Comparison of Visualization Technologies to Support RCAF Training Modernization

Major Jason Munn, Captain Dan Deluce
RCAF Aerospace Warfare Centre
Ottawa, Canada
Jason.Munn@forces.gc.ca,
Daniel.Deluce@forces.gc.ca

Dr. Jerzy Jarmasz
Defence R&D Canada
Toronto, Canada
Jerzy.Jarmasz@drdc-rddc.gc.ca

ABSTRACT

The Royal Canadian Air Force (RCAF) is undertaking a major training system modernization. This transformation looks to leverage emerging training technologies to support more flexible, individualized and streamlined training for its personnel. In order to support training modernization in the RCAF, the RCAF Aerospace Warfare Centre (AWC) seeks to develop evidence-based procedures for the adoption, application and maintenance of the RCAF's Modelling & Simulation (M&S) capabilities in collaboration with Defence Research and Development Canada (DRDC) and their research program on Training for the Future Operational Environment (TFOE). DRDC-led investigations have identified a number of organizational and human factors that impede effective training technology use. One of those factors is a general lack of evidence on the effectiveness and usability of emerging training technologies, consistent with the experiences of the RCAF AWC in this area. This study examines the suitability of different visualization technologies in the context of distributed Collective Training (CT) by comparing the effects on performance (track integrity), workload, and usability of three different visualization configurations (conventional monitor, curved wide-angle monitor, and mixed-reality headset) in simulated air patrol tasks by experienced fast jet pilots. The findings of this study and their implications on visualization technology use are discussed in the context of RCAF AWC efforts to modernize its distributed collective training capability and DRDC activities on supporting training technology adoption in the Canadian Armed Forces (CAF).

ABOUT THE AUTHORS

Major Jason Munn is a M&S Support Officer at the Royal Canadian Air Force Aerospace Warfare Centre with a long background in Maritime Helicopter aviation piloting and commanding missions in the venerable H-3 Sea King helicopter. He holds an undergraduate degree in Aerospace Engineering and interests include Extended Reality (xR) and advancing the use of M&S in the RCAF.

Captain Daniel Deluce is a M&S Support Officer with the Royal Canadian Air Force Aerospace Warfare Centre. He is a qualified CF-18 pilot with experience in operational and training roles. He studied Chemical Engineering and was a commercial pilot before joining the military. His interests include autonomous systems, artificial intelligence, and human machine teaming.

Dr. Jerzy Jarmasz is a Defence Scientist at Defence Research and Development Canada (DRDC). With backgrounds in both electrical engineering and cognitive science, he brings a multidisciplinary approach to all of his projects. His areas of expertise include simulation-based training, training for cognitive skills (e.g., situation awareness, decision-making) and team effectiveness training at the tactical, operational and strategic levels.

Comparison of Visualization Technologies to Support RCAF Training Modernization

Major Jason Munn, Captain Dan Deluce
RCAF Aerospace Warfare Centre
Ottawa, Canada
Jason.Munn@forces.gc.ca,
Daniel.Deluce@forces.gc.ca

Dr. Jerzy Jarmasz
Defence R&D Canada
Toronto, Canada
Jerzy.Jarmasz@drdc-rddc.gc.ca

INTRODUCTION

The Canadian Armed Forces (CAF) have initiated a major Reconstitution effort to strengthen capabilities needed to ensure operational relevance in the current and future security environment (Government of Canada, 2022). The future operating environment is becoming increasingly complex, and therefore training to become a war fighter is also more complex. The use of emerging technologies allows the training to be more immersive and realistic, to tailor training to individual learners to enhance training transfer, and also provides more opportunities to conduct collective mission training and rehearsal. The Royal Canadian Air Force (RCAF) is undertaking a major training system modernization to meet the challenges of preparing for the future operating environment. To support the RCAF training modernization, the RCAF Aerospace Warfare Centre (AWC) seeks to develop evidence-based procedures for the adoption, application and maintenance of the RCAF's Modeling & Simulation (M&S) capabilities. Despite a wealth of data and evidence available, there are few sources for guiding the systematic application and employment of training technologies based on effectiveness evidence. Obtaining empirical evidence for the effectiveness of training technologies is challenging and therefore this step is often ignored. The RCAF aspires to evaluate emerging training technologies, and their suitability for use, in a more thoughtful way. Developing this capability will enable the RCAF to modernize its training system by selecting the right technology and using it in ways that deliver the maximum effects.

Defense Research Development Canada (DRDC) supports the RCAF in assembling a body of knowledge to systematically guide the application of training technologies. This effort involves identifying or obtaining evidence for the effectiveness of technology selection and adoption. Empirical studies are often required to obtain the evidence required for a specific technology. The study presented here is the first step in a nascent collaboration between the RCAF and DRDC on accruing data on the effectiveness of various training technologies. The specific use case under study is a comparison of different visual display configurations for a fighter simulator in support of aerospace controller training. The medium-fidelity simulator is used by qualified fighter pilots to provide simulated air effects in a distributed collective training scenario to train aerospace controllers. Currently, the simulator uses a standard flat screen display. The RCAF is interested in determining whether newer display technologies such as Virtual Reality (VR) and Mixed Reality (MR) can improve the quality of the inputs provided to the Primary Training Audience (PTA) by increasing fidelity/realism, improving usability, or decreasing the workload (task complexity) for the fighter pilot in the simulator. This is in part motivated by a general enthusiasm for MR-based training solutions throughout the aerospace training communities in recent years. Despite the steep increase in interest, MR has apparent usability limitations that relate to visual issues and possible simulator sickness, which are discussed below. Thus, it is of critical interest to the RCAF to determine whether current VR/MR solutions are warranted for consideration in a particular use case.

Review of Literature on VR/MR for Military Training

The use of simulation is a long-established practice in aviation both civilian and military (e.g., see Ali, Guckenberger & Rossi, 2000, and Jentsch, Curtis & Salas, 2011). A specific simulation technology that has been drawing more and more attention in recent years is a family of related technologies often referred to as Extended Reality (xR; Kaplan et al., 2021), which includes Augmented Reality (AR) as well as VR and MR mentioned above. In a recent meta-analysis, Kaplan et al. (2021) found evidence for the different xR technologies being effective as training interventions in a variety of task types and domains. Despite their efforts at a comprehensive overview, the authors were not able to

include any military or aviation studies in their analysis, even while explicitly noting the established value of simulation in general for military aviation.

Furthermore, despite many putative advantages, the use of xR technologies for training also has several potential drawbacks. xR technologies are known to carry the risk of simulator sickness, or cyber sickness (a form of simulator sickness specific to xR headsets; Arcioni et al., 2018; Bos et al., 2021), and may impose additional workload or usability challenges on users relative to more established technologies. Recent studies on the workload burden of xR (Lackey et al., 2016; van Weelden et al., 2022; Xi et al., 2022) have found conflicting or inconclusive results relative to other visualization modalities. One implication of these findings is that user performance, workload and cybersickness are not consistently assessed together in many xR studies. Another implication is that the strengths and weaknesses of xR technologies must be assessed with respect to specific training applications. Specifically, more studies are required with military users.

Two recent studies on xR for military aviation training, notable in part because of their rarity, are McCoy-Fisher et al. (2019) and Severe-Valsaint et al. (2022). These studies assessed the use of VR and MR technologies in US Navy pilot training. McCoy-Fisher et al. examined user acceptance, skill acquisition and simulator sickness with VR/MR-based part task trainers for a variety of US Navy airframes. The study found a degree of positive user acceptance and skill acquisition for certain tasks, together with a degree of simulator sickness that was considered minor or acceptable. Severe-Valsaint et al. focused on the instructional methodology behind the application of VR/MR technologies, finding that the manner in which these technologies are used is crucial to their training effectiveness. However, this last study did not assess workload, user discomfort or other usability issues.

Thus, while new findings on the effectiveness of xR technologies are emerging, the knowledge base is incomplete with respect to the RCAF AWC specific requirements considered here, namely the relative merits of VR/MR compared to traditional (flat and curved screen) displays for an aviation training support (role player) task in a collective military training context.

Overview of Experimental Task and Research Objectives

The simulation tool suite employed by the RCAF Distributed Mission Operations Centre (DMOC) provides friendly (blue), opposing (red) and other (white) entities, including land, sea, air and space units, whose behaviours can be automated or controlled by a Human-In-The-Loop (HITL) operator (the fighter pilot). The specific use case for this study utilizes a fast air configuration for an Offensive Counter Air/Defensive Counter Air (OCA/DCA); 2 versus 2 Beyond Visual Range (BVR) engagement mission set, requiring the fast-air HITL to affect the intercept and engagement of adversaries. To provide appropriate fidelity training to the PTA, the Air Battle Manager (ABM) in this use case, the HITL must be able to emulate the general procedural, physical and technical behaviours and limitations of a typical fighter aircraft currently employed by the RCAF, currently the CF-18. The visual display of the fighter simulator configuration is a key technology to enable the appropriate level of emulation fidelity of those behaviours. The subjective and objective data collected for this analysis will guide the decision for the most appropriate visualization setup, given constraints further detailed below, when configured to provide fast air, OCA/DCA mission sets, for distributed mission training or operations.

Many different types of visual displays and systems have been developed for HITL fighter simulations. Considering constraints including the specific use case, cost, infrastructure and support, not all of these visualization technologies can be considered for the RCAF DMOC. Due to the above constraints, while considering the PTA and a need to limit the physical area required (footprint), HITL systems such as dome projections, or detailed cockpit panels, are not practical, nor necessary. The visualization options that meet the constraints for use at the RCAF DMOC that were examined in this experiment are:

- An ultra-wide curved monitor providing a wide horizontal field of view, but restricted field of regard.
- A single conventional monitor with less horizontal field of view, and similarly restricted field of regard.
- A head mounted MR display providing a field of view similar to the conventional monitor but full field of regard. The MR visualization configuration includes high-fidelity selective pass-through aspects allowing the user to see real world components blended with the virtual.

There exists no specific guidance on when and where to apply specific visualization technologies for the given use case. This study should answer the question of what configuration (or combination of configurations) is best suited for HITL interaction with a simulation system in order to produce the desired training effect for the specific use case of the OCA/DCA mission set for DST; specifically, a 2v2 BVR engagement scenario.

METHOD

Participants and Demographics

A total of 10 participants volunteered for the study. All of the participants were male, ages ranging from 30 to 59 (4 participants reported an age in the 50-54 range, 2 in the 40-44 range, 2 in the 35-39 range, and one each in the 55-59 and 30-34 ranges). As per the inclusion criteria, all were current or former RCAF fast jet pilots, with a mean time since their last qualification as a CF-18 pilot of 5.5 years (SD = 8.5 years). All were qualified as Lead, with an average of 2806 flight hours. The participants had either no or very little previous experience with VR headsets (5 participants reported never using them, and the remaining participants reported using them a few times a month or less). They generally did not make extensive use of simulation recreationally (e.g., games or home flight simulators), with one participant reporting using simulators recreationally a few times a year, and all other participants reporting doing so either never or a few times in the past. However, most used simulators in a professional setting frequently (either daily or weekly, with only 4 participants reporting doing so a few times a month or less). Given the relatively low number of participants, and their relative uniformity, further analyses of demographic data were not performed.

Equipment and Experimental Stimuli

The simulator setup consisted of a commercial gaming chair with a variety of monitors, flight controls and other peripherals. Participants were required to ingress, operate, and egress the simulator. The main variable examined was the configuration of the Heads-Up Display (HUD) presenting the external view. The HUD hardware was configured in three different ways: a curved screen configuration (Figure 1), a software restricted configuration which presents a traditional monitor aspect and size (Figure 2), and completely presented in MR configuration (Figure 3). The MR configuration retained the information presented to the other views through human eye resolution dynamic video occlusions.

The technical details of each visualization configuration (V1, V2 and V3) are:

- V1: HUD is a 49” curved display; Resolution: 5120x1440; Refresh Rate: 59.977 Hz.
- V2: HUD is a 27” conventional display (using the same monitor as V1 configuration but the display area is limited by adjusting display software settings; Resolution: 1920x1080; Refresh Rate: 59.977 Hz.
- V3: The external view represented by the HUD is a Varjo XR-3 Head Mounted Display; Focus area: 27° x 27° at 70 Pixels Per Degree (PPD) uOLED; 1920x1920 pixels per eye; Peripheral area: over 30 PPD LCD; Field of view: Horizontal 115°; Weight: 980 g.

The equipment used by participants included a full set of representative primary flight controls (replica F-18 flight stick grip, a generic Thrustmaster Warthog HOTAS throttle quadrant, and generic rudder pedals). The displays and hardware used in the experiment are depicted in Figures 1, 2 and 3 below.

Procedure

Participants were provided detailed information about the experiment, asked to sign a Voluntary Consent Form prior to the experiment, and allowed to ask any questions they had about the procedure before the study. Participants then received a system familiarization hands-on demo, including a free-flight session using each of the different visualization configurations. During this session, they took control of a blue air entity at altitude and became familiar with basic manipulation of the tactical display (TACPLOT).

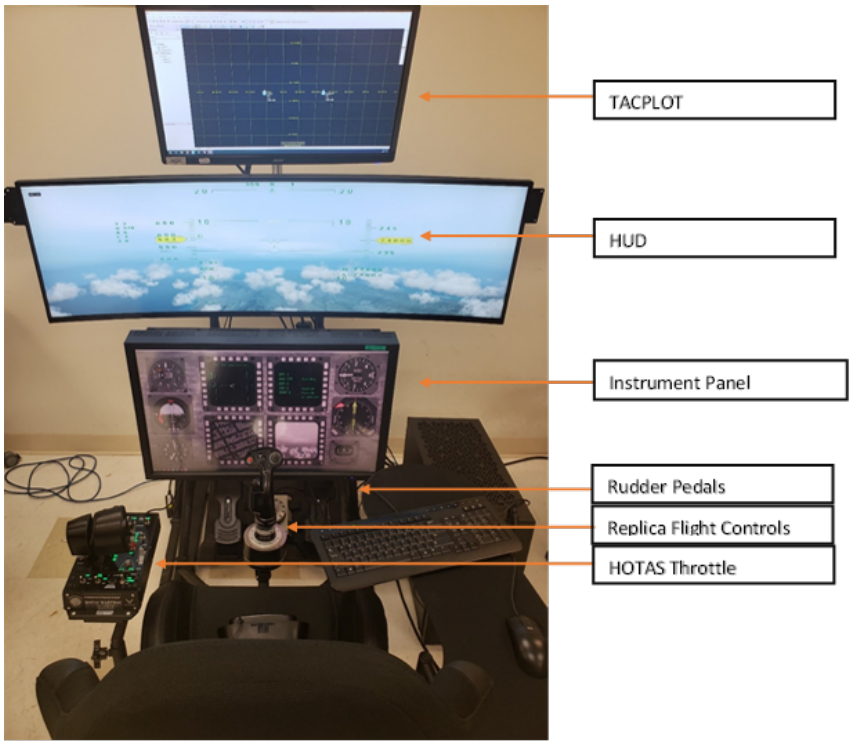

Figure 1: V1 configuration layout (curved display)

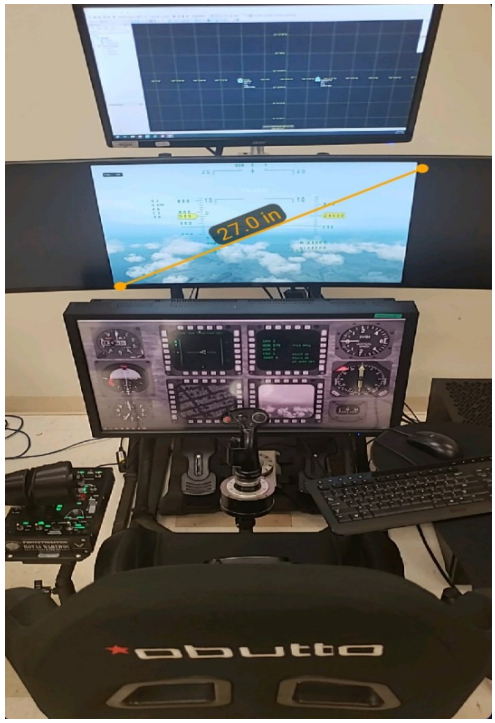

Figure 2: V2 configuration layout (conventional)

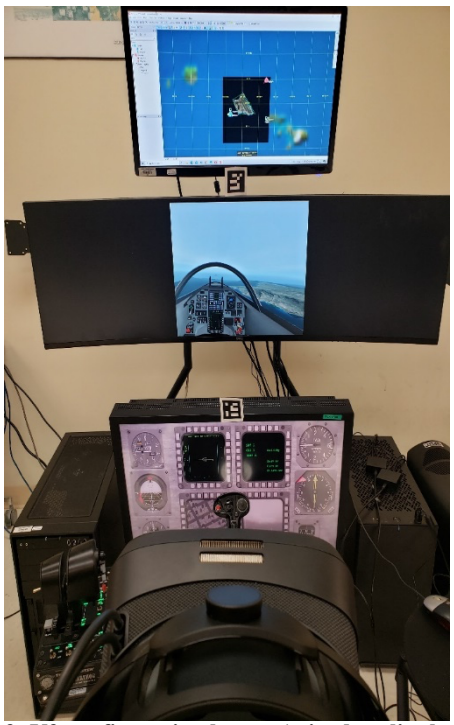

Figure 3: V3 configuration layout (mixed reality headset)

Participants completed 3 different mission scenarios repeating them once in each visual display configuration (V1, V2 and V3), for a total of 9 trials per participant. To control for the possibility of sequence effects with the visualization configurations, the order of presentation was randomized and counter-balanced between participants. At the start of each scenario, participants received a scenario description and instructions on what they had to accomplish. Specifically, the participants were instructed to maintain 2 nautical mile (NM) line abreast formation integrity with a computer generated lead aircraft that had scripted behaviours. As would be available in an actual HITL simulation with PTA, range to the Lead entity was available in 1 NM increments to the participants via a range line hooked to Lead and HITL entities and displayed on the TACPLOT. Each scenario started mid-flight, and did not include take-off or landing phases, as they were not relevant to the study. All 3 scenarios represented simple OCA/DCA BVR engagement profiles as could be expected in a typical simulation. In Scenarios 1 and 3, the Lead aircraft changed heading according to a pre-programmed route and the participant had to maneuver to maintain formation with the Lead, whereas in Scenario 2 the Lead aircraft essentially flew in a straight line, requiring the participant to fly in parallel with the Lead throughout. Although the experiment did not evaluate the participant's proficiency with avionics/weapon systems manipulation during 2 out of the 3 scenarios, participants were asked to manipulate these systems (selecting a target from the radar using appropriate HOTAS switches and firing an air-to-air missile) to measure whether the visualization configuration affects their cognitive workload.

Measures

At the end of each block of scenarios within a given visualization configuration, investigators administered three questionnaires. The first questionnaire was the Simulator Sickness Questionnaire (SSQ; Kennedy et al., 1993) to compare the degree of simulator sickness induced by the different configurations (participants reporting excessive simulator sickness were also able to discontinue their participation if they felt unable to continue). Following the SSQ, the NASA Task Load Index (TLX) workload questionnaire (Hart, 2006) was administered to evaluate the participant's cognitive workload, and the System Usability Scale (SUS; Brooke, 1996) to evaluate the usability of that particular configuration. Since the participants were also subject matter experts in the task, informal debriefs were conducted to gather any additional comments not recorded in the questionnaires that may be useful in evaluating the visual configuration.

RESULTS

Flight Performance (Formation Integrity)

The primary performance measure analyzed was the separation between the participant's aircraft and the Lead aircraft. The average separation in meters between the Lead and the Wingman (participant) trajectories was computed over the duration of each scenario and visualization condition. As noted above, participants were instructed to maintain a 2 NM (3,704 meters) separation from the Lead aircraft. With the basic flying conducted by participants, a deviation of plus or minus 0.5 NM was used as an acceptable tolerance for the line abreast formation. Anything outside of these parameters the aircraft would be considered being out of formation. Thus, the separation data was analyzed to determine whether participants maintained separation within an envelope of 1.5 to 2.5 NM (2778 m to 4630 m). Deviations above 2.5 NM were considered undesirable in each scenario. However, in Scenarios 1 and 3, where the Lead aircraft made frequent changes in heading, requiring participants to momentarily maneuver towards the Lead to re-establish formation, deviations below 1.5 NM were expected and considered appropriate (note also that in Scenario 2, participants typically did not fly within 1.5 NM of the Lead aircraft). Thus, in addition to average separation, the following metrics were computed and analyzed: proportion (%) of a scenario where separation exceeded 2.5 NM (all scenarios and visualizations), proportion (%) of a scenario where separation was below 1.5 NM (all visualizations, Scenarios 1 and 3 only), and average separation within the phases where separation was below 1.5 NM (all visualizations, Scenarios 1 and 3 only). Each of these metrics was subjected to 2-level (scenario × visualization) repeated-measures Analysis of Variance (ANOVAs).

Analysis of the average separations showed main effect of Scenario ($F(2,18) = 28.4$, $p < .05$) only, with Scenario 2 having the highest average separation (4,394.16 m). While the average separation was consistently higher for visualization condition V3 (MR) than the other visualizations in each scenario, this difference was not statistically significant, and there was no interaction between factors.

Analysis of the percentage of each scenario where separation exceeded 2.5 NM also showed a main effect of Scenario only ($F(2,18) = 8.38$, $p < .05$; Scenario 1 = 0.9%, Scenario 2 = 18.7%; Scenario 3 = 7.7%). No clear pattern emerged across visualization conditions. Analysis of the portions of the scenario where participants maneuvered closer than 1.5 NM to regain formation showed a main effect of scenario only, with a higher percentage of Scenario 1 (39.4%) having a separation below 1.5 NM than Scenario 3 (25.0%; $F(1,9) = 19.5$, $p < .05$), and no effect of visualization condition or interactions between scenario and visualization. However, analysis of the average separation between aircraft during the sub-1.5 NM portions of the scenarios showed a main effect of Visualization ($F(2,18) = 15.1$, $p < .05$), with the smallest separations occurring for visualization condition V2 (2180 m) compared to V1 (2339 m) and V3 (2283 m), and no main effect of scenario or interaction between scenario and visualization conditions.

Simulator Sickness Questionnaire (SSQ) Ratings

SSQ scores, computed from the SSQ ratings gathered from participants were compared between visualization conditions (as only one overall SSQ rating was collected in each visualization condition, covering participants experiences with all 3 scenarios, analyses of SSQ ratings per scenario condition were not undertaken). Using non-parametric analysis (Friedman ANOVA), which are generally recommended for SSQ data (Merchant & Kirollos, 2022), a significant difference in SSQ ratings was found between the 3 visualization conditions ($\chi^2 = 9.3$, $p < .05$), with the V3 (MR) condition having the highest mean ratings (M=16.83), followed by the V2 (traditional flat screen) condition (M=4.11) and the V1 (wide-screen) condition (M=1.87). The SSQ ratings for the V3 condition are considered to be within the "concerning" range (15-20), whereas the ratings for conditions V1 and V2 are considered to be negligible (Kennedy et al., 2003).

As only 2 participants reported SSQ ratings in condition V1, and only 2 in condition V2, it is clear that the V3 condition is the only one with meaningful SSQ scores; accordingly, the V1 and V2 conditions were not analyzed any further with respect to SSQ. SSQ in V3 was further analyzed to determine whether order of presentation of the MR condition affected SSQ ratings. As noted above, the order of visualization conditions was varied across participants to minimize order effects: 3 of the participants were presented with the MR condition first (average SSQ rating = 13.7), 3 experienced the MR condition second (average rating = 21.2), and 4 experienced it last (average rating = 15.9). A Kruskal-Wallis ANOVA was performed to compare SSQ ratings in the V3 condition across presentation orders, and no significant difference was found ($H(2) = .52$, $P = .77$).

Workload (NASA-TLX) Ratings

NASA TLX ratings were compared between visualization conditions (as with the SSQ, only an overall NASA TLX rating was obtained for participants across the scenarios in each visualization condition). Both raw scores (ranging from 1 to 20) for each of the 6 TLX sub-scales, as well as a total workload score (using raw scores but reverse-coding the successful performance scale and normalizing the average across subscales to produce a total score from 0 to 100) were compared across the 3 visualization conditions using a repeated-measures ANOVA. Significant differences were found for the Mental Demands sub-scale ($F(2,18) = 11.7$, $p < .05$) and Physical Demands sub-scale ($F(2,18) = 3.6$, $p < .05$) only; in both cases, workload was rated highest for the V3 (MR) condition. Mean ratings for the individual sub-scales (out of 20) and the overall TLX scores (out of 100) per visualization condition are given in Table 1.

Table 1: NASA TLX scores. Bolded subscales marked with an asterisk (*) denote a significant difference between means within the subscale.

TLX sub-scale	Visualization condition (Mean, Standard Deviation)		
	V1	V2	V3
Mental Demand*	**5.2 (1.3)**	**4.9 (1.4)**	**7.2 (1.3)**
Physical Demand*	**2.5 (1.0)**	**2.0 (0.5)**	**4.3 (1.2)**
Temporal Demand	4.4 (1.3)	4.1 (1.3)	4.2 (1.4)
Performance	10.2 (1.7)	10.2 (1.8)	9.7 (1.7)
Effort	7.0 (1.8)	6.5 (1.7)	7.9 (1.7)
Frustration	3.8 (1.3)	3.2 (1.0)	3.3 (0.9)
Total TLX	22.7 (5.5)	20.3 (5.0)	25.4 (5.3)

System Usability Scale (SUS) Ratings

Overall SUS ratings were computed for participants for each visualization condition, and subjected to comparison using repeated measures ANOVA. A significant difference was found between visualizations ($F(2,18) = 3.9$, $p < .05$), with the V3 (MR) visualization having the lowest SUS average rating (60.8, SD = 3.8), and the curved (M = 67.0, SD = 4.1) and conventional (M = 68.0, SD = 2.8) having similarly higher ratings.

Correlations

Pearson correlations were computed between total SSQ scores, each NASA TLX subscale, total NASA TLX score, and SUS ratings within each visualization condition. While various NASA TLX subscales correlated with each other in each condition, the correlations of interest here are those between SSQ and NASA TLX (subscales and total), SSQ and SUS, and NASA TLX and SUS.

SSQ and NASA TLX
Overall, SSQ and workload scores were not related in general. Correlations between SSQ scores and NASA TLX scores were found only for condition V3 (MR), where SSQ significantly correlated with the Performance Scale ($r = .72$, $p < .05$).

SSQ and SUS
SSQ and SUS ratings in each visualization condition were not significantly correlated with each other (all correlations $p > .05$)

SUS and NASA TLX
SUS scores significantly correlated ($p < .05$) with the Frustration and Total NASA TLX scales for visualization conditions V1 (curved screen) and V2 (flat screen) only (no significant correlations for the other subscales, or the MR condition). All significant correlations were negative indicating that usability ratings decreased as Frustration and Total TLX scores increased. The correlations between SUS and NASA TLX ratings are given in Table 2.

Table 2: Correlations between SUS and NASA TLX ratings. Asterisks (*) denote significant correlations (p < .05).

Visualization Condition	NASA TLX scales						
	Mental Demand	Physical Demand	Time Demand	Performance	Effort	Frustration	Total
V1	-.46	-.36	-.54	-.05	-.39	-.91*	-.74*
V2	-.47	-.27	-.56	-.19	-.38	-.76*	-.68*
V3	-.21	-.24	-.19	-.44	-.12	-.27	-.42

DISCUSSION

Discussion of Findings

The three visualizations used for this study did not produce significantly differing performance across the 3 Scenarios, except for the V2 (traditional flat screen) resulting in less separation between participant and Lead aircraft during formation re-establishment maneuvers. However, with respect to the other measures, the V3 (MR) condition was associated with notably higher SSQ ratings, higher workload in two TASA TLX subs-scales (Mental and Physical Demands), and lower SUS (usability) ratings. Thus, for the task used in the present study, no meaningful performance differences were measured between the MR condition and the conventional flat or curved screen conditions. However, more discomfort, higher workload and lower usability were reported in the MR condition, suggesting that it may not be the optimal choice for this use case.

These findings should not be taken to indicate that MR (or xR technologies) in general is not suited to flight tasks or training; indeed, as noted above, a number of studies have indicated the worth of xR technologies in aviation. Rather,

they indicate the importance of ensuring a given technology is suited to its intended task. The use case considered here involved a qualified role-player providing an effect to a PTA in a distributed training context, with a relatively simple flying task. Two of the visualizations used in the study (V1 and V2) did not have peripheral views, and therefore formation was maintained using information provided by the TACPLOT. Had the Lead aircraft tracks been less predictable (e.g., climbing over the water), participants may have been required to make more use of their peripheral view, which may have favoured the MR condition. However, this would have to be verified empirically through an appropriately designed study. As noted below, having a complete body of evidence to support the application of specific technology to specific training use cases is an ongoing challenge facing military training organizations in general, a challenge that is all the more acute in the face of rapid advancements in technology, and the claims that often accompany the marketing of these technologies.

Limitations

An important limitation of this study is the relatively low number of participants (10 only), which consequently reduced the statistical power for the study. This low number is a consequence of the very specialized use case examined and the relatively small number of qualified participants in the CAF. We attempted to mitigate this with a mixed design, with repeated measures for the main variables of interest. However, we were limited in our ability to analyze potentially relevant independent (i.e., between-group) factors such as order of presentation of visualization conditions.

Another limitation pertains to the collection of the SSQ, workload and SUS ratings. These were collected only once per visualization condition (i.e., after all 3 visualizations for that condition were complete), making it impossible to assess the role of specific scenarios in these ratings, or how they might have interacted with visualization conditions. Also, no baseline assessment of SSQ (prior to the scenarios) was collected, which is recommended by some researchers (Merchant & Kirollos, 2022). Neither was subsidence of simulator sickness symptoms after the experiment studied, unlike other studies (e.g., McCoy-Fisher et al., 2019), as it was not considered a variable of interest in this study; however, understanding how simulator sickness subsides for the typical user after particular use cases is a key safety consideration in the use of MR (and simulation in general) as a standard tool in training facilities.

Finally, the tracks produced by the Lead and Wingman (participant) aircraft – that is, the effect meant to be produced for the intended PTA – was not inspected by aerospace controller subject matter experts to ascertain their quality and adequacy for a PTA. This is a next step to be considered in future studies.

Challenges Supporting Training Technology Modernization Decisions

Though the general principle of using simulation for training is beyond question, reliable data on training technology (e.g., sim) effectiveness is limited, in particular for specific use cases. Kaplan et al.'s (2021) meta-analysis of MR technologies in training notes that "the current literature is surprisingly sparse." Furthermore, the available studies constitute a set of disparate findings that don't obviously generalize to other settings, and don't provide clear guidance for military applications. The study of training effectiveness in general is riddled with challenges, from the very specific nature of most training interventions (limiting generalizability) to the challenges in conducting empirical studies (number and access to participants, unwillingness to divert limited training resources to support studies, etc.; for further discussion see Boldovici et al., 2002 & Kirkpatrick, 1979).

The study presented above is one such study on a specific application, with a limited number of participants and statistical power. However, it represents a first step in a systematic collaborative effort by DRDC and the RCAF AWC to investigate cases of relevance to the RCAF with scientific rigour, to help the RCAF and the CAF better prepare personnel for future operations and challenges. In addition to examining other aspects of the distributed simulation training use case established above (most notably the PTA itself), this effort looks to examine training technology effectiveness in other use cases relevant to the RCAF (e.g., pilot training) and the CAF in general. In this sense, the study described above is offered as the first step in a crawl-walk-run effort to provide the CAF with solid evidence to make decisions on training technology.

Once evidence is generated, it must be organized and synthesized to support effective policy and training system decisions for large organizations like the CAF. As noted above, such organization and synthesis are lacking in the current literature. Thus, the study presented here also constitutes an input in DRDC's ongoing efforts to provide the CAF with a systematic body of knowledge about the application of training technologies across various use cases.

This body of knowledge will be leveraged to develop tools (e.g., flowcharts, decision aids) to support policy and process decisions about training technologies. Initial efforts by DRDC in this direction have included the development of a framework for identifying barriers to the application of technologies for military training (Jarmasz & Martin, 2018), a review of technology adoption models with respect to training technologies (Emond et al. 2022) and have identified a variety of factors involved in successful training technology implementation over the whole technology lifecycle (Parkinson, 2022). Further activities similar to the study presented here will enable this broader "training technology adoption toolkit" effort to advance by adapting its application to specific use cases. Such a "toolkit" will be essential for the CAF to leverage the rapidly evolving technology landscape effectively as it modernizes its training approaches to better meet the demands of future operational activities and environments.

CONCLUSION

Despite the wealth of studies on the use of visualization technologies, the rapidly evolving field of xR (including AR, VR and MR) for training in specific use cases (such as various aspects of pilot training) still requires empirical study. In this paper, we present one such study, comparing visualizations with a proven track record in distributed collective training for air operations (conventional flat and curved monitors) against an emerging contender technology (MR headset) for a very specific use case in distributed mission training (fast jet role player). Our study found that, for the specific application considered, the MR technology did not provide an appreciable performance advantage over the established technologies, while introducing increased workload and simulator sickness and decreasing usability in our study sample. However, compared against studies on other use cases, our findings do not indicate the unsuitability of MR for pilot training in general; rather, they highlight the importance of matching technologies to the requirements of specific use cases, and of guiding specific applications with evidence relevant to them.

This study also represents a first step in a collaboration between the RCAF and DRDC on building a body of evidence on the use of various technologies to support the modernization of training in the RCAF, which will also benefit the modernization of training in the CAF. It is intended for this body of evidence to guide procedures and policies on the use of training technologies in the RCAF as well as other parts of the CAF. DRDC is undertaking a program of research to better inform the selection, insertion and employment of training technologies across various use cases and mission sets in the CAF, which includes generating empirical evidence on the performance of technologies as well as developing decision stupor aids based on existing evidence from various sources. These activities would also be of benefit to allied nations who are also engaged with training transformation and dealing with the challenges of making evidence-based decisions on the selection and application of technologies for training.

ACKNOWLEDGEMENTS

The authors wish to acknowledge RCAF AWC staff who contributed to the conduct of this study, namely, Second Lieutenant Keirt Aman, Private Svyatoslav Koval, Co-Op student Ms. Yoojin Jung, Co-Op Student Mr. David Kenyi, Co-Op Student Mr. Sundar Vengadeswaran, Co-Op Student Mr. Cameron Johnston, and Co-Op Student Ms. Nadeen Mortaja. They also wish to acknowledge valuable advice from Dr. Ramy Kirollos (DRDC) on the analysis and interpretation of simulator sickness data.

REFERENCES

Arcioni, B., Palmisano, S., Apthorp, D. & Kim, J. (2018). "Postural stability predicts likelihood of cybersickness in active HMD-based virtual reality." Displays. http://doi.org/10/1016/j.display.2018.07.001.

Boldovici, J.A., Bessemer, D.W., & Bolton, A.E. (2002). The Elements of Training Evaluation. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Bos, J.E., Lawson, B.D., Allsop, J., Rigato, P. & Secci, S. (2021). "Introduction in Guidelines for Mitigating Cybersickness in Virtual Reality Systems. Peer-Reviewed Final Report of the Human Factors and Medicine Panel/Modeling & Simulations Group, NATO STO-TR-HFM-MSG-323: Chapter 2.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), Usability evaluation in industry (pp. 189–194). London: Taylor & Francis.

Emond, B., & Durand, G. (2022). Support to training and education modernization: Training technology adoption. DRDC Contract Report

Government of Canada (2022). CDS/DM Directive for CAF Reconstitution. Accessed on 24 Apr 2023 from https://www.canada.ca/en/department-national-defence/corporate/policies-standards/dm-cds-directives/cds-dm-directive-caf-reconstitution.html

Hart, S. G. (2006). 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). Los Angeles, CA: Sage Publications.

Jarmasz, J., & Martin, B. (2018). Distributed simulation for training: Promises, barriers and pathways. (STO-MP-MSG-159), doi:10.14339/STO-MP-MSG-159

Jentsch, F., Curtis, M., and Salas, E. (2011). Simulation in aviation training. Farnham: Ashgate

Lackey, S.J., Salcedo, J.N., Szalma, J.L, & Hancock, P.A. (2016). The stress and workload of virtual reality training: the effects of presence, immersion and flow. *Ergonomics 59(8),*1-13, DOI: 10.1080/00140139.2015.1122234.

Kaplan A.D., Cruit J., Endsley M., Beers S.M., Sawyer B.D., & Hancock P.A.. The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis. Human Factors. 2021;63(4):706-726. doi:10.1177/0018720820904229

Kennedy, R. S., Drexler, J. M., Compton, D. E., Stanney, K. M., Lanham, D. S., and Harm, D. L. (2003). Configural scoring of simulator Sickness, Cybersickness and space adaptation syndrome: Similarities and differences. *Virtual and adaptive environments: Applications, implications, and human performance issues, 247.*

Kennedy, R. S., Lane, N. E., Berbaum, K. S. and Lilienthal, M. G., (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. International Journal of Aviation Psychology, 3, 203–220, doi:10.1207/s15327108ijap0303_3.

Kirkpatrick, D.L. (1979). Techniques for evaluating training programs. Training and Development Journal, June 1979, 78-92.

Parkinson, G.C. (2022). Training technology support tool development. DRDC Contract Report DRDC-RDDC-2022-C295.

Kirollos, R., & Merchant, W. (2022). An Overview of Cybersickness Self-Report Measures for use in Defence Research and Development Canada Experiments. DRDC Reference Document DRDC-RDDC-2022-D063.

Salas, E., Tannenbaum, S.I., Kraiger, K., Smith-Jentsch, K.A. (2012).The science of training and development in organizations: what matters in practice. Psychological Science in the Public Interest, 13(2), 74–101.

Van Weelden, E., Wiltshire, T.J, Alimardani, M., & Louwerse, M.M. (2022). Comparing Presence, Workload, and Performance in Desktop and Virtual Reality Flight Simulations. In Proceedings of the 2022 HFES 66th International Annual Meeting, p. 2006-2010.

Xi., N., Chen, J., Gama, F., Riar, M., & Hamari, J. (2023). The challenges of entering the metaverse: An experiment on the effect of extended reality on workload. In *Information Systems Frontiers, 25,* 659–680. DOI https://doi.org/10.1007/s10796-022-10244-x