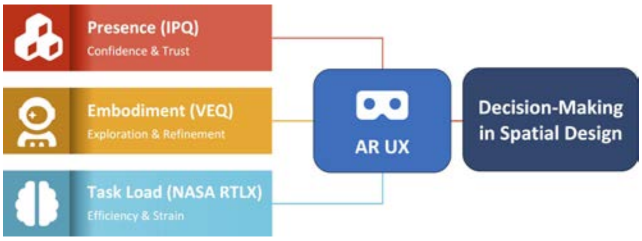

Figure 8. Conceptual Model for Decision-Making in Spatial Design

This exploratory Venn mapping therefore functions as a diagnostic tool: by showing which dimensions require attention, it guides developers in iteratively refining AR applications to move more users into the HPELW region. Figure 8 extends this perspective into a conceptual model, linking AR UX metrics directly to spatial design decision-making.

Unlike previous studies that assess presence, embodiment, and workload as separate constructs, this study works toward a unified multi-metric interpretive framework that visualizes their interaction using a Venn diagram. This integrative approach contributes to a more comprehensive understanding of user experience in augmented reality (AR) decision-making tasks, highlighting balanced states where immersion, embodiment, and cognitive effort converge to support more effective user performance.

The scope of this study focuses specifically on three key experiential quality metrics, namely IPQ, VEQ, and NASA-RTLX, as no prior research has proposed a unified evaluation framework encompassing these instruments. This paper serves as an initial attempt to examine their combined applicability and interrelations within a single study. Future work will extend this effort by exploring deeper correlations among these metrics to systematically develop a comprehensive framework grounded in the findings presented here.

Overall, this integrative analysis demonstrates the value of combining psychometric evaluation with visual mapping. Beyond confirming strengths in immersion and embodiment, it identifies workload as a critical factor to optimize. This multi-metric approach represents a step toward a structured framework for AR UX evaluation, helping developers design applications that not only engage users but also support confident, efficient, and user-centered decision-making.

VII. LIMITATIONS AND FUTURE WORK

This study has several limitations that should be acknowledged. First, the participant sample was limited to only 20 individuals, primarily students. While sufficient for a pilot study, the small and relatively homogeneous sample limits the generalizability of findings to wider populations such as professional designers or homeowners. Future studies should involve a larger and more diverse cohort to strengthen external validity.

Second, this study represents an exploratory step toward developing a multi-metric evaluation framework for assessing AR experience in decision-support contexts. While the proposed combination of immersion, embodiment, and workload measures provides useful insights, it does not yet constitute a comprehensive conceptual model of AR-supported decision-making. In particular, the interrelationships between the sub-metrics, for example, how presence interacts with embodiment or how workload modulates both, were not examined in depth in this pilot and remain an important direction for further investigation.

Third, the AR system tested in this study required manual scanning and marker tracking for tile placement, which may have contributed to increased physical and temporal workload for some participants. Future implementations could benefit from adopting markerless AR, automatic wall detection, or AI-driven alignment techniques to streamline interactions and minimize effort.

Finally, the study focused exclusively on individual use of the AR application. In practice, design decisions often involve collaboration between multiple stakeholders, such as clients, designers, and contractors. Exploring collaborative use cases, where multiple users co-experience customization through shared AR or cross-device platforms, would provide valuable insights into the social and professional applicability of AR in design workflows.

VIII. CONCLUSION

This study employed wall tile selection as a testbed to explore how presence, embodiment, and workload interact to shape AR experiences that underpin decision-making in interior design. Strong immersion (IPQ) and embodiment (VEQ) provided the perceptual and interactive grounding necessary for users to trust virtual previews and feel agency in their design choices, while workload (NASA-RTLX) emerged as a key factor influencing efficiency, as higher task demands risked diverting attention away from evaluating options to managing the interface.

The exploratory Venn mapping and radar profiles extend this analysis by offering a diagnostic lens: they reveal not only where the system succeeds, through high presence and embodiment, but also where improvements are needed, particularly in managing workload. By visually synthesizing these constructs, the approach moves beyond usability to show how experiential dimensions converge to support or hinder decision confidence, agency, and efficiency.

Although limited by its pilot nature and student-based sample, the study demonstrates the feasibility of employing structured, multi-metric evaluations in AR contexts. More importantly, it positions these metrics as a foundation for a future framework, where psychometric evaluation and visual mapping together can guide iterative refinement of AR applications. Ultimately, this multi-metric perspective helps developers move more users into the ideal experiential zone (HPELW), supporting confident, efficient, and user-centered decision- making in interior design and beyond.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to the Department of Interactive Media, Faculty of Information and Communication Technology (FTMK), the Pervasive Computing and Educational Technology (PET) research group, and the Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka, for their support and contributions to this study. The authors also extend their gratitude to fellow researchers who provided valuable insights and assistance throughout this work.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interests regarding the publication of the paper.

AUTHOR CONTRIBUTION

All authors contributed equally to the conception and design of the study, data collection and testing, analysis, and manuscript preparation.

REFERENCES

[1] R. A. G. Aparicio, J. J. R. Aliaga, and D. G. Q. Velasco, “Mobile application for the recommendation of furniture and appliances through augmented reality to improve the user experience in the online shopping process,” in Proc. of the 2022 3rd International Conference on Internet and E-Business, 2022, pp. 1–6, https://doi.org/10.1145/3545897.3545898

[2] H. Wei, L. Tang, W. Wang, and J. Zhang, “Home environment augmented reality system based on 3D reconstruction of a single furniture picture,” Sensors, vol. 22, no. 11, 2022, https://doi.org/10.3390/s22114020

[3] T. L. P. Do, K. Sanhae, L. Hwang, and S. Lee, “Real-time spatial mapping in architectural visualization: A comparison among mixed reality devices,” Sensors, vol. 24, no. 14, 2024, https://doi.org/10.3390/s24144727

[4] S. P. Revathy, A. Harini, and S. S. Sruthika, “Augmented reality in interior design,” Journal of Innovative Image Processing, vol. 6, pp. 305–313, 2024, https://doi.org/10.36548/jiip.2024.3.007

[5] L. Merino, M. Schwarzl, M. Kraus, M. Sedlmair, D. Schmalstieg, and D. Weiskopf, “Evaluating mixed and augmented reality: a systematic literature review (2009-2019),” in Proc. of 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2020, pp. 438–451.

[6] Y.-S. Chang, K.-J. Hu, C.-W. Chiang, and A. Lugmayr, “Applying mobile augmented reality (AR) to teach interior design students in layout plans: evaluation of learning effectiveness based on the arcs model of learning motivation theory,” Sensors, vol. 20, no. 1, 2020, https://doi.org/10.3390/s20010105

[7] A. H. Rasouli and S. Banihashemi, “Augmented reality in architecture and construction education: state of the field and opportunities,” International Journal of Educational Technology in Higher Education, vol. 19, 2022, https://doi.org/10.1186/s41239-022-00343-9

[8] J. G. Lee, J. Seo, A. Abbas, and M. Choi, “End-users’ augmented reality utilization for architectural design review,” Applied Sciences, vol. 10, no. 15, 2020, https://doi.org/10.3390/app10155363

[9] E. Yigitbas, A. Nowosad, and G. Engels, “Supporting construction and architectural visualization through BIM and AR/VR: A systematic literature review,” in Proc. of INTERACT’23, 2023, https://doi.org/10.48550/arXiv.2306.12274

[10] S. Graser, F. Kirschenlohr, and S. Böhm, “User experience evaluation of augmented reality: a systematic literature review,” in Proc. of CENTRIC 2024, 2024, https://doi.org/10.48550/arXiv.2411.12777

[11] E. J. Baker, J. A. Abu Bakar, and A. N. Zulkifli, “Mobile augmented reality elements for museum hearing impaired visitors’ engagement,” Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, no. 2-12, pp. 171–178, 2017.

[12] U. C. Pendit, S. B. Zaibon, and J. A. Abu Bakar, “Enjoyable informal learning at cultural heritage site using mobile augmented reality: Measurement and evaluation,” Journal of Telecommunication, Electronic and Computer Engineering, vol. 8, no. 10, pp. 13–21, 2016.

[13] J. J. Cummings and J. N. Bailenson, “How immersive is enough? a meta-analysis of the effect of immersive technology on user presence.” Media Psychology, vol. 19, no. 2, pp. 272–309, 2016, https://doi.org/10.1080/15213269.2015.1015740

[14] G. Makransky, T. S. Terkildsen, and R. E. Mayer, “Adding immersive virtual reality to a science lab simulation causes more presence but less learning,” Learning and Instruction, vol. 60, pp. 225–236, 2019, https://doi.org/10.1016/j.learninstruc.2017.12.007

[15] M. Slater, “A note on presence terminology,” Presence Connect, 2003.

[16] L. Mejia-Puig and T. Chandrasekera, “The presentation of self in virtual reality: A cognitive load study,” Journal of Interior Design, vol. 48, no. 1, pp. 29–46, 2023, https://doi.org/10.1111/joid.12234

[17] N. Wenk, J. Penalver-Andres, K. A. Buetler, T. Nef, R. M. Müri, and L. Marchal-Crespo, “Effect of immersive visualization technologies on cognitive load, motivation, usability, and embodiment,” Virtual Reality, vol. 27, no. 1, pp. 307–331, 2023, https://doi.org/10.1007/s10055-021-00565-8

[18] M. Cloutier et al., “Augmented reality in extra-vehicular activities: optimizing alert detection and cognitive workload,” in Proc. of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, no. 1, pp. 1349–1352, 2024, https://doi.org/10.1177/10711813241276460

[19] T. Schubert, F. Friedmann, and H. Regenbrecht, “The experience of presence: factor analytic insights,” Presence: Teleoper. Virtual Environ., vol. 10, no. 3, pp. 266–281, 2001, https://doi.org/10.1162/105474601300343603

[20] D. Roth and M. E. Latoschik, “Construction of the virtual embodiment questionnaire (VEQ),” IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 12, pp. 3546–3556, 2020.

[21] S. G. Hart, “NASA-task load index (NASA-TLX); 20 years later,” in Proc. of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, no. 9, pp. 904–908, 2006, https://doi.org/10.1177/154193120605000909

[22] M. Pike and E. Ch’ng, “Evaluating virtual reality experience and performance: a brain based approach,” in Proc. of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry - Volume 1, 2016, pp. 469–474, https://doi.org/10.1145/3013971.3014012

[23] F. Buttussi and L. Chittaro, “Effects of different types of virtual reality display on presence and learning in a safety training scenario,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 2, pp. 1063–1076, 2018.

[24] C. Merz, M. E. Latoschik, and C. Wienrich, “Breaking immersion barriers: smartphone viability in asymmetric virtual collaboration,” in Proc. of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2025, pp. 1-7, https://doi.org/10.1145/3706599.3719814

[25] K. Batra et al., “XRXL: A system for immersive visualization in large lectures,” in Proc. of 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR), 2025, pp. 370–380.

[26] M. Warsinke, F. Vona, T. Kojić, J.-N. Voigt-Antons, and S. Möller, “Digital twins for extended reality tourism: user experience evaluation across user groups,” in Proc. of International Conference on Extended Reality, 2025, p. 22–41, https://doi.org/10.1007/978-3-031-97769-5_3

[27] J. Sauro, and J. Lewis, “How to Estimate the Standard Deviation for Rating Scales,” Accessed: Sep. 9, 2025, [Online] Available: https://measuringu.com/rating-scale-standard-deviations/

[28] R. Takiar, “The relationship between the SD and the range and a method for the identification of the outliers,” BOMSR, vol. 11, pp. 62–75, 2023.

[29] M. Hertzum, “Reference values and subscale patterns for the task load index (TLX): a meta-analytic review,” Ergonomics, vol. 64, no. 7, pp. 869–878, 2021.

[30] M. Sulistiyono, J.W. Hasyim, B. Bernadhed, F. Liantoni, and A. Sidauruk, “Comparative study of markerbased and markerless tracking in augmented reality under variable environmental conditions,” Journal of Soft Computing Exploration, vol. 5, no. 4, pp. 413–422, 2024.