Int. J. Human–Computer Studies 216 (2026) 103895

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Train like you're there: Use of immersive visualizations while training to operate a satellite improves outcomes in operations

Savannah L. Buchner [a,*], Jennifer Wood [b], Gabriella Schauss [a], Blake C. Bateman [a], Hanspeter Schaub [b], Allison P.A. Hayman [a]

[a] Bioastronautics Laboratory, Ann & H.J. Smead Department of Aerospace Engineering Sciences, University of Colorado, Boulder, CO 80303, USA
[b] Autonomous Vehicle Systems Laboratory, Ann & H.J. Smead Department of Aerospace Engineering Sciences, University of Colorado, Boulder, CO 80303, USA

HIGHLIGHTS

- VR training shows promise in improving remote supervision with traditional displays.
- VR training improved level 2 situation awareness (comprehension) during operations.
- VR training did not improve level 3 situation awareness (projection).
- VR training has higher perceived utility than training with a traditional display.
- Training with 3D visualizations improves operational performance.

ARTICLE INFO

Keywords:
Virtual reality
Supervisory control
Remote operations

ABSTRACT

Remote supervisory-based operations, like satellite operations, are becoming increasingly complex, creating a need for improved training. Challenges exist in training operators to have the necessary skills and understanding to perform complex satellite operations, such as on-orbit servicing. Virtual reality (VR), which creates a sense of immersion, has been demonstrated as a learning tool across many fields, however, it has not been investigated for remote supervision environments. This paper investigates the effects of immersion and 3D visualizations as a way to improve training for a satellite rendezvous mission. The transfer of skills after training in three different displays, but then performing the task in a traditional display is compared through human subject testing ($n = 45$) on measures of situation awareness (SA), workload, usability, performance, and subjective utility. The three displays include an immersive VR display, a computer screen with 3D visualizations, and a representative traditional display with only 2D elements. It was found that training in VR improved level 2 SA ($p = 0.006$) and increased usability ($p = 0.005$) during operational tasks. VR training also had higher perceived subjective utility, particularly in aspects such as understanding collision likelihood ($p = 0.022$) and orbital motion ($p = 0.041$). Training in 3D screen visualizations led to improved performance in more difficult scenarios ($p = 0.034$) and subjective understanding of event awareness ($p = 0.024$). No differences were found between workload and levels 1 and 3 SA. This research demonstrates the benefits of using VR as a training tool for complex supervisory operations and may lead to improved operational outcomes.

1. Introduction

Satellite operations are complex and unlike many environments we experience in our daily lives. Planned future spaceflight missions, such as satellite inspections or repairs, will create new difficulties for operators due to the high degrees of uncertainty, and the unintuitive and complex relative orbital motion. The close proximity of satellites in rendezvous operations also results in smaller time windows, possibly only seconds, in which the operator must select and initiate the appropriate procedure, which might include collision avoidance maneuvers (Sellmaier and Frei, 2022).

During these missions, operators act as remote supervisors. Here they monitor and send intermittent commands to an onboard computer, but often do not directly control the satellite. Remote supervision is

* Corresponding author.
Email address: savannah.buchner@colorado.edu (S.L. Buchner).

https://doi.org/10.1016/j.ijhcs.2026.103895
Received 29 July 2025; Received in revised form 28 June 2026; Accepted 14 July 2026
Available online 15 July 2026
1071-5819/

an important control modality for operations and exploration tasks beyond spaceflight, including transportation, robotics, and manufacturing. Beyond the challenges of spaceflight, remote operations can present additional issues to new operators due to the lack of environmental context (Chen et al., 2007) and understanding of spacecraft systems. This often leads to compromised situation awareness (SA), which could harm mission completion. SA can be defined as having 3 levels where level 1 is "the perception of critical elements in the environment", level 2 is "the comprehension of their meaning", and level 3 is "the projection of their status into the future" (Endsley, 1988). Due to the complexities and risks of current and planned satellite operations, it is important that operators have appropriate training, as failures can lead to the loss of spacecraft.

Training is defined as the process by which individuals can gain job-relevant skills, knowledge, or competency (Hawley et al., 2006). Effective training ensures operators have an accurate sense of system capabilities and adequate practice in putting together and understanding complex information, and can lead to the development of appropriate mental models. Mental models are "mechanisms whereby humans generate descriptions of the system purpose and form, explanations of the system functioning and observed system states, and predictions of future system states" (Rouse and Morris, 1985). Anticipating future states, enabling pattern recognition, and information-seeking have all been attributed to mental models (Morris and Rouse, 1985). Additionally, mental models can aid in improving SA, particularly at the higher levels (Endsley, 2021a). For many operations, especially in novel environments, it is critical for operators to include consideration of imperfect information and the possible impacts of that uncertainty in mental models of their actions. For example, in satellite operations, there may be uncertainties during a satellite burn or uncertainties inherent in a given sensor. These imperfections in knowledge of the current spacecraft state may affect the appropriate actions required and can impact future states. It is apparent that appropriate mental models will include an appreciation of uncertainties in the system, and that this must be taught as part of the training program. Inappropriate or inadequate training can make it more difficult for operators to build these mental models and maintain SA during the task (Hawley et al., 2006), which can harm mission performance.

Current supervisory control training often involves classroom education, simulators, and on-the-job training (OJT) (Hawley et al., 2006). Simulations and OJT are critical to build mental models, skills, and judgment to complete a task without errors in both nominal and off-nominal situations (Sheridan, 2012). Current spaceflight operations training varies between organizations, and may take between three months and multiple years to complete (Reiter, 2012; Liebowitz and Lightfoot, 1987; Stathopoulos and Oezdemir, 2018; Kolbeck et al., 2018; Decoust et al., 2018; Dittemore and Bertels, 2011). For example, command controller training at one satellite mission control facility consists of ten weeks of classroom-based lectures in conjunction with OJT (Reiter, 2012). In contrast, NASA space shuttle mission controllers often completed over 100 simulations, over the course of many years, to become certified (Dittemore and Bertels, 2011). However, these comprehensive training environments have some downsides. Simulations can be expensive and require many people to run. In addition, simulators can become obsolete compared to current technology due to the large cost of updates (Dittemore and Bertels, 2011). It has been suggested that incorporating new technology into training as an alternative to simulations can reduce total cost and time (Dittemore and Bertels, 2011).

One promising technology that can improve training is virtual reality (VR). VR has been studied and used for training in medicine (Seymour et al., 2002; Aïm et al., 2016), firefighting (Wheeler et al., 2021), human spaceflight (Garcia et al., 2020), education (Mayer et al., 2023), among other disciplines (Li et al., 2020; Xie et al., 2021). VR has been discussed for use in aerospace education, including for teaching in-orbit control operations (Meng et al., 2024). Additionally, VR has been proposed as a supplement to traditional training for satellite operations (Laskey et al., 2024), finding that VR simulations of a human repairing the satellite were a viable addition, but that no clear benefits were shown in measures such as usability and satisfaction. In both situations, the training effectiveness and skill transfer to operational displays were not studied, representing a gap. VR has been demonstrated to facilitate skill transfer to the real world in other, non-supervisory control contexts, Seymour et al. (2002), and lead to improvements in mental models (Thompson et al., 2021) and perceived learning (Makransky and Lilleholt, 2018). VR can provide a customizable, adaptable, and immersive environment to simulate tasks that might otherwise be costly, unsafe, or unintuitive to replicate in the real world (Xie et al., 2021). This allows users the opportunity to train for a variety of scenarios and in a variety of environments, both nominal and off-nominal. During VR training, trainees can make mistakes with no real-world risk to safety or expensive systems, and can build a better understanding of how actions and outcomes are connected. VR has also been shown to induce realistic stress responses, which can lead to operational realism and enhance training (Dammen et al., 2022).

Although using immersive displays for training and learning is generally promising across a wide variety of fields, little work has been done to apply these principles to remote supervisory operations, where the consequences of inappropriately conveyed uncertainty in this environment are profound for the operator's human perception and performance, and may be exacerbated due to the lack of environmental context. Further, while there has been an increased interest in VR for satellite operations and training (Sittner et al., 2023; Gad et al., 2023), or for other remote supervision paradigms (Lager and Topp, 2019), it may be difficult for some organizations to adopt VR displays for operations. This may be due to a variety of reasons, such as long lead times in technology transitions, increased adoption costs compared to including 3D visualizations on existing computers, difficulty incorporating the technology into classified environments, and the lack of strong benefits over other novel displays (Buchner et al., 2025b,a). If VR may not be feasible for operations, there are open questions about its utility for training, particularly regarding its impact on later performance in a traditional desktop environment.

Our objective is to compare training across display modalities with different degrees of immersion or 3D visualization on a satellite supervision task. The implementations include an immersive VR display, a 3D visualization on a 2D monitor, and a baseline display with traditional 2D graphical representations. After training, participants completed the satellite task in the baseline display, representing current satellite operations display modalities. This is to simulate the condition where VR is not used in operations, but may be used for training. Participant SA, performance, and workload during their second visit are compared to understand the effect of the training modality on task execution in the operational environment. Additionally, their responses to display usability and utility of training are compared. We hypothesize that training in the two environments with 3D visualizations will improve SA and performance in the second visit over those who trained in the baseline, traditional display, as well as have a higher perceived training utility. Furthermore, we hypothesize that VR will lead to further improvements in these metrics compared to those who were trained in 3D visualizations alone. We do not hypothesize there will be a difference in workload scores or usability based on training conditions, due to the fact that supervision is generally low workload and all operators will be using the same display.

2. Methods

In this work, three displays are designed and implemented for a satellite supervision task, where operators are trained on how to supervise a satellite during the proximity portion of a satellite rendezvous mission. The three display designs are compared through a human subject evaluation. This experiment has been described previously (Buchner et al., 2025a).

2.1. Scenario

Basilisk,[1] a high fidelity, flight-proven, physics-based satellite simulation tool, was used to develop the underlying trajectories used for the simulation (Kenneally et al., 2020). Satellite operations of this kind take hours; thus participants experience the simulated satellite rendezvous in faster-than-real-time, compressing the time by a factor of 15, to enable training and testing on multiple scenarios during a laboratory experiment. Compressed training has previously been found to provide greater realism to trainees (Miller et al., 1997), and has been shown to yield performance benefits during slow-continuous process simulations (Worm et al., 2012).

In every scenario, the set-up consists of two satellites in orbit about Earth: a non-operational client satellite and an active servicer satellite sent to inspect the client. The operator's goal is to service the client satellite or abort to avoid a collision. The client satellite is tumbling, non-operational, and cannot communicate with the Earth or the approaching servicer. This leads to uncertainties in its attitude and exact location.

The servicer satellite begins on a drift orbit passing by the client satellite. At some point, the servicer satellite performs a burn to enter a proximity orbit that relatively circles the client. The operator is able to select from one of three proposed locations, separated by 15min of flight time. The optimal burn location is influenced by lighting conditions, battery levels, and other factors. Different burn locations may lead to servicing the satellite or colliding. However, a thruster burn at any location can temporarily increase the uncertainty.

This burn uncertainty, combined with the location and attitude uncertainty of the client satellite, forms a keep-out zone. The servicer must avoid entering it to minimize collision risk. The keep-out zone grows and shrinks based on environmental factors and the actions initiated by an operator. The servicer satellite is equipped with a simulated proximity sensor. This sensor has better performance in the sunlight, causing the keep-out zone to be reduced; when it is in the Earth's shadow, the keep-out zone increases due to worsened sensor performance. To mitigate the degradation in sensor performance, the operator can turn on a 'light', which simulates activating an improved sensor to reduce the uncertainty. A consequence of employing the improved sensor is draining the battery at an increased rate, which limits the amount of time it can be used. Nominally, the satellite battery would drain in darkness and charge in sunlight.

Post burn, the servicer inspects the client satellite while avoiding entering the keep-out zone. For the purpose of the experiment, entering the keep-out zone counts as a collision, ending the trial. As the keep out zone includes the combined uncertainty in the spacecraft position and rotational measurements, that would exist in a real mission, entering it leads to a probability of collision. In real operations entering the keep out zone is a mission safety violation, and does not always lead to a physical impact; we elected to treat them the same to stress to participants the criticality of not allowing the satellites to enter the keep out zone. To avoid entering the keep out zone, an operator can elect to abort the mission provided they meet the fuel, battery, and time requirements.

Different scenarios are created with different initial orbits, dates, and lighting conditions, as well as a range of fuel and battery levels for the servicer satellite. Varying these parameters influences what burn locations are optimal, how often and when the light should be used, and if the operator will be forced to abort. Each trial lasts up to seven minutes; initiating an abort maneuver or colliding causes the trial to end earlier. Eight experimental scenarios and two training scenarios are developed. The difficulty level of the scenarios is varied based on the number of different ways to achieve a successful mission (i.e, any burn location or amount of light use that will lead to mission success would be characterized as easy, versus only one specific burn location and specific amounts of light use to succeed would be characterized as hard). In two of the trials, an abort is the most likely outcome.

2.2. Display design

Three different training displays are designed: a VR display, a 3D screen visualization (Scr. Viz.), and a Baseline display as seen in Fig. 1. All three displays contain the same information, but are presented to the operator in different ways. A Meta Quest 2 headset and controllers are used to interact with the VR display. The other two displays are designed for use with a computer monitor and mouse. The Baseline display mimics a traditional satellite telemetry display and is also the display used for experimental visits. All displays are built using the Unity 2022.3.21f development platform, and are based on previously developed displays (Buchner et al., 2025b).

The VR display can be seen in Fig. 1(a). The visualizations are based on the Vizard spacecraft simulation visualization software application (Wood et al., 2018), including satellite models, relative orbital motion, accurate Earth models, and appropriate sun-based lighting. Additional visualization elements include the keep-out zone, color-coded by collision risk, and an onboard light that is visible when the operator turns the 'light' sensor on to reduce uncertainty in the relative positions of the two satellites.

Beyond the 3D visualization, text is included in the VR display to convey critical information, alerts, and user inputs. A heads-up display (HUD) contains the range of the client servicer from the servicer, the rate at which their positions are converging or diverging, the time to the next thruster burn, the time to possible collision, as well as time to sunrise/sunset. This HUD panel is always located on the periphery of the headset. This allows operators to access critical information through eye movements, reducing information access effort (Wickens, 1993) and minimizing blockage of the visualization. Satellite states, like the fuel and battery levels, move with their associated satellite. This is done to minimize the amount of text in the HUD and preserve immersion (Iacovides et al., 2015; Rosyid et al., 2021; Marre et al., 2021; Caroux and Isbister, 2016). Cautions and warnings are displayed at the top of the screen, designed to be salient and easily noticeable by an operator. Alerts can be minimized to remove them from the field of view and maximized again (O'Hara and Fleger, 2020; Federal Aviation Administration, 2004). Alerts are triggered automatically for events such as low battery or collision potential. Finally, user input options are also displayed automatically when appropriate on a panel above the HUD.

The participant controls the visual elements (such as toggling items on and off) through a radial menu system and their hand controllers, as seen in Fig. 2. The radial menu system (Fig. 2a) can be used by the participant to change the viewpoint, view alerts, turn on or off the light, or abort. Critical operations, such as triggering an abort or toggling the light, can be done directly through physical buttons on the hand controllers, as outlined in Fig. 2(b). Triggering an abort maneuver requires a confirmation to ensure it is not an error. The left controller can be used to change the user's view through panning, zooming, and teleportation. Viewpoints can also be changed through natural movements, such as walking around the provided 10 by 10 foot area of floor or by turning their head while wearing the headset. The right controller is used to navigate the menu and submit user input. The display and controls underwent human factors testing before the experiment to ensure the text was readable and the controllers were acceptable.

The Scr. Viz. display maintains the same 3D visualizations as in the VR display but is not immersive (i.e., presented in a full field of view head-mounted display) and can be seen in Fig. 1(b). In the Scr. Viz., the 3D visualization is embedded as a panel within the larger display. All text, alerts, and requests for user input are arranged alongside the visualization panel to allow for a consistent scan pattern. The participant can interact with the 3D visualization using a mouse to pan (click + drag) and zoom (scrolling), thereby changing the viewpoint. Users customize and perform actions on the Scr. Viz. display by left-button clicking on

[1] https://avslab.github.io/basilisk.

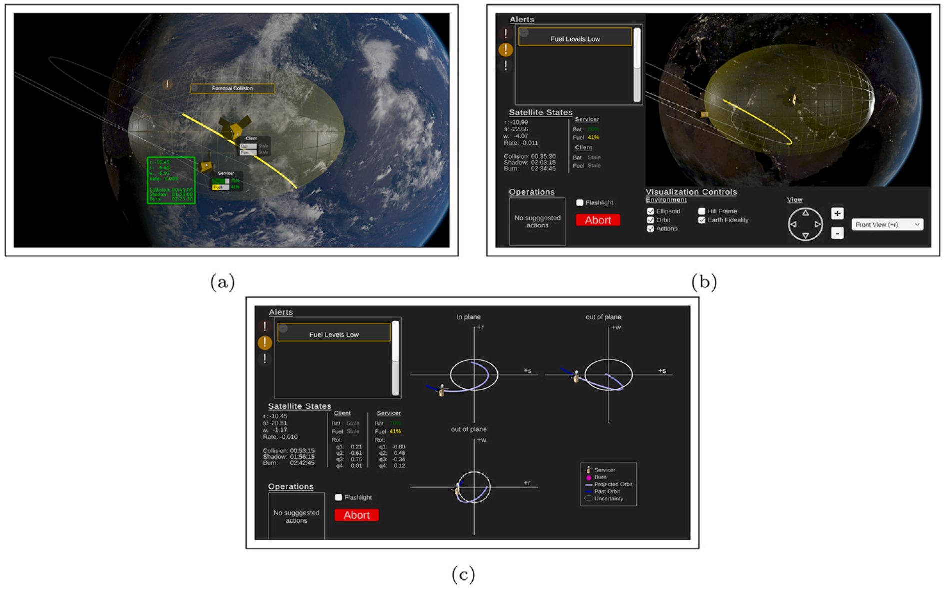

Fig. 1. The three different display designs participants are trained in: (a) VR Display. The green HUD is near the edge of the readable field of view. (b) Screen visualization display. (c) Baseline Display. This represents the traditional display and is the display used in the experimental visit.

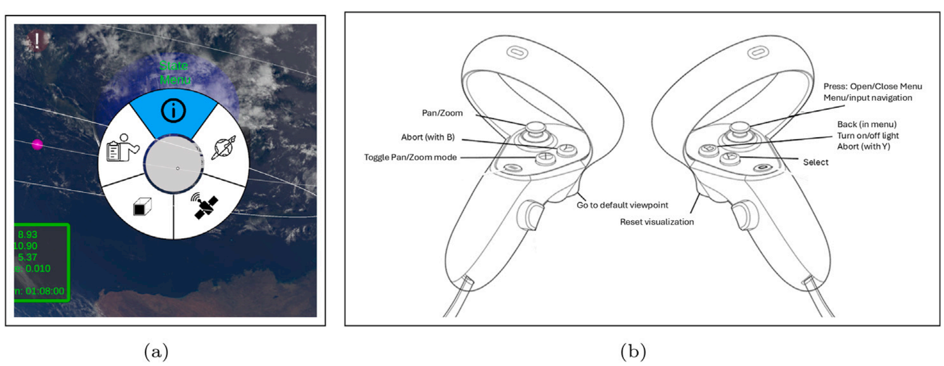

Fig. 2. User navigation and controls for the VR system: (a) a close up of the top level radial menu (b) the Quest hand controller mapping. From (Buchner et al., 2025a).

a mouse to interact with GUI elements including buttons, check boxes, scroll bars, and drop-down menus.

The Baseline display (in Fig. 1c) is designed to be consistent with traditional satellite monitoring displays. This means that operators primarily have to rely on telemetry presented in a text-based form and some 2D graphical representations. This display is similar to the Scr. Viz. display, but the 3D visualization panel is replaced by a series of plots containing the relative positions in the different orbital planes. Similarly to the Scr. Viz., all control interactions with the Baseline display are performed with a mouse, left-clicking to customize the view or select an action.

2.3. Experimental design

This study was approved by the University of Colorado Institutional Review Board (#24-0250). Forty-five participants (18 female, 27 male; ages 18–38, median age 23 years; 25 with prior orbital mechanics education experience; 15 with prior satellite operational experience; 21 graduate students, 17 undergraduate students, 5 with completed undergraduate degrees) completed the study; 2 participants voluntarily did not complete the study and thus were removed. Before enrolling, participants were screened to ensure their vision was correctable to 20/20, that they were not colorblind, and that they had a score less than 90% on the Motion Sickness Susceptibility Questionnaire (Golding, 1998). This was used to identify participants who might be highly susceptible to simulator sickness. Participants underwent two visits, each between one and eight days apart (average 3.5 days apart). The first visit was a training session. In this visit, participants were randomly assigned to one of three training conditions: Baseline, Scr. Viz., and VR. The second visit was the experimental visit, where participants were asked to perform satellite operations using the Baseline display (Fig. 1c), as would be consistent with displays for actual operators.

During the training day, informed consent was obtained from all participants; they were left naïve towards the alternate displays, but were aware of the high level project goals. Afterward, both visits followed a similar series of events, as seen in Fig. 3. Participants were then familiarized with the task and their assigned display. This was done through a PowerPoint presentation that reviewed the task's motivations and goals, as well as how to use their display. They then completed three familiarization trials. The first trial was primarily an opportunity to become accustomed to the controls and display, and in this trial, there was no collision risk present. The test operator ensured that test subjects practiced using all the controls and could read the displays. All participants ended the trial by performing an abort so they would be familiar with

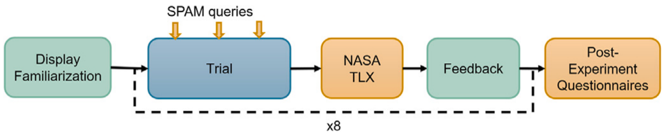

Fig. 3. Experimental design flowchart for a single visit. Both visits followed this flowchart. The Orange boxes indicate data collection through surveys or queries.

the abort process. The remaining two familiarization trials mimicked a real trial by including surveys and queries. The last two training trials had identical initial states, and participants were encouraged to make different choices in each to experience the wide range of outcomes resulting from their decisions. These initial trials were included to ensure that participants understood how to use the displays and perform the task before the real trials began and to eliminate issues of performance due to lack of familiarity with the display. Additionally, participants had to get 50% of the SA queries, including 1 per SA level, (described below), correct in order to move on.

After familiarization and practice, participants completed eight experimental trials. The scenarios in each trial were presented in a random order. During each trial, SA was measured using the situation presence assessment method (SPAM) (Durso et al., 1995). SPAM is a real-time SA assessment designed to mimic a control room environment and has previously been used in other experiments in operational settings, such as air traffic control (Durso et al., 1999; Loft et al., 2013, 2015; Cunningham et al., 2015; Mirchi et al., 2015; Fujino et al., 2020). During each experimental trial, participants would hear an auditory tone a maximum of three times (fewer if the trial was shortened due to the incidence of abort or collision) to cue participants. At this point, participants had up to twenty seconds to provide a verbal confirmation that they were ready to answer a SPAM query, preventing questioning from occurring while they were occupied making an input. The test operator then verbally asked three randomly selected questions, one for each SA level. Examples of questions include “Is the battery less than 50%”, “Is the servicer currently in the sunlight?”, “Will there be a collision alert in 30min?”. A full list of questions is in the appendix. Participants were instructed to respond as quickly and as accurately as possible and were allowed to reference the display to answer. For each question, the participants’ first verbal response was recorded, and the test operator recorded the accuracy and the time it took the participant to respond. Audio and video capture were used as a backup to ensure accuracy if the test operator was unable or unsure of the correct answer.

After each trial, participants completed the NASA Task Load Index (TLX) (Hart et al., 1988) on a computer to measure workload. During this, they rated the mental, physical, temporal, performance, effort, and frustration aspects of workload, each on a 21-point scale. Participants were then provided feedback on their performance, including information on whether they made an appropriate burn decision, and whether the end state was optimal.

After all trials were completed, participants completed the weighting comparison for NASA TLX (Hart et al., 1988). They then completed the System Usability Scale (Brooke, 1996), which is a 10-question survey that results in a combined usability score from 0 to 100. They also completed a custom Likert-style subjective utility survey about how their assigned display from the training day impacted their performance during the experimental visit; the full text is in the appendix. Additionally, they completed a demographics survey, which included information on their sex, familiarity with orbital mechanics, VR experience, and operational experience. These were coded as binary: little to none, or moderate to high. After the trials on the experimental visit, participants also completed the Balloon Risk Analog Task (BART) (Lejuez et al., 2002) to assess their risk aversion. Participants were compensated for their time and had the opportunity to receive cumulative performance bonuses (up to ± $1 per trial).

2.4. Statistical analysis

The primary objective of this paper is to understand how the display used in training yielded differences during operations. This is measured by the participant’s SA, performance, workload, usability, and subjective utility when operating on a traditional display. Comparing participants’ second experimental visit data, grouped by training display condition, allows for an understanding of whether training in alternative immersive modalities enables operators to gain context and intuition more easily, facilitating improved operations. The first experimental visit data of all subjects was assessed in previous work (Buchner et al., 2025a), by comparing outcomes of the task when performed on the different displays. This allowed for assessments and comparisons of the different displays alone, not their transferability.

In total, 360 trials over 45 participants were collected. For SA, each level was analyzed independently using the percent correct of queries answered across all trials, resulting in one data point per participant per SA level, for a total of 45 data points. Due to the overall lower accuracy, an accuracy based, instead of time based, analysis was conducted. Previous work (Bowden and Loft, 2013) has found that SPAM accuracy is correlated with other SA assessments. Likewise, for usability and utility, only one score was collected per participant, for a total of 45 data points per assessment type. For performance and workload, each trial yielded its own measure, so 360 data points were analyzed. A criterion of $\alpha = 0.05$ was used to assess the significance of all tests and all test assumptions were verified.

Before beginning the analysis, SA, performance, and workload were inspected to see if the trial order, particular scenario, or particular SPAM query had an effect. No effect of trial order, or SPAM query was identified, but for performance and workload, there was a dependence on specific scenarios (i.e., regardless of the order in which the scenarios were presented, some scenarios had consistently higher/lower workload and performance than others). Thus, scenario was included as a factor in the model. In addition, the effect of participant background was assessed, including sex, orbital mechanics knowledge, satellite operational experience, VR familiarity, and BART score as these may influence how a participant responds (i.e., orbital mechanics experience may result in better SA performance, and an increased risk tolerance may lead to different performance outcomes. These were included in the models as random effects, as appropriate, after visual inspection of the data.

For SA, linear mixed effects models were used to compare the effect of the training condition on SA, with the training display condition as a fixed effect. The random effects included are different per SA level. For level 1 SA, participants’ orbital mechanics experience and sex were relevant; for level 2 SA, their orbital mechanics experience; and for level 3 SA, their operational experience. The model was fit using the lme4 package in R via penalized maximum likelihood estimation (Bates et al., 2015). The significance of display modality was assessed using an F test with a type III ANOVA with a Satterthwaite approximation for degrees of freedom and was implemented using lmerTest package in R (Kuznetsova et al., 2017). If a significant main effect was found, post-hoc tests were conducted between all pairwise comparisons using estimated marginal means (emmeans package in R (Lenth et al., 2023)) with a Tukey p-value correction and Kenward-Roger degrees of freedom correction.

Performance is on a 12-point ordinal scale, as such; a cumulative linked mixed model approach was taken (using clmm in R (Christense, 2023)). This score encompasses the burn decision, other actions taken, and end state. The full metric calculation is described in the appendix. The training display and its interaction with trial difficulty were included as fixed effects, and the number of balloons collected in the BART task (mean: 17.6, standard deviation = 3.7, range 8-25) (Reed et al., 2012), which has been shown to be correlated with risk, was also included as a random effect. A significant interaction between training display and difficulty was found, indicating that training display modality may not yield consistent changes across all difficulty levels. Thus, we separated the data by difficulty level and analyzed each

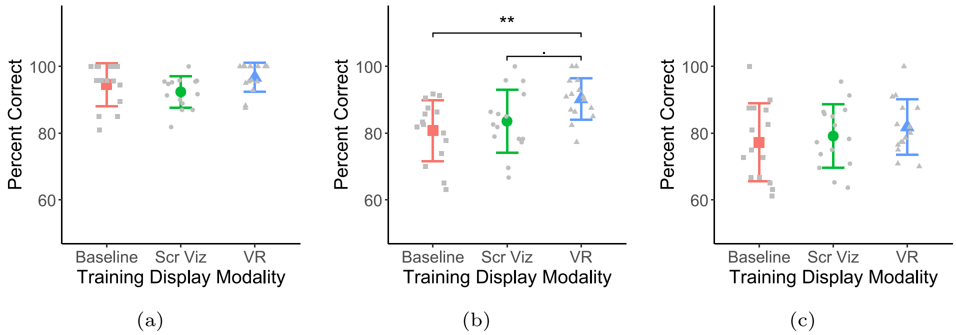

Fig. 4. The level 1 (a), 2 (b), and 3 (c) SA results. All figures show the participant averaged percent of SPAM queries at that level answered correctly. The data mean, standard deviation error bars, and significance are noted between the conditions.

separately, keeping the same fixed and random effects. A type III ANOVA was used to compare performance across conditions (anova.clmm in the RVAideMemoire package in R (HERVE, 2011)). Post-hoc tests were conducted between all pairwise comparisons using estimated marginal means (Lenth et al., 2023) with a Tukey p-value correction.

Additionally, the end states were compared across training conditions, including the number of scenarios aborted, the number of scenarios resulting in a collision, and the number of successful scenarios. This was done using a Kruskal-Wallis one-way analysis of variance.

For workload, the unique weighted NASA TLX score for each trial was analyzed. As with SA, a linear mixed effects model was fit using the training display condition as the fixed effect, with participant and scenario included as random effects to account for the repeated measures and differences across each scenario. Likewise, for usability, a linear mixed effects model was fit with the training display as a fixed effect and sex as a random effect.

The subjective utility questions were each on a 5-point Likert scale, requiring a nonparametric Kruskal-Wallis one-way analysis of variance. Post-hoc tests were conducted using Dunn's test with a Holm correction (Dinno, 2014).

3. Results

Our results show a significant difference in level 2 SA for the percent of questions answered correctly (F(2, 40.048) = 5.83, p = 0.006, η^2 = 0.23). Post-hoc comparisons found that those who trained in VR improved over those who trained in the Baseline display (t(41.4) = −3.22, p = 0.007, d = −1.20). Those who trained in VR approached, but did not reach, statistically significant improved performance compared to those who trained in Scr. Viz. (t(36.7) = −2.17, p = 0.09, d = −0.98). There was no difference between those who trained in Scr. Viz. and the Baseline screen condition (t(40.9) = −0.52, p = 0.86, d = −0.22). Level 1 SA trended toward, but did not reach, significance (F(2, 42) = 2.45, p = 0.09, η^2 = 0.11). There were no changes in level 3 SA (F(2, 41.61) = 0.73, p = 0.49, η^2 = 0.03). Fig. 4 shows the level of SA achieved for all groups across the three levels.

Performance is shown in Fig. 5. No difference was found between training modalities for the Easy (χ^2(2) = 3.84, p = 0.15) and Medium (χ^2(2) = 3.49, p = 0.17) difficulty trials. However, for the Hard difficulty trials, performance differed depending on the training condition (χ^2(2) = 6.78, p = 0.034). Post-hoc comparisons on the Hard difficulty trials only show differences between those trained in Scr. Viz. and VR (z = 2.55, p = 0.03), with those trained in Scr. Viz. performing better. No differences were found between Baseline and Scr. Viz. (z = −11.66, p = 0.22) and Baseline and VR (z = 0.89, p = 0.65). Additionally, there are no differences in the number of aborts (H(2) = 1.21, p 0.55, η^2 = −0.02), number of collisions (H(2) = 2.16, p = 0.33, η^2 = 0.003), or number of successes (H(2) = 0.84, p = 0.65, η^2 = −0.03) between participants trained in different conditions.

There are no significant differences in workload (F(2, 41.9) = 0.40, p = 0.67, η^2 = 0.02). Results are shown in Fig. 6(a).

Usability was significantly different depending on the screen condition in which participants were trained (F(2, 42.23) = 5.91, p = 0.005, η^2 = 0.21), as seen in Fig. 6(b). Post-hoc tests found participants rated the Baseline screen more usable if they had trained in VR than if they had trained in either the Baseline condition (t(41) = −2.90, p = 0.016, d = −1.06) or the Scr. Viz. condition (t(41) = −2.79, p = 0.021, d = −1.02). No differences were found between how participants rated the usability between those who trained in the Baseline and Scr. Viz. conditions (t(41) = −0.11, p = 0.99, d = −0.04).

Finally, for subjective utility (Fig. 7), there are differences between how people perceived the effectiveness of their training to help them understand the relative orbital motion (H(2) = 6.39 p = 0.041, η^2 = 0.10), understand collision likelihood (H(2) = 7.62, p = 0.022, η^2 = 0.13), and promote event awareness (H(2) = 7.36 p = 0.024, η^2 = 0.13). In addition, their perception of understanding uncertainties was trending toward significance (H(2) = 5.88, p = 0.053, η^2 = 0.09). Post-hoc tests found differences in understanding orbital motion between those who trained in the Baseline and VR displays (z = −2.47, p = 0.039), indicating VR was perceived to help participants understand orbital mechanics more, and no differences between Baseline and Scr. Viz. (z = −1.67, p = 0.19) or Scr. Viz. and VR (z = −0.81, p = 0.42). For promoting the understanding of collision likelihood, differences are found between Baseline and VR (z = −2.68, p = 0.022), indicating VR facilitated a perceived improved understanding of how likely a collision is to occur. No differences are found between Baseline and Scr. Viz. (z = −1.91, p = 0.11) or Scr. Viz. and VR (z = −0.76, p = 0.44). For both of these cases, VR had higher ratings than Baseline. Finally, for promoting event awareness, differences are seen between Baseline and Scr. Viz. (z = −2.73, p = 0.019), where visualizations were perceived to improve the participant's understanding of critical events. No differences are found between Baseline and VR (z = −1.41, p = 0.32) or Scr. Viz. and VR (z = 1.32, p = 0.19). Scr. Viz. had higher subjective ratings than Baseline.

Participants' subjective comments regarding how they believed their training display impacted their performance in the baseline display indicated that they found benefits in learning from VR and from the Scr. Viz. displays. For VR, participants' comments included: *"The overall orbit is much easier to see in VR, which is helpful."*, *"I felt like I was better able to visualize the trajectories based on the 2D graphs today [on the baseline display] because I had already seen the 3D equivalents"*, and *"My training on the first day helped me greatly in my decision-making process. I had developed a process on day one that was still applicable today. I looked for the same flight telemetry metrics to base my decisions on today as I did on day one. The first day also helped me to better visualize the 2D displays in front of me because I had already seen the 3D simulation and knew what to picture the graphs as."*. These were echoed among the other participants. Similar comments were reported among those trained with the

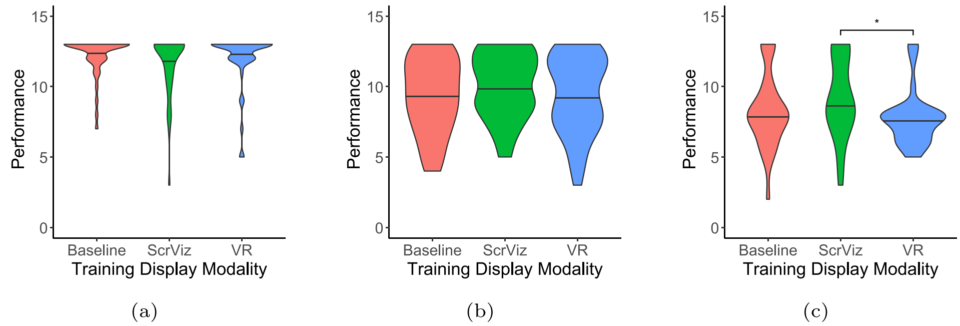

Fig. 5. Participant's performance on Easy (a), Medium (b), Hard (c) trials. The violin plot is shown with the median score, and significance is noted between the conditions.

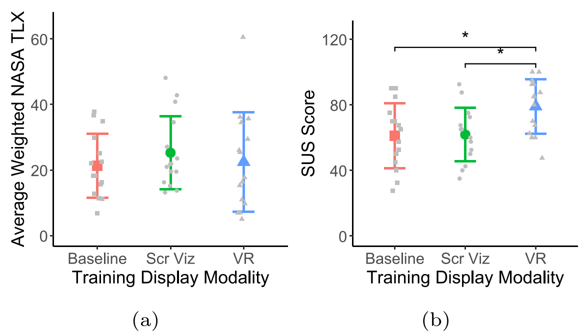

Fig. 6. The workload (a), and usability (b) results. Workload shows the participant-average weighted TLX score, and usability shows the system usability scale score. The data mean, standard deviation error bars, and significance are noted between the conditions.

3D visualization in Scr. Viz. *"I think the transition from 3D to 2D made it much easier to understand the 2D display than I feel it would have been if I could only see the 2D perspectives."*, *"The first day of training was more visual and enhanced my ability to understand orbit approaches and the associated nuances."* Participants did not provide comments about training in the baseline display.

4. Discussion

This study compares the effects of training display modality with different degrees of 3D visualizations and immersion for a satellite supervision task, and is one of the first studies to investigate training in VR for supervisory control tasks. This helps address the gap in understanding how VR may be used for satellite operation training, what benefits it may offer, and if training operators with VR might enhance their use of traditional desktop tools in their work. The Baseline display training condition is a control and helps to set a reference point for the familiarization that occurs with repeated interactions on the same system by using a traditional satellite operations display set-up, and is also consistent with current satellite operations and training. Further improvements over the Baseline display in any metric indicate additional benefits achieved from training with a system with enhanced displays.

Training in VR improves level 2 SA (comprehension) and approached significance in level 1 SA (perception), indicating a benefit for VR-based training. However, no improvement in level 3 SA (projection) was found. This is in partial support of our hypothesis that training in VR would improve all levels of SA. The lack of significance in level 1 SA is likely due to the fact that all groups had high SA perception scores, making it difficult to discern differences in perception across the groups. This indicates that all groups are able to achieve the high level of perception needed for operations. This may be due to the task and query design, as the types of perception questions that came from the goal directed task analysis could be answered without visualizations.

The difference in level 2 SA may be attributed to improved mental models of the operational task. Mental models are a contributing factor to SA, especially level 2 and level 3 (Endsley, 2021a). Previous work on VR training has not specifically studied the transfer of this training and how it affects SA in later, traditional operations, nor mental models of operations. However, biology education research has suggested that learning in VR leads to improved mental models of a concept over non-immersive displays (Thomas and Wickens, 2001); likewise, 3D visualizations have been shown to improve mental models over 2D for electron orbits (Chen et al., 2015). This could be one potential mechanism for how VR training is able to improve SA. Level 3 SA requires a more highly developed mental model, which often comes from increased experience and expertise (Endsley, 1996). It could be possible that the time spent undergoing training was not enough to develop the mental models sufficiently to improve level 3 SA, but enough for level 2 SA.

Being able to improve any aspect of SA through VR is important to achieve safer operations. High SA is critical to maintaining safety and performance (Endsley, 1990, 1996). Low SA has been attributed to many human-caused accidents in other fields like aviation (Jones and Endsley, 1996; Durso et al., 1998), maritime (Baker et al., 2021), and nuclear power plant operations (Solberg et al., 2023).

Differences in performance are seen among the Hard trials, where training in the Scr. Viz. display led to better performance than training in VR. In the Easy and Medium trials, participants had a high level of success in all training conditions, and no differences in performance were seen. While participants trained in VR performed worse than those trained with Scr. Viz. on the Hard trials, there was no difference in performance between those trained in Baseline or in VR. This is a positive, as it indicates that VR does not harm performance compared to traditional training displays, and so would not be detrimental to use. Some of these differences, or lack of differences, may be attributed to being unfamiliar with the display. While participants who were trained in VR had a familiarization session with the Baseline display, it is possible additional time could have yielded different results. The Scr. Viz. display had many elements in common with the Baseline, including the same interface to display alerts, data, and methods to take action; the only difference between these two displays was the visualization of the data. These differences in familiarity may have contributed to some of the differences seen in performance with the VR display. No differences between VR and traditional training on subsequent performance are consistent with findings in literature (Daling and Schlittmeier, 2024; Kaplan et al., 2021), which have found VR to often be as good as, but not better, for promoting performance. However, previous research has not considered 3D visualizations as an in-between, as was done in this study.

Finding no differences in workload is also a positive result and matches our hypothesis. Any increase in workload would be indicative of a poor training experience that left participants unprepared to

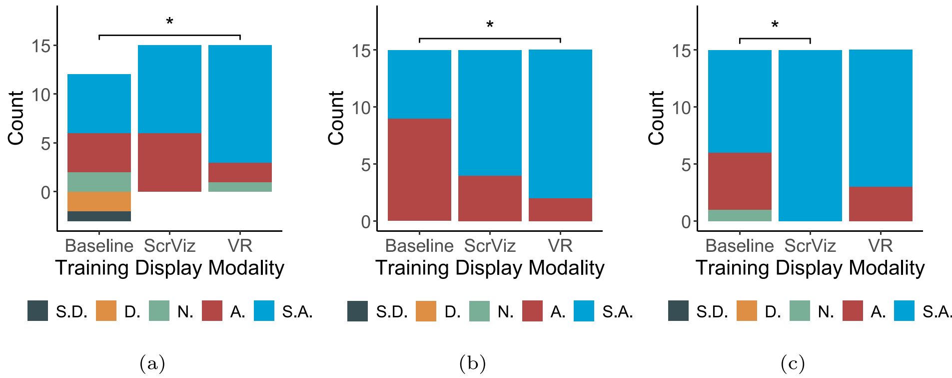

Fig. 7. The subjective utility questions, showing the results of (a) "I found that the training from the first visit was effective in enabling me to understand the relative orbital motion of the satellites today." (b) "I found that the training from the first visit was effective in enabling my understanding of collision likelihood today.", and (c) "I found that the training from the first visit was effective in enabling me to understand mission critical events today." (S.D. = Strongly Disagree, D. = Disagree, N. = Neither Agree nor Disagree, A. = Agree, S.A. = Strongly Agree).

understand the Baseline display without increased effort. Supervisory control and monitoring tasks are typically lower in workload (Hooey et al., 2018; Huey and Wickens, 1993; Grier, 2015) to begin with, and it was not expected that any one training modality could reduce this further.

The results in support of using VR in training also come from subjective participant assessments. Participants trained in VR gave the Baseline display higher usability scores compared to those trained in the Baseline display, contrary to our hypothesis. This indicates that the prior context developed by people trained in VR enabled them to synthesize and understand the traditional display more easily. How training modality influences the usability of a traditional display has not previously been studied, and represents an important finding of this study. The utility questionnaires and comments further highlight the potential of VR or 3D visualizations for training and help explain some of the differences in other metrics. These results also support the idea that VR and visualizations may be improving mental models of the scenario. As previously mentioned, future satellite operations may be more complex and require quicker decisions to avoid collisions and have harder to understand orbits (Selkowitz et al., 2017). Being able to promote understanding of aspects like collision likelihood or orbital motion, which VR did, is critical to improving operators' understanding of the scenario. These are reflected in the comments about the importance of understanding the 3D orbits before transitioning to operating on the Baseline display. In sum, these results indicate that those who trained in VR and Scr. Viz. perceived them as being very useful for understanding the task better in a traditional environment. These perceptions may also be indicative of the inferred improvements in mental models achieved through training in VR. This finding agrees with previous training studies where participants subjectively preferred VR (Al-Hor et al., 2024; Chen et al., 2025).

4.1. Limitations and future work

There are some limitations to the approach taken for this study. The training timeline, scenario, and task used are both sped up and simplified compared to actual operations. Here, training took place over one day instead of the typical three months to a year that is required for complex operations (Reiter, 2012; Liebowitz and Lightfoot, 1987; Stathopoulos and Oezdemir, 2018; Kolbeck et al., 2018; Decoust et al., 2018; Dittemore and Bertels, 2011). Despite this, the results show that VR is able to provide benefits from just one day of training, highlighting its potential. Furthermore, training sessions were not evenly spaced between participants, although they were bounded between one and eight days. No trends were observed between the duration between training sessions and any of the metrics, indicating these differences did not impact the results. Future work should include repeated training sessions, more consistent timing between sessions, or long time between sessions, to further understand how and when VR can provide an impact. This may also lead to differences in aspects like level 3 SA, which may come from an increased amount of time training. Additionally, the simplification of the scenario may influence some of the results; for example, in SA level 1 and the easy trials' scores are generally high regardless of modality, and no differences are found. These simplifications were required, though, to make the experiment feasible. The faster than real time speed of the trials may influence aspects such as reducing boredom, or adding stress, potentially influencing the measured outcomes. However, the speed was consistent between the two visits, allowing us to conduct a fair comparison of metrics across trials to understand differences based on training modalities. Future work should increase the complexity of these trials and should study the effect of training scenario speed on outcomes.

There are some limitations regarding the SA measurement, particularly when using an accuracy, not response time, based measurement. SPAM may have less sensitivity than other objective measurements, although it is equally predictive of performance, Endsley (2021b), influencing the results, which may also explain some of the lack of differences observed. Furthermore, future work should aim to capture mental models of the participants, not just correlate them to SA. This can aid in understanding how and whether the development and quality of mental models can play a role in the outcomes.

Additionally, while our participant pool drew from people with a background representative of many satellite operators the subjects had differing amounts of previous knowledge that may have affected their training experience. We included aspects of participant background as random effects in the statistical modeling to account for this potential confound; thus it is unlikely that these differences would impact our results. Future work should consider doing a controlled experiment to see if VR benefits certain types of backgrounds, such as complete novices or more expert users, more than others.

The displays used for this experiment are representative displays of satellite operations, but simplified and constrained. There is a wide variety of current displays, sharing similar principles, but also a wide variety of potential VR interface designs. While we made an effort to base our display decisions on literature and existing guidelines, different decisions made could have an impact on the outcomes. In the future, different types of VR and 3D interfaces could be explored to see how they can improve training. Finally, while this experiment is geared towards satellite operations, these results may be able to translate to other supervisory control paradigms such as air traffic control, maritime, other transportation, and manufacturing.

5. Conclusion

This study compares three training displays, including an immersive VR display, a 3D screen visualization, and a representative traditional display for satellite operations. SA, performance, workload, usability, and subjective utility were compared after training and during a human subject experiment where participants worked with a traditional display. The results of this work indicate that VR is a promising training mechanism for satellite operations as it improves level 2 SA and usability in traditional displays. Training in VR also has a higher perceived subjective utility towards understanding critical aspects of satellite operations. However, when VR training is compared with training in the 3D screen visualization, Scr. Viz, performance is harmed. No differences were found in the other SA levels and workload. VR training may be able to promote improved mental models, safer operations, and improved understanding of traditional displays. Future work should continue to assess VR potential for use in training, including the duration of VR training needed to improve mental models further, and if this can lead to level 3 SA benefits. Additional future work includes studying how to effectively incorporate VR into a training timeline for the most effective results, and understanding how display design choices play a role in the results. These results are important for satellite applications where using VR or visualizations may not be possible, whether due to long lead times in technology transitions, increased adoption costs compared to including 3D visualizations on existing computers, or difficulty incorporating the technology into classified environments. These conclusions can inform training for satellite operations, as well as other supervisory control scenarios like robotics, manufacturing, and transportation.

CRediT authorship contribution statement

Savannah L. Buchner: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jennifer Wood:** Writing – review & editing, Software. **Gabriella Schauss:** Writing – review & editing, Investigation. **Blake C. Bateman:** Writing – review & editing, Investigation. **Hanspeter Schaub:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition. **Allison P.A. Hayman:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Savannah Buchner reports that financial support was provided by SpaceWERX. Jennifer Wood reports that financial support was provided by SpaceWERX. Gabriella Schauss reports that financial support was provided by SpaceWERX. Blake Bateman reports that financial support was provided by SpaceWERX. Hanspeter Schaub reports that financial support was provided by SpaceWERX. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This material is based upon work supported by a SpaceWERX STTR under grant number AWD-23-04-0184. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Space Force.

Appendix A. SPAM queries

The following lists the possible SPAM queries asked to operators. Queries were randomly selected from this list; however, queries were only allowed to be selected at times when they were appropriate to the situation (e.g., “Will you receive new information before the next collision?” would only be asked if there was a collision potential.) Queries with a ‘/’ indicate that either option may be selected.

Level 1:

- Is the servicer’s battery greater/less than 50%?
- Is the servicer’s fuel greater/less than 50%?
- Is the servicer currently in the sunlight/shadow?
- Is the servicer’s flashlight on/off?
- Are there currently any cautions/warnings?
- Is the next planned burn more/less than 30min away?
- Is there more/less than 30min until the lighting conditions change?
- Is the servicer above/below the client satellite?
- Is the servicer ahead/behind of the client satellite?
- Is there a potential collision in the next 30min?
- Does the servicer’s current trajectory enter the keep out zone?

Level 2:

- Is the servicer’s battery level currently increasing/decreasing?
- Is the keep out zone currently increasing/decreasing in size?
- Is the servicer’s flashlight currently aimed towards the client?
- Is there currently enough battery/fuel/time to complete an abort burn?
- Does the orbit enter the keep out zone from ahead/behind of the client satellite?
- Does the orbit enter the keep out zone from above/below the client satellite?
- Is the time to collision increasing/decreasing?
- Is the portion of the orbit line in the keep out zone increasing/decreasing?

Level 3:

- If no new unplanned actions are taken will there be a collision alert in 30min?
- If no new unplanned actions are taken, will there be enough battery/fuel to complete an abort burn in 30min?
- If no burns occur, will the servicer enter the keep out zone in the next 30min?
- If there is no burn, will the servicer cross orbits with the debris satellite?
- If no new unplanned actions are taken, will the servicer’s battery be greater/less than 50% in 30min?
- If no action is taken, will you receive a new low battery/fuel caution/warning within the next 30min?
- If no new action is taken, will the keep out zone be growing/shrinking in 30min?
- If no new unplanned action is taken, will the servicer’s battery be increasing/decreasing in 30min?
- Will the servicer be in the sun/shadow at the early/default/late burn location?

Appendix B. Utility survey

Please read EACH of the following statements, and indicate the extent to which you agree with each. The system refers to the display that you worked with today.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I found the necessary information easy to find in this system	○	○	○	○	○
I found this system allowed me to be aware of mission critical events	○	○	○	○	○
I found this system enabled me to understand the uncertainties associated with the mission	○	○	○	○	○
I found this system promoted my understanding of collision likelihood	○	○	○	○	○
I found this system enabled me to understand the relative orbital motion of the satellites	○	○	○	○	○
I found this system allowed me to make appropriate operation decisions	○	○	○	○	○

Please provide any additional comments on this display and your experience performing the task.

Please read EACH of the following statements, and indicate the extent to which you agree with each. The training from the first visit refers to your previous visit.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I found that the training from the first visit was effective in enabling me to understand mission critical events today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling me to understand the uncertainties today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling my understanding of collision likelihood today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling me to understand the relative orbital motion of the satellites today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling me to make appropriate operation decisions today.	○	○	○	○	○

Please provide any additional comments about how your training on the first day impacted your performance today.

Did you prefer the display from visit 1 or 2? Why?

Appendix C. Performance

Table C.1
Burn decision.

Outcome	Score
No Selection	1
Poor	2
Fair or Good	3
Excellent	4

Table C.2
End state.

Outcome	Score
Collision – No Abort Attempt	1
Collision – Abort Attempt	2
Abort with >15 min to collision	3
Serviced with <5 min to collision or <25% battery	4
Abort – light use < 75%	5
Abort – light use > 75%	6
Serviced – Battery < 30%	7
Serviced – Battery 30–50%	8
Serviced – Battery > 50%	9

Total Performance = End State(battery, outcome) + Burn Decision.

Participants' performance score on each trial is the sum of two parts: burn decision and satellite end state. Burn decision relates to how effective the participants' burn selection enables overall mission success, and is described in Table C.1. During training, participants were instructed to always select a burn location, so no burn selection is an indication of either poor SA or poor understanding of appropriate actions. A 'Poor' burn decision consists of a burn location where it would be impossible to service the satellite. 'Fair' and 'Good' burn decisions are locations where servicing a satellite was possible, but different burn decisions would be more effective. 'Excellent' burn decisions were selections that put the participant in the best position to service the client. Not every scenario has all possible outcomes and for each trial, the location of a 'Poor', 'Fair', 'Good', or 'Excellent' may differ. The satellite end state refers to the final serviced state (collided, aborted, or serviced), and is in Table C.2. The granularity of the serviced state was added by evaluating participants' abort attempts, time to collision when aborted, use of flashlight, and end battery level. These different actions and metrics provide indirect information about the participants' understanding of the task, awareness of the environment, and optimization of the outcome. Like the burn decision, not every trial had the full spectrum of possible scores for the satellite end state. The granularity of the scores is intended to objectively evaluate participant actions. These performance subdivisions allow for the delineation of performance but can be generalized across scenarios. It should be noted that not all scenarios can achieve all of the performance values.

Data availability

Data will be made available on request.

References

Aïm, F., Lonjon, G., Hannouche, D., Nizard, R., 2016. Effectiveness of virtual reality training in orthopaedic surgery. Arthrosc.: J. Arthrosc. Relat. Surg. 32, 224–232. https://doi.org/10.1016/j.arthro.2015.07.023

Al-Hor, M., Almahdi, H., Al-Theyab, M., Mustafa, A.G., Seed Ahmed, M., Zaqout, S., 2024. Exploring student perceptions on virtual reality in anatomy education: insights on enjoyment, effectiveness, and preferences. BMC Med. Educ. 24, 1405. https://doi.org/10.1186/s12909-024-06370-6

Baker, A.L., Fitzhugh, S.M., Forster, D.E., Schaefer, K.E., 2021. Communication metrics for human-autonomy teaming: lessons learned from us army gunnery field experiments. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 65, SAGE Publications Inc, pp. 1157–1161, https://doi.org/10.1177/1071181321651159

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Bowden, V., Loft, S., 2013. Situation Awareness Measurement Techniques for Submarine Track Management. Technical Report FA23861214028., University of Western Australia.

Brooke, J., 1996. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation in Industry. CRC Press. pp. 6.

Buchner, S., Wood, J., Schauss, G., Bateman, B., Schaub, H., Hayman, A., 2025a. Benefits of virtual reality and 3D visualizations for remote satellite supervisory control. Virtual Real. 30, 8. https://doi.org/10.1007/s10055-025-01263-5

Buchner, S.L., Rindfuss, A., Wood, J., Schaub, H., Hayman, A.P., 2025b. Impacts of 3D visualizations and virtual reality in display designs for remote monitoring of satellite operations. Front. Virtual Real. 6, https://doi.org/10.3389/frvir.2025.1487281. Frontiers.

Caroux, L., Isbister, K., 2016. Influence of head-up displays' characteristics on user experience in video games. Int. J. Hum.-Comput. Stud. 87, 65–79. https://doi.org/10.1016/j.ijhcs.2015.11.001

Chen, J.Y.C., Haas, E.C., Barnes, M.J., 2007. Human performance issues and user interface design for teleoperated robots. IEEE Trans. Syst. Man Cybern. C (Appl. Rev). 37, 1231–1245. https://doi.org/10.1109/TSMCC.2007.905819

Chen, S.C., Hsiao, M.S., She, H.C., 2015. The effects of static versus dynamic 3D representations on 10th grade students' atomic orbital mental model construction: evidence from eye movement behaviors. Comput. Hum. Behav. 53, 169–180. https://doi.org/10.1016/j.chb.2015.07.003

Chen, Y., He, Y., Zou, X., Cai, H., Yiu, H.H.E., Ming, W.K., 2025. Undergraduate nursing students' preferences for virtual reality simulations in nursing skills training: a discrete choice experiment. Digit. Health 11, 20552076251339009. https://doi.org/10.1177/20552076251339009

Christense, R.H.B., 2023. Ordinal—regression models for ordinal data.

Cunningham, J.C., Battiste, H., Curtis, S., Hallett, E.C., Koltz, M., Brandt, S.L., Lachter, J., Battiste, V., Johnson, W.W., 2015. Measuring situation awareness with probe questions: reasons for not answering the probes. Procedia Manuf. 3, 2982–2989. https://doi.org/10.1016/j.promfg.2015.07.840

Daling, L.M., Schlittmeier, S.J., 2024. Effects of augmented reality-, virtual reality-, and mixed reality-based training on objective performance measures and subjective evaluations in manual assembly tasks: a scoping review. Hum. Factors 66, 589–626. https://doi.org/10.1177/00187208221105135. SAGE Publications Inc.

Dammen, L.V., Finseth, T.T., McCurdy, B.H., Barnett, N.P., Conrady, R.A., Leach, A.G., Deick, A.F., Van Steenis, A.L., Gardner, R., Smith, B.L., Kay, A., Shirtcliff, E.A., 2022. Evoking stress reactivity in virtual reality: a systematic review and meta-analysis. Neurosci. Biobehav. Rev. 138, 104709. https://doi.org/10.1016/j.neubiorev.2022.104709

Decoust, C., Magalhaes, M., Rolland, P., Pizolato, K., Farias, M., 2018. From submarine to satellite operations: a training success story, AIAA, https://doi.org/10.2514/6.2018-2357. https://arc.aiaa.org/doi/epdf/10.2514/6.2018-2357.

Dinno, A., 2014. dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums. Comprehensive R Archive Network Pages, pp. 1.3.6. https://doi.org/10.32614/CRAN.package.dunn.test

Dittemore, G., Bertels, C., 2011. The final count down: a review of three decades of flight controller training methods for space shuttle mission operations. In: AIAA SPACE 2011 Conference & Exposition. American Institute of Aeronautics and Astronautics, Long Beach, California, https://doi.org/10.2514/6.2011-7190

Durso, F., Truitt, T., Hackworth, C., Crutchfield, J., Ohrt, D., Nikolic, D., Moertl, P., Manning, C., 1995. Expertise and chess: a pilot study comparing situation awareness methodologies.

Durso, F.T., Truitt, T.R., Hackworth, C.A., Crutchfield, J.M., Manning, C.A., 1998. En route operational errors and situational awareness. Int. J. Aviat. Psychol. 8, 177–194. https://doi.org/10.1207/s15327108ijap0802_6

Durso, F.T., Hackworth, C.A., Truitt, T.R., 1999. Situation awareness as a predictor of performance in en route air traffic controllers. https://doi.org/10.1037/e402702004-001

Endsley, M.R., 1988. Design and evaluation for situation awareness enhancement. In: Proceedings of the Human Factors Society Annual Meeting, vol. 32, SAGE Publications, pp. 97–101, https://doi.org/10.1177/154193128803200221

Endsley, M.R., 1990. Predictive utility of an objective measure of situation awareness. In: Proceedings of the Human Factors Society Annual Meeting, vol. 34, SAGE Publications, pp. 41–45, https://doi.org/10.1177/154193129003400110

Endsley, M.R., 1996. Toward a theory of situation awareness in dynamic systems. Hum. Factors 37, 33.

Endsley, M., 2021a. Situation Awareness Measurement: How to Measure Situation Awareness in Individuals and Teams. Human Factors and Ergonomics Society.

Endsley, M.R., 2021b. A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. Hum. Factors 63, 124–150. https://doi.org/10.1177/0018720819875376. SAGE Publications Inc.

Federal Aviation Administration, 2004. ARAC WG Report Far/Jar 25.1322 & AC/ACJ 25.1322. Technical Report.

Fujino, M., Lee, J., Hirano, T., Saito, Y., Itoh, M., 2020. Comparison of SAGAT and SPAM for seeking effective way to evaluate situation awareness and workload during air traffic control task. In: Proceedings of the 2020 HFES 64th International Annual Meeting, Virtual, https://doi.org/10.1177/1071181320641442. https://journals.sagepub.com/doi/epdf/10.1177/1071181320641442.

Gad, R., Martin, S., Olbrich, M., Fischer, M., Baci, S., Ruecker, F., Sfaxi, R., 2023. Towards leveraging augmented and virtual reality for spacecraft mission operations at ESOC. In: 17th International Conference on Space Operations, Dubai, UAE.

Garcia, A.D., Schlueter, J., Paddock, E., 2020. Training astronauts using hardware-in-the-loop simulations and virtual reality. In: AIAA SciTech 2020 Forum. American Institute of Aeronautics and Astronautics, Orlando, FL. https://doi.org/10.2514/6.2020-0167, https://arc.aiaa.org/doi/10.2514/6.2020-0167.

Golding, J.F., 1998. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. Brain Res. Bull. 47, 507–516. https://doi.org/10.1016/s0361-9230(98)00091-4

Grier, R., 2015. How high is high? A metanalysis of NASA TLX global workload scores. https://doi.org/10.1177/1541931215591373

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Advances in Psychology. North Holland, vol. 52 Human Mental Workload, pp. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hawley, J.K., Mares, A.L., Giammanco, C.A., 2006. Training for Effective Human Supervisory Control of Air and Missile Defense Systems. https://doi.org/10.1037/e512092010-001

HERVE, M., 2011. RVAideMemoire: Testing and Plotting Procedures for Biostatistics. 10.32614/CRAN.package.RVAideMemoire. Comprehensive R Archive Network Pages, Institution, 0.9–83–7.

Hooey, B.L., Kaber, D.B., Adams, J.A., Fong, T.W., Gore, B.F., 2018. The underpinnings of workload in unmanned vehicle systems. IEEE Trans. Hum.-Mach. Syst. 48, 452–467. https://doi.org/10.1109/THMS.2017.2759758

Huey, B.M., Wickens, C.D., 1993. Workload Transition: Implications for Individual and Team Performance. Technical Report., National Research Council.

Iacovides, I., Cox, A., Kennedy, R., Cairns, P., Jennett, C., 2015. Removing the HUD: the impact of non-diegetic game elements and expertise on player involvement. In: Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play, Association for Computing Machinery, New York, NY, USA, pp. 13–22, https://doi.org/10.1145/2793107.2793120

Jones, D.G., Endsley, M.R., 1996. Sources of situation awareness errors in aviation. Aviat. Space Environ. Med. 67, 507–512.

Kaplan, A.D., Cruit, J., Endsley, M., Beers, S.M., Sawyer, B.D., Hancock, P.A., 2021. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: a meta-analysis. In: Human Factors, vol. 63. SAGE Publications Inc, pp. 706–726. https://doi.org/10.1177/0018720820904229

Kenneally, P.W., Piggott, S., Schaub, H., 2020. Basilisk: a flexible, scalable and modular astrodynamics simulation framework. J. Aerosp. Inf. Syst. 17, 496–507. https://doi.org/10.2514/1.I010762. American Institute of Aeronautics and Astronautics.

Kolbeck, S., Amodio, C., Burkhardt, V., 2018. Job analysis and collaborative training for spacecraft operators in future control rooms. In: AIAA, https://doi.org/10.2514/6.2018-2472

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Softw. 82, 1–26. https://doi.org/10.18637/jss.v082.i13

Lager, M., Topp, E.A., 2019. Remote supervision of an autonomous surface vehicle using virtual reality. IFAC-Pap. 52, 387–392. https://doi.org/10.1016/j.ifacol.2019.08.104

Laskey, L., Keebler, J.R., Cairns, P.M., Lopez, G., 2024. Evaluating the effectiveness of game-based virtual reality in satellite ground control operations education and training. Int. J. Aviat. Aeronaut. Aerosp. 11, https://doi.org/10.58940/2374-6793.1841

Lejuez, C.W., Read, J.P., Kahler, C.W., Richards, J.B., Ramsey, S.E., Stuart, G.L., Strong, D.R., Brown, R.A., 2002. Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (BART). J. Exp. Psychol. Appl. 8, 75–84. https://doi.org/10.1037//1076-898x.8.2.75

Lenth, R.V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., Singmann, H., 2023. emmeans: Estimated Marginal Means, aka Least-Squares Means.

Li, M., Sun, Z., Jiang, Z., Tan, Z., Chen, J., 2020. A virtual reality platform for safety training in coal mines with AI and cloud computing. In: Discrete Dynamics in Nature and Society 2020, e6243085. Hindawi, https://doi.org/10.1155/2020/6243085

Liebowitz, J., Lightfoot, P., 1987. Training NASA satellite operators: an expert system consultant approach. Educ. Technol. 27, 41–47. Educational Technology Publications, Inc.

Loft, S., Morrell, D., Huf, S., 2013. Using the situation present assessment method to measure situation awareness in simulated submarine track management. Int. J. Hum. Factors Ergon. 2, 33–48. https://doi.org/10.1504/IJHFE.2013.055975

Loft, S., Bowden, V., Braithwaite, J., Morrell, D.B., Huf, S., Durso, F.T., 2015. Situation awareness measures for simulated submarine track management. Hum. Factors 57, 298–310. https://doi.org/10.1177/0018720814545515. SAGE Publications Inc.

Makransky, G., Lilleholt, L., 2018. A structural equation modeling investigation of the emotional value of immersive virtual reality in education. Educ. Technol. Res. Dev. 66, 1141–1164. https://doi.org/10.1007/s11423-018-9581-2. Germany Springer.

Marre, Q., Caroux, L., Sakdavong, J.C., 2021. Video game interfaces and diegesis: the impact on experts and novices' performance and experience in virtual reality. Int. J. Hum.-Comput. Interact. 37, 1089–1103. https://doi.org/10.1080/10447318.2020.1870819. Taylor & Francis _eprint.

Mayer, R.E., Makransky, G., Parong, J., 2023. The promise and pitfalls of learning in immersive virtual reality. Int. J. Hum.-Comput. Interact. 39, 2229–2238. https://doi.org/10.1080/10447318.2022.2108563. Taylor & Francis.

Meng, T., Mao, R., Sun, S., 2024. Exploration and reflections on empowering aerospace control education through digitalization. IFAC-Pap. 58, 53–58. https://doi.org/10.1016/j.ifacol.2024.08.461

Miller, L., Stanney, K., Guckenberger, D., Guckenberger, E., 1997. Real-time training. Ergon. Des. 5, 21–24. https://doi.org/10.1177/106480469700500306

Mirchi, T., Vu, K.P., Miles, J., Sturre, L., Curtis, S., Strybel, T.Z., 2015. Air traffic controller trust in automation in NextGen. Procedia Manuf. 3, 2482–2488. https://doi.org/10.1016/j.promfg.2015.07.509

Morris, N.M., Rouse, W.B., 1985. The effects of type of knowledge upon human problem solving in a process control task. IEEE Trans. Syst. Man Cybern. SMC-15 698–707. https://doi.org/10.1109/TSMC.1985.6313453

O'Hara, J., Fleger, S., 2020. Human-System Interface Design Review Guidelines. Technical Report BNL–216211-2020-FORE, 1644018, https://doi.org/10.2172/1644018

Reed, S.C., Levin, F.R., Evans, S.M., 2012. Alcohol increases impulsivity and abuse liability in heavy drinking women. In: Experimental and Clinical Psychopharmacology, vol. 20. American Psychological Association, US, pp. 454–465. https://doi.org/10.1037/a0029087

Reiter, J.M., 2012. Multi-mission operator training practices. In: SpaceOps 2012 Conference. American Institute of Aeronautics and Astronautics, Stockholm, Sweden, https://doi.org/10.2514/6.2012-1275361

Rosyid, H.A., Pangestu, A.Y., Akbar, M.I., 2021. Can diegetic user interface improves immersion in role-playing games? In: 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), pp. 200–204, https://doi.org/10.1109/ICEEIE52663.2021.9616732. https://ieeexplore.ieee.org/abstract/document/9616732.

Rouse, W.B., Morris, N.M., 1985. On Looking Into the Black Box: Prospects and Limits in the Search for Mental Models. Technical Report AD-A 159 080.

Selkowitz, A.R., Lakhmani, S.G., Chen, J.Y.C., 2017. Using agent transparency to support situation awareness of the autonomous squad member. Cogn. Syst. Res. 46, 13–25. https://doi.org/10.1016/j.cogsys.2017.02.003

Sellmaier, F., Frei, H., 2022. Operations of on-Orbit Servicing Missions. Spacecraft Operations. Springer International Publishing, Cham, pp. 491–529. https://doi.org/10.1007/978-3-030-88593-9-25

Seymour, N.E., Gallagher, A.G., Roman, S.A., O'Brien, M.K., Bansal, V.K., Andersen, D.K., Satava, R.M., 2002. Virtual reality training improves operating room performance. Ann. Surg. 236, 458–464.

Sheridan, T.B., 2012. Human supervisory control of automation. In: Handbook of Human Factors and Ergonomics. John Wiley & Sons, Incorporated, New York, United States.

Sittner, F., Hartmann, O., Montenegro, S., Friese, J.P., Brubach, L., Latoschik, M.E., Wienrich, C., 2023. An update on the virtual mission control room. In: 37th Annual Small Satellite Conference, Logan Utah. SSC23-P5-25.

Solberg, E., Nystad, E., McDonald, R., 2023. Situation awareness in outage work – a study of events occurring in U.S. nuclear power plants between 2016 and 2020. Saf. Sci. 158, 105965. https://doi.org/10.1016/j.ssci.2022.105965

Stathopoulos, F., Oezdemir, K., 2018. Human factor and knowledge management in 24/7 multi-mission satellite operations. In: 2018 SpaceOps Conference. American Institute of Aeronautics and Astronautics, Marseille, France, https://doi.org/10.2514/6.2018-2474

Thomas, L.C., Wickens, C.D., 2001. Visual displays and cognitive tunneling: frames of reference effects on spatial judgments and change detection. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 45, SAGE Publications Inc, pp. 336–340, https://doi.org/10.1177/154193120104500415

Thompson, M., Uz-Bilgin, C., Tutwiler, M.S., Anteneh, M., Meija, J.C., Wang, A., Tan, P., Eberhardt, R., Roy, D., Perry, J., Klopfer, E., 2021. Immersion positively affects learning in virtual reality games compared to equally interactive 2D games. Inf. Learn. Sci. 122, 442–463. https://doi.org/10.1108/ILS-12-2020-0252. Emerald Publishing Limited.

Wheeler, S.G., Engelbrecht, H., Hoermann, S., 2021. Human factors research in immersive virtual reality firefighter training: a systematic review. Front. Virtual Real. 2.

Wickens, C.D., 1993. Cognitive factors in display design. J. Wash. Acad. Sci. 83, 179–201. Washington Academy of Sciences.

Wood, J., Margenet, M.C., Kenneally, P., Schaub, H., Piggott, S., 2018. Flexible basilisk astrodynamics visualization software using the unity rendering engine. In: AAS Guidance and Control Conference, Breckenridge, CO.

Worm, G.I.M., van der Wees, M., de Winter, J.C.F., de Graaf, L., Wieringa, P.A., Rietveld, L.C., 2012. Training and assessment with a faster than real-time simulation of a drinking water treatment plant. Simul. Model. Pract. Theory 21, 52–64. https://doi.org/10.1016/j.simpat.2011.09.007

Xie, B., Liu, H., Alghofaili, R., Zhang, Y., Jiang, Y., Lobo, F.D., Li, C., Li, W., Huang, H., Akdere, M., Mousas, C., Yu, L.F., 2021. A review on virtual reality skill training applications. Front. Virtual Real. 2.