This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Feeling the Stakes: Realism and Ecological Validity in User Research for Computer-Assisted Interventions

Sue Min Cho, Winnie Wu, Ethan Kilmer, Russell H. Taylor, and Mathias Unberath

Johns Hopkins University, Baltimore MD, USA
scho72@jhu.edu

Abstract. User research is increasingly recognized as an essential strategy for ensuring the usability, safety, and effectiveness of emerging technologies in surgery. From a human-centered perspective, user studies are key to evaluating how technology-assisted interventions affect human behavior and system perceptions. For feasibility and scalability, these studies are typically conducted in controlled, desk-based lab settings. However, these settings often lack ecological validity, raising questions about how well they capture the actual surgical environment's emotional, perceptual, and interactive complexities. Previous work in human-centered assurance for image-based navigation, for example, described office-like laboratory studies where participants were asked to assess the adequacy of image-based 2D/3D registration, revealing that evaluators struggled to identify misalignments reliably. For that same task in robotic surgery, this study investigates whether–and how–the environment in which user studies are administered influences user behavior and performance. Specifically, we compare a conventional office-like lab to a high-fidelity mock operating room (mock OR) with an active robotic system, where the latter is contextually more relevant to the surgical task. Twenty-one participants first trained in an office, then were randomly assigned to either return to the office or proceed to the mock OR. Although task performance did not differ significantly, likely due to task difficulty, participants in the mock OR showed significantly higher interaction, perceived stakes, and NASA-TLX workload changes, despite completing the same task. These findings suggest that realistic, contextually relevant environments modulate user responses and behavior, with important implications for how user studies are designed, interpreted, and applied in computer-assisted interventions.

Keywords: Ecological validity · user-centered evaluation · computer-assisted intervention · surgical navigation · surgical automation · robot-assisted surgery.

1 Introduction

User research is increasingly recognized as essential for ensuring the usability, safety, and effectiveness of emerging technologies in image-guided surgery [3,

8]. From a human-centered perspective, user studies elucidate how technology-assisted interventions affect human behavior and system perceptions [9]. For feasibility and scalability, however, these studies are typically conducted in office-like labs. Yet, especially in the context of surgery, such controlled environments can often lack ecological validity.

Ecological validity—defined as the degree to which results obtained in a controlled setting generalize to actual real-world contexts [10]—is essential in ensuring that an experimental evaluation truly captures the conditions, stakes, and cognitive-emotional loads characteristic of real environments. Thus, ecological validity is a particularly important concept in surgical or medical context, where the differences between controlled, de-risked and real-world settings can be substantial. This challenge is known: Previous work in human-robot interaction (HRI), for example, has consistently revealed that laboratory studies risk oversimplifying complex real-world scenarios [1], potentially inflating usability metrics or downplaying genuine user stress. Although the value of realistic testing is widely recognized, logistical and financial barriers often prompt researchers to continue relying on simplified setups, thus limiting the external validity and clinical relevance of their findings [2].

In the context of surgery in general and human-centered assurance in image-based navigation in particular, prior studies investigated how humans assess the quality of 2D/3D registration and examined the influence of different visualization paradigms on user performance [4]. While these initial user studies demonstrated that human evaluators can detect misalignments on average, they also highlighted that human detection cannot be performed reliably without considering higher-fidelity and higher-stakes conditions.

Building on these insights, this paper explores how experimental environments themselves influence user experience and decision-making in computer-assisted interventions. Specifically, we focus on the "cost of failure" that looms large in surgical tasks: even minor errors can carry substantial clinical repercussions. Consequently, it is vital not only to measure system accuracy (e.g., 2D/3D registration quality) but also to understand how users behave under heightened stress and realistic OR conditions, and whether their behavior under stress differs from that observed in less realistic conditions.

Here, we present a comparative study evaluating users performing a 2D/3D registration assessment task in two distinct environments: (1) a conventional office-like laboratory setting and (2) a high-fidelity mock OR equipped with an active robotic system (Fig. 1). We aim to elucidate how environmental realism influences objective performance (e.g., accuracy, click behavior) and subjective experiences (e.g., perceived stakes, workload).

2 Hypotheses

H1 (Objective Measures):

H1a (Task Performance): There is a difference in main task performance (accuracy, specificity, sensitivity) between the two groups.

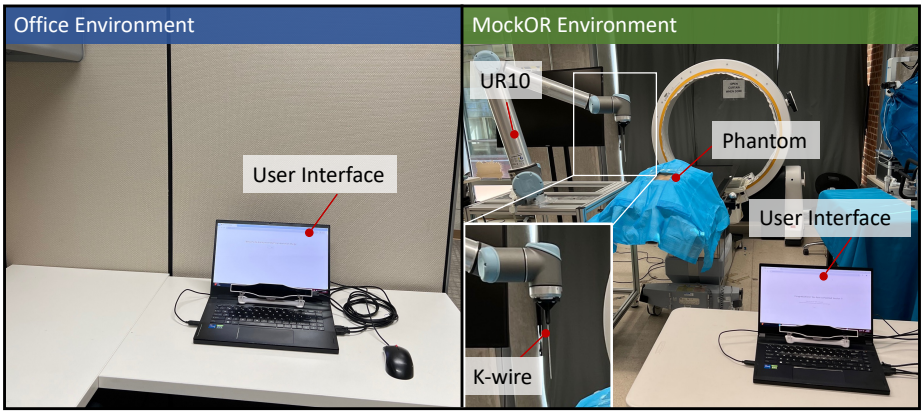

Fig. 1. Office environment (left), as a more conventional, non-clinical setting. Representative views of the two experimental environments. MockOR environment (right), featuring an active robot and a draped phantom to simulate clinical conditions.

H1b (Interaction Behavior): There is a difference in click counts between the two groups, reflecting differing levels of task engagement.

H2 (Subjective Measures):

H2a (Confidence): There is a difference in participants' confidence in their decisions between the two groups.

H2b (Perceived Stakes, Distress, Engagement, and Worry): There is a difference in perceived stakes, distress, engagement, and worry for the main task between the two groups.

H2c (NASA-TLX Delta): There is a difference in the change in NASA-TLX scores (Main Task – Training Task) between the two groups.

3 Methods

3.1 Participants

We recruited 21 volunteers from a local university population. After a manipulation-check item asked after the main task—"I was in a simulated operating room with an active robot" on a 5-point Likert scale—we excluded those who responded in a manner inconsistent with their assigned condition, yielding a final sample of $N = 9$ in the mock OR Mean age = 24.11 years (SD = 4.89) group and $N = 9$ 24.44 years (SD = 4.67) in the office group.

3.2 Experimental Protocol

All participants began with a training session (Session 1) in a standard office-like laboratory. Subsequently, each was randomly assigned to one of two conditions

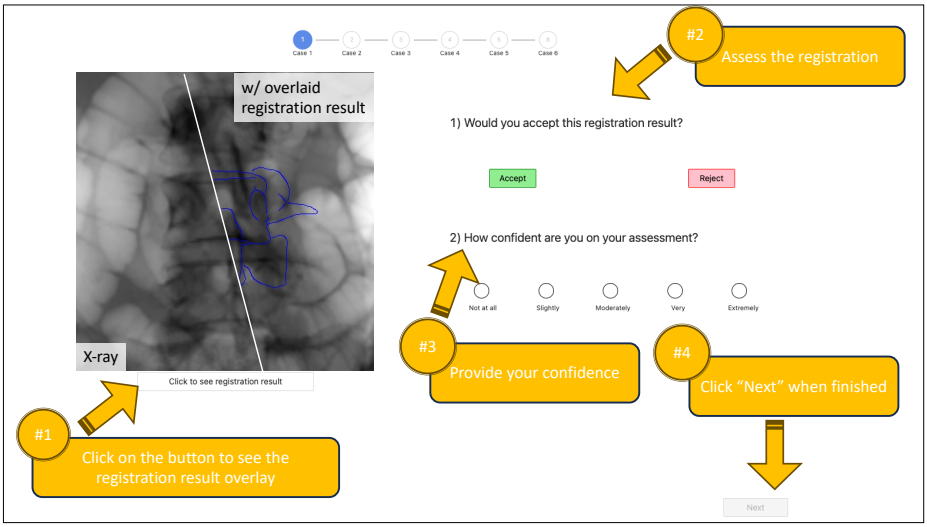

Fig. 2. Experimental user interface and interaction instructions provided to participants, with an example of an X-ray image and its 2D/3D registration result overlaid.

for Session 2: (1) to return to the office environment (Office group), or (2) move to a high-fidelity mock OR (MockOR group) equipped with an active UR10 robotic system. Across both sessions, participants performed single-vertebra 2D/3D registration assessment tasks on X-ray images with varying misalignments. These tasks were designed to simulate a vertebroplasty procedure in which precise needle placement into a vertebra is critical. An initial 6-case set was used in Session 1 for practice. Then, participants completed another 6 cases in Session 2 (with random ordering and offsets).

In the mock OR condition, a UR10 robot with a mounted K-wire adapter performed a predefined trajectory whenever the participant accepted the registration alignment, simulating an interventional step. In contrast, participants in the office environment saw only on-screen pop-up messages. After each trial, participants indicated their final accept/reject decision, confidence level, and, at the end of Session 2, completed a post-task survey assessing subjective stress and stakes. The study was approved by the local Institutional Review Board, and all participants provided informed consent.

Robot and User Interface We employed a 6-DOF UR10 manipulator (Universal Robots, Odense, Denmark) with a custom 3D-printed needle adapter. A custom Robot Operating System (ROS) script interfaced with our in-house developed, web-based user interface (built with Next.js) via a rosbridge websocket, sending joint configuration commands to the UR10. This user interface (Fig. 2) displayed the X-ray images and allowed participants to interactively click to see the registration result overlay before accepting or rejecting the registration. The

phantom-based setup allowed the manipulator to mimic a needle insertion upon user acceptance of the registration result. The user was seated outside of the robot's workspace, and a trained UR10 operator was present to ensure safety throughout all robot motions.

3.3 Measures

Objective Measures. We assessed two types of objective measures. First, Performance Metrics (H1a) were derived by comparing each participant's final decision (accept or reject) to a known ground truth (correct or incorrect registration), thereby yielding Accuracy, Sensitivity, and Specificity. Second, Interaction Behavior (H1b) was quantified through Click Counts, which captured how frequently participants interacted with the user interface and were hypothesized to reflect task engagement.

Subjective Measures. Several subjective measures were collected. Under Confidence (H2a), participants rated their confidence on a 5-point scale after each decision. For Perceived Stakes, Distress, Engagement, and Worry (H2b), we administered specific items related to Stake (decision carefulness, felt responsibility, perceived decision impact) to gauge how consequential participants perceived their decisions, and we adapted the Short Stress State Questionnaire (SSSQ) [7] to measure distress, engagement, and worry on a 5-point Likert scale. We also administered the NASA Task Load Index (NASA-TLX) [6] for both the training task and the main task, then computed a Δ-NASA-TLX composite score as (Main Task TLX) − (Training Task TLX) (H2c) to capture any changes in perceived workload across the two conditions. This approach is commonly adopted in human factors studies to isolate the effect of task manipulation from individual baseline biases [5]. Consequently, our primary focus is on the delta's ability to capture the shift in subjective workload from an easier version of the task to a more complex, realistic one.

3.4 Statistical Analysis

All analyses were performed in Python (pandas, SciPy). For each measure, we checked the normality (Shapiro–Wilk test) for each group (MockOR vs. Office) and variance homogeneity (Levene's test) if both groups appeared normal. Based on these checks, we used an independent t-test (equal variance or Welch's) if the data were approximately normal and Mann–Whitney U if the data violated normality assumptions. We report p-values and used an $\alpha = 0.05$ significance threshold. We also report effect sizes and used Cohen's d for parametric tests and rank-biserial correlation for Mann–Whitney U.

Table 1. Comparison of Key Objective and Subjective Measures between MockOR and Office Groups. Means (± SD) are shown, together with the statistical test, p-value, and effect size (Cohen's d for t-tests, rank-biserial r for Mann–Whitney U).

Measure	MockOR	Office	Test	p-value	Effect Size
Objective Measures (H1)					
Accuracy	0.73 ± 0.18	0.70 ± 0.15	t	0.66	d=0.20
Sensitivity	0.70 ± 0.33	0.57 ± 0.27	MWU	0.25	r=-0.29
Specificity	0.77 ± 0.27	0.83 ± 0.24	MWU	0.61	r=0.13
Click Counts	11.02 ± 5.44	5.45 ± 4.46	t	**0.022***	d=**1.12**
Subjective Measures (H2)					
Confidence	3.48 ± 0.51	3.33 ± 0.54	t	0.53	d=0.29
Stake (Careful)	4.56 ± 0.53	4.00 ± 0.00	MWU	**0.012***	r=**-0.56**
Stake (Responsibility)	3.78 ± 1.30	3.89 ± 0.60	MWU	0.89	r=-0.05
Stake (Impact)	4.33 ± 0.50	3.00 ± 1.00	MWU	**0.0062****	r=**-0.70**
Distress	2.48 ± 0.65	2.26 ± 0.52	MWU	0.53	r=-0.19
Engagement	4.38 ± 0.46	3.92 ± 0.51	MWU	0.12	r=-0.44
Worry	2.22 ± 1.12	2.11 ± 1.31	MWU	0.50	r=-0.20
NASA-TLX Δ	16.57 ± 7.59	7.59 ± 8.35	t	**0.030***	d=**1.13**

t = independent-samples t-test; MWU = Mann–Whitney U test;
n.s. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

4 Results

4.1 Does the environment influence users' objective measures? (Hypothesis 1)

H1a (Task Performance). Table 1 and the metrics excerpt below show that we observed similar accuracy between the MockOR group (M=0.73, SD=0.18) and the Office group (M=0.70, SD=0.15); an independent t-test indicated $p = 0.66$, with Cohen's $d = 0.20$, suggesting no statistically significant difference and a small effect size. Mean sensitivity for the MockOR group was 0.70 (SD=0.33), compared to 0.57 (SD=0.27) for the Office group. A Mann–Whitney U test yielded $p = 0.25$, rank-biserial $r = -0.29$, indicating no significant difference, though the MockOR group showed a slight numerical advantage. For specificity, the MockOR group averaged 0.77 (SD=0.27), whereas the Office group averaged 0.83 (SD=0.24). The Mann–Whitney test yielded $p = 0.61$, $r = 0.13$, again showing no significant difference. Taken together, these results fail to support **H1a**, as no significant difference in main task performance metrics was observed between the two groups.

H1b (Interaction Behavior). For mean click counts, MockOR participants had 11.02 clicks (SD=5.44), whereas Office participants had 5.45 clicks (SD=4.46). An independent t-test showed a statistically significant difference ($p = 0.022$) with Cohen's $d \approx 1.12$ (a large effect size). Hence, MockOR participants performed significantly more clicks, suggesting higher interaction or engagement in

that environment. This result supports the hypothesis that there is a difference in click counts, confirming **H1b**.

4.2 Does the environment influence users' subjective measures? (Hypothesis 2)

H2a (Confidence). As shown in Table 1, participants' self-reported confidence scores did not differ significantly. MockOR participants reported an average confidence of 3.48 (SD=0.51), whereas Office participants reported an average of 3.33 (SD=0.54) ($p = 0.53$, t-test, Cohen's $d = 0.29$). Thus, the result fails to support **H2a**, as no significant difference in confidence was observed.

H2b (Perceived Stakes, Distress, Engagement, Worry) Table 1 shows that the MockOR group reported higher ratings for perceived stakes than the Office group, with significant Mann–Whitney results for decision carefulness (MockOR M=4.56, SD=0.53 vs. Office M=4.00, SD=0.00, $p = 0.0124$, $r = -0.56$) and perceived impact (MockOR M=4.33, SD=0.50 vs. Office M=3.00, SD=1.00, $p = 0.0062$, $r = -0.70$), but not for felt responsibility (MockOR M=3.78, SD=1.30 vs. Office M=3.89, SD=0.60, $p = 0.89$, $r = -0.05$). In other words, participants in the MockOR condition felt a significantly greater need to be careful and perceived the impact of their decisions to be significantly higher than those in the Office condition, although the two groups did not differ in how responsible they felt for the outcome. There were also no significant differences between the groups in distress (MockOR M=2.48, SD=0.65 vs. Office M=2.26, SD=0.52), engagement (MockOR M=4.38, SD=0.46 vs. Office M=3.92, SD=0.51), or worry (MockOR M=2.22, SD=1.12 vs. Office M=2.11, SD=1.31), with p-values of 0.53, 0.12, and 0.50, respectively. Hence, **H2b** is partially supported: the environments differed in perceived stakes but not in distress, engagement, or worry.

H2c (NASA-TLX Delta) We computed Δ NASA-TLX by subtracting the scores for the training task from the main task. As shown in Table 1, the MockOR group reported a higher delta (M=16.57, SD=7.59) compared to the Office group (M=7.59, SD=8.35). Shapiro–Wilk tests confirmed normality for both groups, and Levene's test suggested equal variances, allowing for a standard t-test, which yielded $p = 0.0296$, $t = 2.39$, and Cohen's $d = 1.13$. The statistically significant difference in NASA-TLX deltas supports **H2c**, indicating a greater perceived workload change in the MockOR environment compared to the Office.

5 Discussion and Conclusion

We conducted a comparative user study in which participants performed a 2D/3D registration task either in a conventional office-like lab or in a high-fidelity mock OR with an active robotic system. The primary objective was to determine

whether, and how, heightened realism in the testing environment affects user performance, engagement, perceived stakes, and workload in computer-assisted interventions.

Our results indicate no significant differences in key performance metrics (accuracy, sensitivity, specificity). This aligns with prior work on the challenging nature of human-based registration assessment tasks—for instance, similar performance levels were observed in a pelvis-registration study [4], and here we see a comparable outcome in vertebra-focused registration. The consistent difficulty across both environments likely reflects the inherent complexity of such registration tasks, particularly when no additional technical assistance is provided.

Despite these equivalent performance outcomes, participants in the mock OR exhibited significantly higher click counts, perceived stakes, and NASA-TLX workload deltas, suggesting increased engagement and a heightened sense of consequence under more realistic conditions. Notably, there was no significant difference in confidence, distress, or worry, implying that while participants recognized the mock OR environment as more critical, this did not translate into heightened anxiety or self-doubt.

These findings underscore the importance of ecological validity in user studies for computer-assisted interventions. Although measured performance may appear similar in simplified lab settings, the presence of realistic cues (e.g., a functioning robotic system, clinically relevant surroundings) seems to amplify user engagement and elevate the perceived cost of errors. Such intensification of scrutiny can be advantageous for both system design and training protocols.

While our focus was on high-level performance and subjective responses to isolate the effect of environmental context, a more detailed analysis of user errors and interaction patterns could offer valuable insights in future work, particularly in understanding how different environments influence strategy and decision-making. Incorporating complementary modalities such as eye-tracking or physiological sensing could offer finer-grained insight into users' cognitive and affective states. Such signals can deepen our understanding of how environmental context shapes both user behavior and subjective experience beyond task performance.

Our sample size, while sufficient for observing notable effects, may have constrained the detection of subtler differences. Moreover, this study focused on a single 2D/3D registration task; future research could broaden the scope to multi-step procedures, dynamic navigation tasks, or tool manipulation in different surgical domains. Expanding the participant pool to include domain experts, especially in areas with well-characterized novice–expert distinctions, may also reveal important differences in perception, strategy, and adaptation to the environment.

Cost and resource constraints inevitably limit the availability of high-fidelity mock ORs, prompting the question of how to replicate or approximate these elements of realism in more accessible formats. Exploring alternatives such as Virtual Reality or modular simulation setups could elicit similar real-world behaviors without requiring full-scale facilities. Additional variables like time pres-

sure, ambient noise, or team coordination demands may further capture the complexity of actual operating rooms.

Overall, our study highlights that office-based user experiments, while pragmatic and easy to conduct, can underestimate the psychological and behavioral dynamics that emerge in higher-stakes settings. Incorporating realistic environmental elements into user research can produce engagement patterns and perceptions that are more indicative of genuine surgical scenarios. By approximating real OR conditions more closely, researchers and developers can better refine computer-assisted intervention technologies, ultimately fostering safer and more effective transitions from the lab to clinical practice.

Acknowledgments. This study was supported by a Catalyst Award from Johns Hopkins University.

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

References

1. Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., Belpaeme, T.: From characterising three years of hri to methodology and reporting recommendations. In: 2016 11th acm/ieee international conference on human-robot interaction (hri). pp. 391–398. IEEE (2016)
2. Berkowitz, L., Donnerstein, E.: External validity is more than skin deep: Some answers to criticisms of laboratory experiments. American psychologist **37**(3), 245 (1982)
3. Brown, E.J., Fujimoto, K., Blumenkopf, B., Kim, A.S., Kontson, K.L., Benz, H.L.: Usability assessments for augmented reality head-mounted displays in open surgery and interventional procedures: A systematic review. Multimodal technologies and interaction **7**(5), 49 (2023)
4. Cho, S.M., Grupp, R.B., Gomez, C., Gupta, I., Armand, M., Osgood, G., Taylor, R.H., Unberath, M.: Visualization in 2d/3d registration matters for assuring technology-assisted image-guided surgery. International Journal of Computer Assisted Radiology and Surgery pp. 1–8 (2023)
5. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50, pp. 904–908. Sage publications Sage CA: Los Angeles, CA (2006)
6. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. Elsevier (1988)
7. Helton, W.S.: Validation of a short stress state questionnaire. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 48, pp. 1238–1242. Sage Publications Sage CA: Los Angeles, CA (2004)
8. Lin, Z., Lei, C., Yang, L.: Modern image-guided surgery: A narrative review of medical image processing and visualization. Sensors **23**(24), 9872 (2023)
9. Manzey, D., Röttger, S., Bahner-Heyne, J.E., Schulze-Kissing, D., Dietz, A., Meixensberger, J., Strauss, G.: Image-guided navigation: the surgeon's perspective on performance consequences and human factors issues. The International Journal of Medical Robotics and Computer Assisted Surgery **5**(3), 297–308 (2009)

10. Subramanian, S., De Moor, K., Fiedler, M., Koniuch, K., Janowski, L.: Towards enhancing ecological validity in user studies: a systematic review of guidelines and implications for qoe research. Quality and User Experience **8**(1), 6 (2023)