Toward Augmenting Army Aviation Collective Training with Game-Based Environments

Lauren Reinerman-Jones, Martin S. Goodwin, Andrew J. Wismer, Brian F. Goldiez
University of Central Florida
Institute for Simulation and Training (UCF IST)
Orlando, FL
mgoodwin@ist.ucf.edu, lreinerm@ist.ucf.edu, awismer@ist.ucf.edu, bgoldiez@ist.ucf.edu

Robert A. Crapanzano
U.S. Army Program Executive Office for Simulation, Training, and Instrumentation (PEO STRI)
Orlando, FL
robert.a.crapanzano.mil@mail.mil

ABSTRACT

Maintaining the highest levels of training and readiness is an ongoing concern for today's warfighters. A rapidly evolving threat landscape, ever-present fiscal concerns, and the move toward virtualization are driving the need for more efficient training methods. The U.S. Army is addressing this need by investigating the potential of game-based systems to augment traditional simulation-based aviation collective training. Game-based training is one component of the U.S. Army Aviation Combined Arms Training Strategy (2016), which highlights the use of Training Aids, Devices, Simulations, and Simulators (TADSS) as key, low-cost tools to prepare Army aviation forces for future combat. However, the effectiveness of game-based training requires further investigation, and its use as an adjunct to aviation collective training has not been adequately evaluated. The goal for the present study was to determine the potential for the low physical fidelity Virtual Battlespace 3 (VBS3) games-for-training system to augment aviation collective training conducted in the medium physical fidelity Aviation Combined Arms Tactical Trainer (AVCATT). Evaluation efforts focused on the cognitive fidelity of these training systems. Twenty-seven expert pilots participated in a realistic collective air assault mission scenario first in either the VBS3 or AVCATT training environment and then in a high fidelity Operational Flight Trainer (OFT) serving as a real world analog environment. Each environment was evaluated in terms of presence, simulation sickness, workload, performance, and HRV. The cognitive fidelity of the OFT corresponded more closely with the AVCATT than VBS3. Objective performance was comparable between the AVCATT and VBS3 and did not lead to performance differences in the OFT. This paper concludes by discussing potential ways to augment collective aviation training with lower fidelity game-based systems and by proposing design improvements for simulated collective training environments.

ABOUT THE AUTHORS

Lauren Reinerman-Jones, Ph.D. is the Director of Prodigy, which is one lab at the University of Central Florida's Institute for Simulation and Training, focusing on assessment for explaining, predicting, and improving human performance and systems.

Martin S. Goodwin, Ph.D. has over 30 years of experience in the research and development of dynamic instructional systems, simulation and gaming technology integration, and evaluation methodologies to improve learning, engagement, and retention in virtual environments.

Andrew J. Wismer, M.A. is a Human Factors & Cognitive Psychology Ph.D. Candidate at the University of Central Florida with research experience in categorization, decision making, and learning in dynamic environments.

Brian F. Goldiez, Ph.D. is the Deputy Director of the University of Central Florida's Institute for Simulation and Training and a Research Associate Professor at UCF. Dr. Goldiez has over 40 years of modeling and simulation experience spanning Government, Industry, and Academia. His principal focus has been oriented to optimizing technology to enhance human performance.

Robert A. Crapanzano, MAJ, is an Assistant Product Manager at the U.S. Army's Program Executive Office for Simulation, Training, Instrumentation and is a doctoral student at Embry Riddle Aeronautical University.

Toward Augmenting Army Aviation Collective Training with Game-Based Environments

Lauren Reinerman-Jones, Martin S. Goodwin, Andrew J. Wismer, Brian F. Goldiez
University of Central Florida
Institute for Simulation and Training (UCF IST)
Orlando, FL
mgoodwin@ist.ucf.edu, lreinerm@ist.ucf.edu, awismer@ist.ucf.edu, bgoldiez@ist.ucf.edu

Robert A. Crapanzano
U.S. Army Program Executive Office for Simulation, Training, and Instrumentation (PEO STRI)
Orlando, FL
robert.a.crapanzano.mil@mail.mil

INTRODUCTION

Maintaining the highest levels of training and readiness is an ongoing concern for today's warfighters. For the United States, the Global War on Terrorism has been characterized by the Army Force Generation (ARFORGEN) training model with rapid successions of training, deployment, and reset periods. As the threat landscape continues to evolve, training requirements shift in preparation for future, anticipated combat scenarios. Technological innovations are changing the characteristics of modern warfare.

The Aviation Combined Arms Training Strategy (2016) highlights Training Aids, Devices, Simulations, and Simulators (TADSS) as key, low cost, tools to prepare the Army Aviation force for future combat. TADSS cost less than equivalent aircraft or flight hours and can perform training tasks that would otherwise be extremely costly or dangerous to perform. For example, pilots can perform engine failure emergency scenarios, simulate combat operations against surface to air missiles, or fire dozens of virtual Hellfire missiles that would costs hundreds of thousands of dollars in the real world. TADSS are available in a range fidelities. Device fidelity is the degree of similarity between a simulated environment and the environment being simulated; high realism generally translates to high fidelity (Alessi, 1988). Fidelity can be broadly classified as either physical or cognitive (Liu, Macchiarella, & Vincenzi, 2008). Physical fidelity corresponds to how well a training context physically reproduces the performance context. Cognitive fidelity concerns how well the training context stimulates the main psychological effects and mechanisms (e.g., workload) involved in the performance context (Kozlowski & DeShon, 2004). Three systems currently utilized by the Army are described below and each varies in fidelity.

The Operational Flight Trainer (OFT) is a high-fidelity flight simulator staged on a 3 degree-of-freedom (DOF) vibration platform that is further mounted on a 6-DOF electric motion system to support full motion simulation (See Figures 1-C). A 200° x 45° visual display, the inclusion of pilot and co-pilot chin windows, and a high-definition projection system supports out-the-window imagery. The OFT includes actual replications of the UH-60A/L's cyclic and collective controls for the flight control system. The OFT also uses a physics-based model and high-fidelity software modeling to accurately simulate onboard aircraft components (L-3 Link Simulation & Training, 2010).

The Reconfigurable Collective Training Device (RCTD) is a moderate physical fidelity training device containing flight controls, displays, and switches that match the tactile feel and most functionality of the actual simulated aircraft (see Figure 1-A). Instead of the traditional projection system seen in most flight simulators, the RCTD uses augmented reality helmet mounted displays (HMDs) to blend the physical cockpit with the virtual environment. The transmissive lenses of the HMD allows pilots to see the physical cockpit when they look at the cockpit display and see the virtual

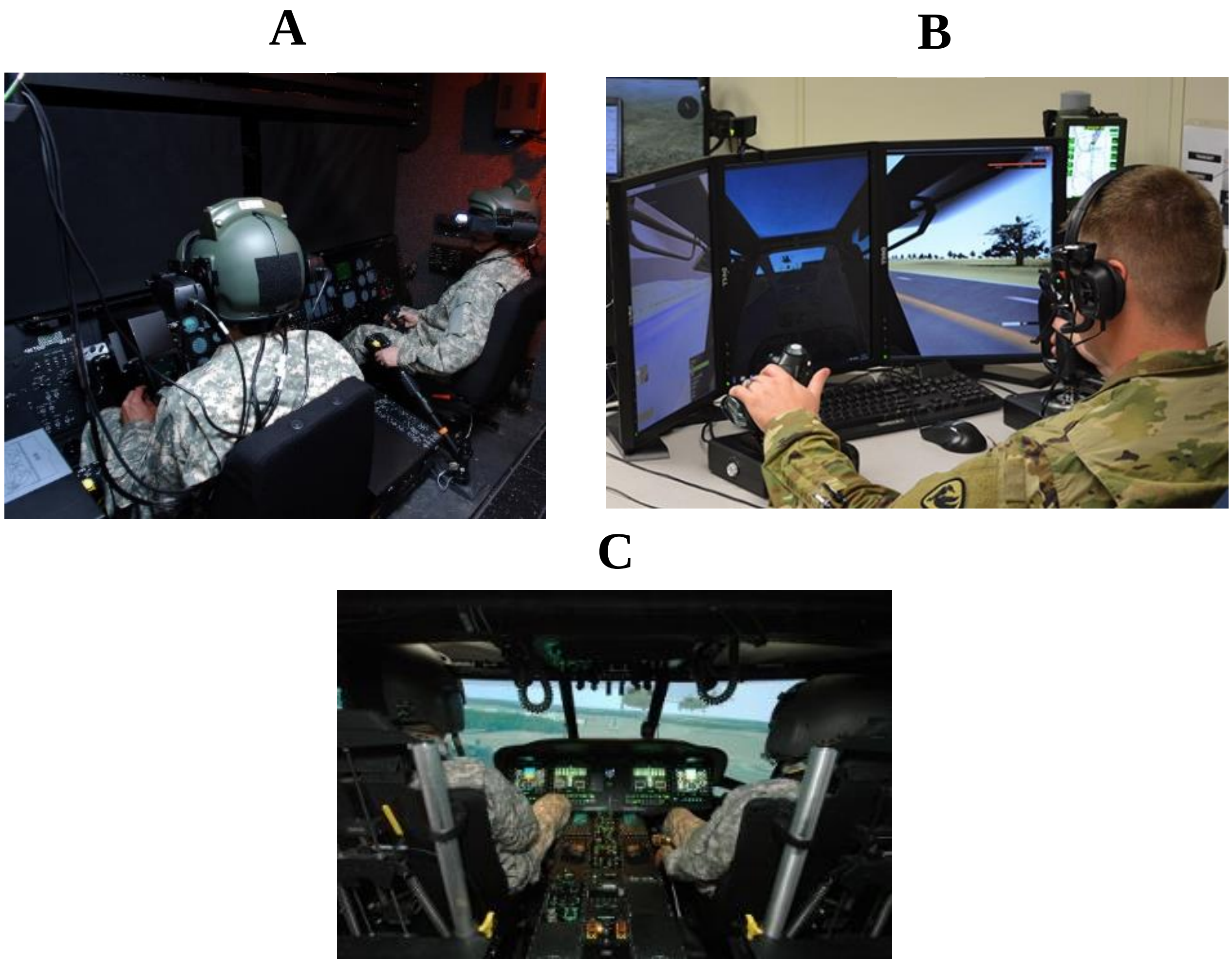

Figure 1. Images of RCTD (Figure A; L-3 Link Simulation & Training, 2012), VBS3 (Figure B; Morris, 2015), and the OFT (Figure C; Tony, 2013).

world when they look above the dashboard (Rockwell Collins, 2017). The RCTD cockpit is reconfigurable and can be set up to simulate the UH-60A/L Blackhawk (as in the present study), AH-64 Apache, CH-47 Chinook, or UH-72 Lakota helicopters. Furthermore, the RCTD can be networked with the Armor Close Combat Tactical Trainer, the remotely piloted aircraft Universal Mission Simulator, VBS3, or various other TADSS.

Virtual Battle Space 3 (VBS3) is a low physical fidelity, first-person, games-for-training system developed by Bohemia Interactive Simulations. It can operate on a laptop or desktop computer (see Figure 1-B). Pilots can use the game with a mouse and a keyboard or they can add additional peripheral devices such as multiple monitors or commercial-off-the-shelf (COTS) video game flight controllers. In the game, pilots can train in scenarios with various military and civilian entities. In exercises, VBS3 can be networked with other systems to support collective training with teammate aircraft or other operators from infantry, armor, artillery, or headquarters units. VBS3 allows for repeated training of field tactics without costly ammunition, travel time, or risk of injury or damage ("Virtual Battlespace 3," 2017). While VBS3 has a significant network capability and low cost, it has limited functionality compared to the actual aircraft. Most avionics and weapon systems are not simulated in VBS3 and the controls differ from an actual UH60A/L.

However, fiscal concerns are driving the need for more efficient training methods. Cost-effective training strategies are needed to prepare warfighters for the challenges of highly dynamic military operations without straining the defense budget. Therefore, game-based simulations like VBS3 are potential solutions to this problem. Game-based simulation has become more sophisticated and may provide viable training options for some applications. The use of game-based systems to augment higher fidelity simulation-based training could help optimize training resources by focusing on the cognitive aspects of training simulations. Game-based simulation costs less due to lower physical fidelity. A game-based environment that stimulates a similar level of cognitive fidelity as higher physical fidelity training environments could provide a cost-effective supplement for certain training tasks. This approach may also enhance return on investment, advance training objectives, and inform the design of future training environments.

Previous research analyzed the relationship between individual performance and flight training device fidelity (e.g., Sotomayor & Proctor, 2009). However, little or no research exists concerning the relationship between collective aviation training and device fidelity (see Whitney, Temby, & Stephens, 2014). Further, work along these lines typically falls into Training Effectiveness Evaluations (TEE) and current practices of executing a TEE are based upon Kirkpatrick's Model whereby training TEEs are less concerned about improving a single training program and more concerned about proving the efficacy of specific individual factors that influence training effectiveness. This focus on proving instead of improving necessitates the use of a TEE approach based on a research methodology instead of a standard evaluation methodology. It is from this perspective that the interdisciplinary TEE approach called Assessing Simulated Systems Empirically for Training (ASSET) was developed (see Goodwin, Reinerman-Jones, Goldiez, & Crapanzano, 2017). That approach leverages an evaluation paradigm better aligned with the purpose and objectives of TEEs in modern, technology-enabled training environments. ASSET increases the breadth of evaluation efforts to more fully capture the range of factors that contribute to training effectiveness in dynamic, interactive simulation training environments. ASSET follows the procedures and rigor of a research methodology, with some modification to optimize its use to conduct TEEs in simulation training environments (see Figure 2).

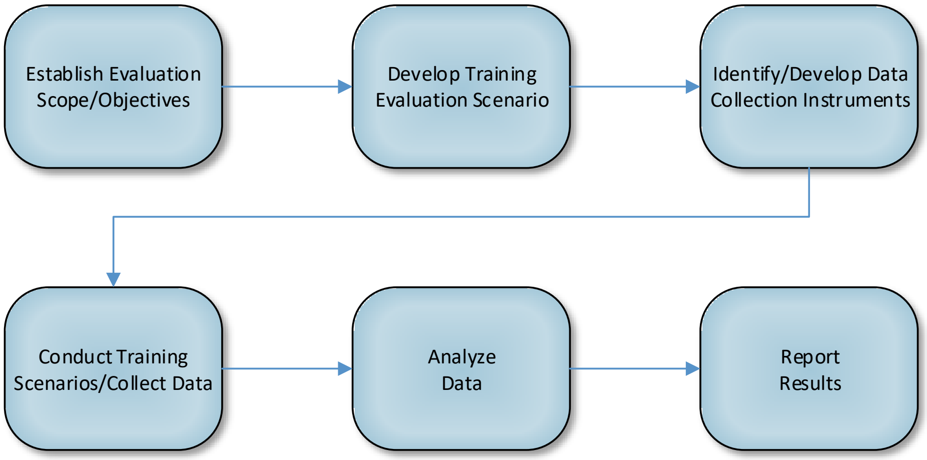
Establish Evaluation Scope/Objectives
Develop Training Evaluation Scenario
Identify/Develop Data Collection Instruments
Conduct Training Scenarios/Collect Data
Analyze Data
Report Results

Figure 2. ASSET Evaluation Approach

Study Objectives

The present work compared two TADSS – the RCTD and VBS3 - for a set of collective aviation training tasks (i.e., team and teams-of-teams training). The RCTD is the stationary version of the Aviation Combined Arms Tactical Trainer (AVCATT), which is the current Program of Record for Army aviation collective training. VBS3 is the principal component of the Army's games-for-training initiative. These environments were first compared to each other and then evaluated for their transfer of training to the operational setting. The OFT served as the real-world analog test environment. Following the ASSET approach, the scope and objectives of the training evaluation were determined. The goal for the present study was to determine if the low physical fidelity VBS3 games-for-training system could provide comparable levels of cognitive fidelity as the medium physical fidelity RCTD for a set of aviation collective training tasks. Cognitive fidelity is believed to be associated with positive transfer of training (De Winter, Wieringa, Dankelman, Mulder, Van Paassen, & De Groot, 2007; Hochmitz & Yugiler-Gavish, 2011; Kozlowski & DeShon, 2004), and as such, forms an important basis for evaluating training effectiveness. Critically, it is unclear if there would be a significant disadvantage if any of the collective training tasks currently performed in the RCTD were augmented with game-based systems such as VBS3. Augmenting training with game-based systems could have significant cost savings, but identifying the training and performance outcomes associated with level of cognitive fidelity associated with each environment is needed first.

METHOD

Participants

The study involved 27 previously qualified Army aviators recruited from different USAACE schoolhouses at Fort Rucker, Alabama. Fifteen participants completed the RCTD condition and 12 participants completed the VBS3 condition. Participant ages ranged from 26 to 53 (*M* = 35.59, *SD* = 6.35). Pilots had anywhere from 4 to 27 years of experience (*M* = 14.07, *SD* = 6.03) and up to six deployments (*M* = 2.48, *SD* = 1.67). Flight hours for the UH-60A/L and rotary aircrafts in general ranged from 35 to 6600 hours (*M* = 1493.22, *SD* = 1466.14) and 200 to 7100 hours (*M* = 1801.22, *SD* = 1562.20), respectively.

Experimental Design

The experimental approach consisted of a 2 (Simulated Training Environment: RCTD and VBS3) x 1 (Real-World Analog Environment: OFT) mixed design with repeated measures on the Real-World Analog Environment. Performance was first compared among simulated training environments and then within the OFT based on the preceding simulated training environment. The primary metric of interest consisted of change scores compared among simulated training environment conditions. Change scores were calculated by subtracting the value of a dependent measure in the simulated training environment from the corresponding dependent measure evaluation in the OFT; change scores provide a measure of correspondence between training and performance contexts.

Each participant completed two experimental sessions over the course of one day. Study participation time was approximately 7 hours (3.5 hours each session, morning and afternoon) with a lunch break between sessions. The morning session involved a mission scenario conducted in one of the two simulated training environments (RCTD or VBS3). Only one flight simulated environment could be run per week due to aviator availability, facility space, flight equipment availability, and Army support staff availability. Participants run during the first week of data collection completed the RCTD simulated training environment while participants in the second week completed VBS3. Each afternoon session involved a similar mission scenario conducted in the near-real-world analog environment (OFT). Each environment – training and OFT - used two-person crews comprised of a pilot and a copilot. Study participants were the pilots. Copilots were study confederates who were briefed on their proper role in the study. All experimental sessions were conducted at Fort Rucker, Alabama under the operation of the USAACE Directorate of Simulation.

Experimental Scenarios

Mission scenarios, involving a flight of UH-60A/L Blackhawk helicopters engaged in a collective air assault mission, formed the basis of the training evaluation. The mission scenarios were developed in conjunction with the U.S. Army Program Executive Office for Simulation, Training, and Instrumentation (PEO STRI) and the U.S. Army Aviation Center of Excellence (USAACE) Directorate of Simulation (DOS). These scenarios consisted of operationally demanding tasks, cognitive decision-making points, and flight metrics that formed a set of performance measures for the evaluation. Operational tasks focused on mission events from standard operating procedures or specific items covered in mission/crew briefings. Decision-making points involved the pilot's specific choices and reactions to changing mission scenario conditions.

Equated mission scenarios were used for the VBS3 and RCTD simulated training environments and were similar with respect to terrain, weather conditions, and approximate flight time and distance. The mission scenario positioned the pilot participant as an Air Mission Commander of a General Support Aviation Battalion (GSAB) leading a flight of two UH-60A/L aircraft engaged in collective air assault operations. Remaining roles were played by other aviators, researchers, or simulator operators.

The mission scenario employed a narrative involving hostile conditions where enemy forces attempted to seize key terrain. The mission involved the air assault of quick reaction force (QRF) soldiers from a pickup zone (PZ) into a landing zone (LZ). Along the way, pilot tasks included reporting several ACP waypoints, performing fuel checks, and requesting status updates and clearances. The end of the scenario included a casualty evacuation (CASEVAC) event, of which the participant was not pre-briefed. A comparable scenario with similar mission tasks and decision points was used in the OFT.

Dependent Measures

Study measures included subjective measures of presence, simulation sickness, and workload, seven performance metrics relating to flight tasks, cognitive decisions, and CASEVAC flight time, one physiological measure of heart rate variability, and phenomenological interviews following mission completion. In accordance with the applied ASSET approach, this interdisciplinary suite of measures was employed to provide a more comprehensive picture of training effectiveness.

Presence is a self-report measure that quantifies subjective experiences of involvement and immersion within a virtual environment (Witmer & Singer, 1994). Presence was assessed using the Presence Questionnaire by Witmer and Singer (1998). The 7-point scale questionnaire assesses presence along seven factors: Realism, Possibility to Act, Quality of Interface, Possibility to Examine, Self-Evaluation of Performance, Sounds, and Haptic. Presence was assessed after the participant completed the training and test simulated environments.

Simulation sickness is another self-report measure that assesses the level of discomfort experienced by participants in each simulated environment (SSQ; Kennedy et al., 1993). The SSQ is clustered into three factors: Nausea, Oculomotor Discomfort, and Disorientation. The 4-point scale questionnaire assesses the degree to which each symptom is affecting the participant in the present moment (1 = “none”; 4 = “severe”). The SSQ was administered as a baseline before the study (not reported here) and after both experimental sessions, and raw (unweighted) scores served as a comparison between simulated environments.

The NASA-Task Load Index (TLX; Hart & Staveland, 1988) was administered after the training and test simulated environments to assess participants’ perceived workload while performing the mission scenario. The NASA-TLX is composed of six subscales measuring workload across the dimensions of Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration, and Performance.

Mission performance was assessed along three dimensions: tasks, decisions, and CASEVAC Flight Time. There were ten mission tasks that were broken down into five categories (with the number of corresponding tasks in parentheses): ACP Reports (3), Clearances (2), Fuel Checks (2), Landings (1), and Status Requests (2). There were four decisions that participants could correctly make in the scenario and the proportion of total correct decisions served as the sixth performance metric. Cognitive decisions involved situational judgment tasks and responses to changing mission parameters. Finally, the flight time for the CASEVAC portion of the mission served as the seventh and final dimension of performance. Flight time incorporated elements of a participant’s route selection, airspeed, and approach/landing to the base airfield.

Heart rate variability (HRV) was captured using a Procomp Infiniti System by Thought Technology Ltd. (see Figure 3). Participants wore sensor leads on their right index and ring fingers. These sensors were used because of the availability of associated software development kits (SDKs) that enabled data to be logged locally, in real time, and synchronized with the simulated environments. The sensors presented minimal interference with the performance of the tasks. HRV is generally associated with changes in cognitive workload and engagement in effortful information processing (Jorna, 1993). Increases in cognitive workload of task demands are associated with decreases in HRV (an inverse relationship; Mulder, de Waard, & Brookhuis, 2004).

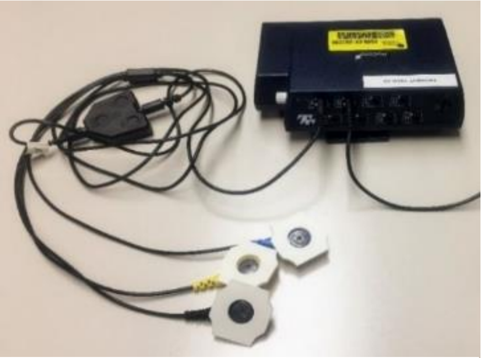

Figure 3. Procomp Infiniti System

Finally, phenomenological interviews were conducted with pilots at the end of each experimental session, following the method of Petitmengin (2006) and implementation guidelines of Bockelman, Reinerman-Jones, and Gallagher (2013). Interview questions were designed for the participant to focus on their real-time subjective experience of completing the mission scenario in the simulated training environment. Copilot interviews were also conducted to provide an additional metric regarding pilot performance within each simulated training environment. Findings from the interviews are reviewed in the discussion section to shed light on differences found within other study measures.

Procedure

Upon entering the Seneff Aviation Warfighting Simulation Center at Fort Rucker, Alabama, participants completed an informed consent and were paired (randomly) with a copilot. Participants were provided mission materials including an air assault route map and kneeboard packet for the simulated mission scenario. Participants were then provided an overview of the study by the lead researcher and a mission briefing by an Army operations officer.

Upon completion, copilots were taken to a separate room. Participants completed a demographics survey, restrictions checklist, and baseline simulation sickness questionnaire on a Microsoft Surface tablet, while copilots were briefed on their role in the study and were provided with a script for the crew brief. Then, each study participant and copilot were reunited and escorted to the appropriate simulated training environment where ECG and GSR sensors were applied to the participant. The participant received a brief familiarization period with the simulation device, followed by an aircrew briefing by the copilot. Once ready to begin, a physiological baseline was conducted. This was a five minute resting baseline where the participant was instructed to remain still, look forward, and stay silent. This baseline served as a reference against which potential physiological changes during the mission could be measured.

After the baseline, the mission scenario began. Sessions included a video recording of the simulator instrument panel. Pilots completed the mission scenario which consisted of ten tasks involving radio calls, fuel checks, pickups and drop-offs. Each scenario also included four decision points and an unanticipated CASEVAC.

At the end of the mission, copilots were sent to a separate room for a post-mission briefing while the ECG and GSR sensors were removed from the pilot and video recording ended. The pilot then completed the Presence Questionnaire, Simulation Sickness Questionnaire, and NASA-TLX. Phenomenological interviews were conducted separately for the pilot and copilot. When finished, the participant and copilot were dismissed for a lunch break. After the lunch break, participant pairs reconvened in the lobby. The general procedure that followed was identical to the morning session with two exceptions: 1) there were no initial surveys to complete in the lobby, and 2) each pair of participants was taken to an OFT serving as a real-world analog environment. PQ and NASA-TLX surveys were completed following the mission scenario, as before. Participants were encouraged to perform an after action review before being dismissed. The study took participants approximately 7 hours or less to complete, and participants were compensated at their regular hourly wage.

RESULTS

Data were analyzed using SPSS 24. Data were analyzed separately within training conditions and within the OFT based on the preceding environment, but the main area of interest was in the change scores (difference in a dependent variable between training and test environments). Analyses consisted of independent samples t-tests conducted between simulated training environment conditions. When unequal variance or normality assumptions were violated, Welch’s t-tests or Mann Whitney U tests were used, respectively. Traditional analyses are accompanied by an estimated Bayes Factor obtained through JASP computer software (JASP Team, 2016). The Bayes Factor (BF_{10}) provides an odds ratio for the alternative/null hypothesis (values < 1 favor the null hypothesis and values > 1 favor the alternative hypothesis). Bayes Factors provide information on the strength of the evidence for or against differences between the two simulated training environment conditions.

Demographics

There were no significant differences between simulated training environment conditions on any of the demographic measures (all *p*’s > .104; all *BF*’s < 0.68). Conditions were similar with respect to average age, years in Army, number of deployments, lifetime UH-60A/L hours, and lifetime rotary aircraft hours (see Table 1).

Table 1. *M* (and *SD*) of Demographic Variables by Condition

Condition	Age	Years in Army	Total Deployments	Lifetime UH-60A/L Hours	Lifetime Rotary Hours
RCTD (*N* = 15)	34.67 (6.58)	12.47 (6.99)	2.60 (2.13)	1309.33 (1124.90)	1470.40 (1224.26)
VBS3 (*N* = 12)	36.74 (6.14)	16.08 (3.99)	2.33 (0.89)	1723.08 (1834.81)	2214.75 (1877.51)

Differences between Simulated Training Environments

Significant differences were found between the RCTD and VBS3 simulated training environments with respect to presence, frustration, and disorientation. First, the RCTD was rated higher than VBS3 on five factors of presence (realism, possibility to act, quality of interface, possibility to examine, and haptic), with the strongest support for a difference in possibility to act (*BF* = 1260.17). Second, average frustration levels were significantly higher in VBS3 (*M* = 70.42, *SD* = 30.34) than in the RCTD (*M* = 34.33, *SD* = 22.75), *t*(25)= -3.54, *p* = .002, *BF* = 20.85. Third, the disorientation factor of the SSQ revealed significantly higher average disorientation (e.g., dizziness, vertigo) in the RCTD simulated training environment (*M* = 10.20, *SD* = 2.70) compared to VBS3 (*M* = 8.25, *SD* = 1.96), *t*(24.81) = 2.17, *p* = .040, *BF* = 1.70. There were no other significant differences (all *p*'s > .069; all *BF*'s < 1.86; See Table 2).

Table 2. Statistical Analyses between Simulated Training Environments

Dependent Measure		RCTD *M* (*SD*)	VBS3 *M* (*SD*)	Test Statistic	*p*-value	*BF*
Presence	Realism	29.33 (7.39)	20.50 (4.87)	*t*(25) = 3.56	.002	21.91
	Possibility to Act	19.47 (2.92)	13.42 (2.81)	*t*(25) = 5.43	< .001	1260.17
	Quality of Interface	14.60 (3.00)	9.67 (2.77)	*t*(25) = 4.39	< .001	125.51
	Possibility to Examine	12.93 (2.05)	10.33 (3.00)	*t*(25) = 2.67	.013	4.21
	Self-Evaluation	9.40 (1.88)	8.17 (2.73)	*t*(25) = 1.39	.177	0.73
	Sound	14.00 (2.85)	12.17 (3.59)	*t*(25) = 1.48	.151	0.80
	Haptic	7.27 (3.20)	4.58 (2.28)	*t*(25) = 2.45	.022	2.92
SSQ	Nausea	9 (5)*	8 (1.75)*	*U* = 121.50	.121	0.87
	Oculomotor	12.13 (3.98)	10.33 (3.28)	*t*(25) = 1.26	.219	0.64
	Disorientation	10.20 (2.70)	8.25 (1.96)	*t*(24.81) = 2.17	.040	1.70
NASA-TLX	Mental Demand	75 (20)*	67.5 (15)*	*U* = 67	.277	0.56
	Physical Demand	27.67 (19.99)	41.67 (28.63)	*t*(25) = -1.50	.147	0.81
	Temporal Demand	53.00 (22.35)	49.17 (20.65)	*t*(25) = 0.46	.651	0.39
	Effort	56.00 (22.69)	69.17 (17.69)	*t*(25) = -1.65	.112	0.96
	Frustration	34.33 (22.75)	70.42 (30.34)	*t*(25) = -3.54	.002	20.85
	Performance	27.33 (24.46)	42.08 (30.11)	*t*(25) = -1.41	.172	0.74
Tasks	ACP	3 (N/A)*	3 (1)*	*U* = 67.5	.277	1.85
	Clearance	2 (0)*	2 (1)*	*U* = 71	.373	0.75
	Fuel Check	2 (0)*	2 (0)*	*U* = 86	.867	0.40
	Landing	1 (1)*	1 (0)*	*U* = 61.5	.167	1.35
	Status Request	2 (0)*	2 (0)*	*U* = 86	.867	0.40
Decisions	Out of 4	2 (1)*	3 (1)*	*U* = 63.5	.200	0.65
CASEVAC Flight Time	In seconds	640.53 (89.52)	729.91 (150.72)	*t*(24) = -1.89	.070	1.30
HRV	Change from baseline	9.56 (34.12)	19.13 (18.19)	*t*(22) = -0.81	.429	0.48

Note. Asterisk (*) signifies the displayed values are median (and interquartile range) corresponding to nonparametric tests.

Differences in OFT based on Preceding Environment

In the OFT, when preceded by VBS3, participants self-evaluated their performance higher (*Mdn* = 13) than when preceded by the RCTD (*Mdn* = 12). This difference in self-evaluation of performance (Presence factor measuring speed of adjustment to virtual environment and proficiency in interaction) was significant, *U* = 43, *p* = .019, but the evidence for this difference is weak (*BF* = 2.06). In addition, there were several effects that approached significance: HRV [*t*(22) = -1.80, *p* = .086, *BF* = 1.16], realism [*t*(25) = -1.80, *p* = .085, *BF* = 1.15], and frustration [*t*(18.49) = 2.06, *p* = .054, *BF* = 1.28], though none of these effects received positive support from the Bayes Factors. There were no other significant or near significant differences in the OFT based on the preceding simulated training environment (all *p*'s > .107; all *BF*'s < 0.93). As a whole, performance in the OFT was not dependent on the previous simulated training environment. See Table 3 below for the full set of statistical analyses in the OFT based on the preceding environment.

Table 3. Statistical Analyses in OFT based on Preceding Environment

Dependent Measure		RCTD ***M* (*SD*)**	VBS3 ***M* (*SD*)**	Test Statistic	***p*-value**	***BF***
Presence	Realism	37.87 (5.21)	41.00 (3.41)	*t*(25) = -1.80	.085	1.15
	Possibility to Act	22.27 (2.40)	23.75 (2.34)	*t*(25) = -1.61	.120	0.92
	Quality of Interface	16.33 (3.44)	18.00 (1.54)	*t*(20.27) = -1.68	.108	0.87
	Possibility to Examine	15.60 (2.72)	15.58 (2.11)	*t*(25) = 0.02	.986	0.36
	Self-Evaluation	12 (1)*	13 (2)*	*U* = 43	.019	2.06
	Sound	14.87 (3.11)	15.08 (3.00)	*t* (25) = -0.18	.857	0.36
	Haptic	9.87 (3.48)	9.33 (4.08)	*t* (25) = 1.59	.124	0.38
SSQ	Nausea	8.27 (1.10)	8.50 (1.38)	*t*(25) = -0.49	.629	0.39
	Oculomotor	8.47 (1.64)	9.17 (1.80)	*t*(25) = -1.06	.302	0.54
	Disorientation	7.80 (1.15)	8.00 (1.28)	*t*(25) = -0.43	.672	0.39
NASA-TLX	Mental Demand	59.00 (26.13)	56.25 (26.12)	*t*(25) = 0.27	.788	0.37
	Physical Demand	30.00 (25.43)	35.83 (29.45)	*t*(25) = -0.55	.586	0.40
	Temporal Demand	41.67 (28.20)	45.83 (32.04)	*t*(25) = -0.36	.722	0.38
	Effort	48.67 (29.79)	34.58 (23.30)	*t*(25) = 1.34	.192	0.69
	Frustration	23.67 (26.69)	8.33 (9.85)	*t*(18.49) = 2.06	.054	1.28
	Performance	20 (55)*	12.5 (15)*	*U* = 65	.236	0.74
Tasks	ACP	3 (N/A)*	3 (0)*	*U* = 88	.943	0.56
	Clearance	2 (1)*	2 (0)*	*U* = 73.5	.427	0.61
	Fuel Check	2 (0)*	2 (0)*	*U* = 85	.829	0.44
	Landing	1 (N/A)*	1 (N/A)*	*U* = 90	1.000	N/A
	Status Request	2 (1)*	2 (0)*	*U* = 73.5	.427	0.61
Decisions	Out of 4	3 (1)*	3 (2)*	*U* = 86.5	.867	0.37
CASEVAC Flight Time	In seconds	527.07 (31.13)	519.36 (42.41)	*t*(23) = 0.53	.605	0.41
HRV	Change from baseline	24.54 (41.34)	54.28 (37.93)	*t*(22) = -1.80	.086	1.16

Note. Asterisk (*) signifies displayed values are median (and interquartile range) corresponding to nonparametric tests.

Correspondence between OFT and Simulated Training Environments

Additionally, change scores (OFT – training score) were evaluated as a measure of correspondence between simulated training environments and the test environment. Change scores were calculated by subtracting the value of one condition's dependent measure in the simulated training environment from the corresponding condition's dependent measure in the OFT. Change scores close to zero signify high correspondence, with change scores farther from zero suggesting more substantial differences between training and test environments.

Overall, RCTD and VBS3 were rated significantly different for four of the seven presence factors (realism, possibility to act, quality of interface, and self-evaluation of performance) with differences in the possibility to examine factor approaching significance, *t*(25) = 2.04, *p* = .052, *BF* = 1.58. The presence factor that received the most support for a difference between the two training environments was realism, *t*(25) = -5.81, *p* < .001, *BF* = 2939.76. There were no significant differences between simulated training environments for sounds [*t*(25) = 1.67, *p* = .107, *BF* = 0.99] or haptic [*t*(25) = 1.59, *p* = .124, *BF* = 0.91]. These results reflect the pattern of differences seen among simulated training environments and the lack of differences seen in the OFT based on the preceding environment.

There was a significantly higher correspondence in simulation sickness levels between VBS3 and the OFT than between RCTD and the OFT for disorientation, *t*(25) = -2.27, *p* = .032, *BF* = 2.01, and nausea, *U* = 47.5, *p* = .037, *BF* = 1.29, with the difference in oculomotor approaching significance, *t*(25) = -1.88, *p* = .072, *BF* = 1.28. With respect to workload, change scores revealed a higher correspondence between the OFT and RCTD than with VBS3 for effort, *t*(25) = 2.75, *p* = .011, *BF* = 4.80, frustration, *t*(25) = 5.16, *p* < .001, *BF* = 679.44, and performance, *t*(25) = 2.13, *p* = .043, *BF* = 1.79. The other three workload factors were not significant (all *p*'s > .377; all *BF*'s < 0.50). See Figure 4 for average workload scores by condition.

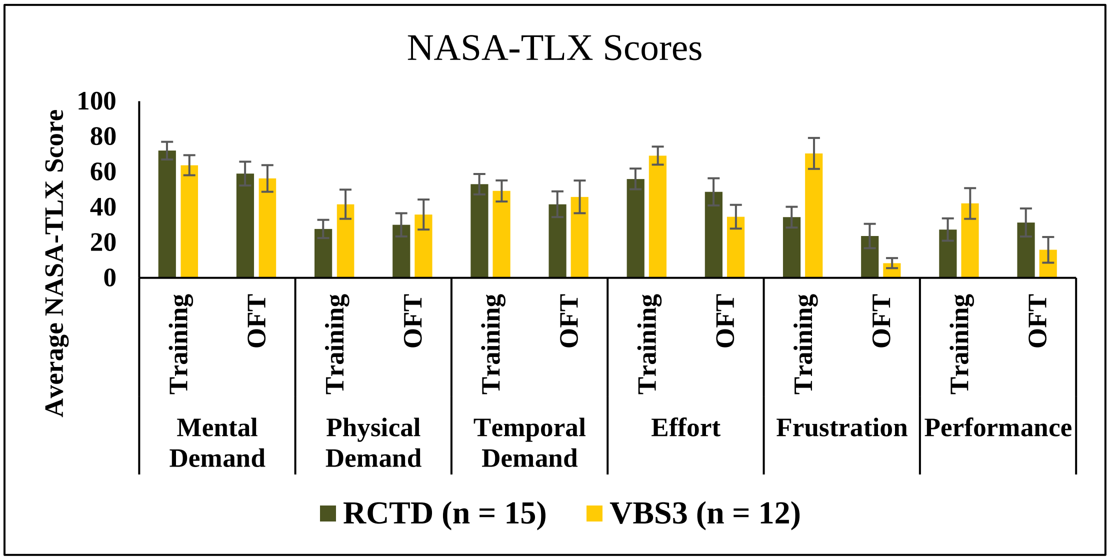

Figure 4. Average NASA-TLX Scores by Simulated Training Environment Condition

There were no significant differences in performance task change scores between RCTD and VBS3 (all *p*'s > .125; all *BF*'s < 2.63). The lack of task performance differences between the two simulated training environment conditions suggests similarities in their respective effectiveness for collective aviation training. There were also no significant differences in decisions, *U* = 66, *p* = .256, *BF* = 0.66, or CASEVAC flight time change scores, *U* = 47, *p* = .129, *BF* = 1.32. Figure 5 shows the average deviation from optimal performance by task type for each condition.

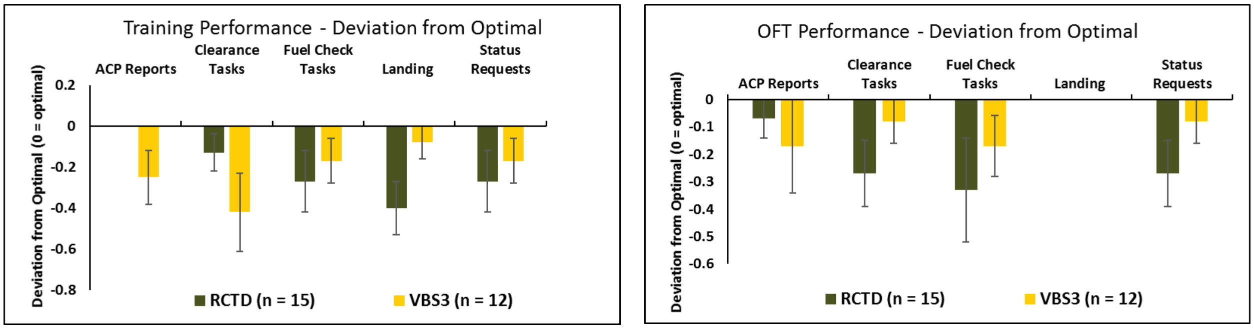

Figure 5. Average Task Performance by Category and Condition in Training and OFT

Finally, there were no significant differences between RCTD and VBS3 with respect to heart rate variability change scores, *t*(21) = -0.95, *p* = .352, *BF* = 0.53. HRV results mirror the cognitive demand factor from the NASA-TLX suggesting similar levels of cognitive demand between RCTD and VBS3.

Phenomenological Interviews

A hermeneutical analysis was conducted involving a key word analysis of pilot interview responses. Seven aspects of the training simulated environment were coded– graphics, controls, responsiveness, flight representation, starting focus, progression focus, and training platform (see Table 4). The results of the coding are displayed in Table 5 as the number of participants classified into each category. Coding allowed application of quantitative principals to qualitative data; however, it is important to note the value in also looking at qualitative data in its original linguistic, textual form.

Table 4. Pilot Interview Coding Categories

Category	Description
Graphics	The overall quality of the display of the simulated environment [Good/Neither/Bad]
Controls	The usability of the controls [Good/Neither/Bad]
Responsiveness	The ability of the game environment to respond appropriately to control inputs, i.e. minimal lag [Good/Neither/Bad]
Flight Representation	How well the game system accurately displayed flight parameters, e.g. airspeed, torque, altitude [Good/Neither/Bad]
Starting Focus	Where was the majority of the pilot's attention focused at the beginning of the mission? [Mission-focus/Both/Game-Focus]
Progression Focus	Where was the majority of the pilot's attention focused as the mission progressed? [Mission-focus/Both/Game-focus]
Training Platform	Does the pilot believe the simulated environment is a good flight training platform for the UH-60? [Yes/No]

Table 5. Number of Participants Classified per Category by Training Simulated Environment

Category	RCTD (*n* = 15)			VBS3 (*n* = 12)		
	Good	Neither	Bad	Good	Neither	Bad
Graphics	2	8	5	6	5	1
Controls	2	11	2	0	5	7
Responsiveness	2	11	2	0	2	10
Flight Representation	4	8	3	0	5	7
	Mission	Both	Game	Mission	Both	Game
Starting Focus	5	2	8	3	1	8
Progression Focus	10	4	1	2	5	5
	Yes		No	Yes		No
Training Platform	15		0	9		3

Pilot responses revealed trends in the data and supported findings from other measures. For example, there was a clear trend toward a focus on mission aspects of flight in RCTD pilots compared to pilots of the game-based systems. Additionally, the large number of responses classifying system aspects as "bad" correspond to some of the increases in frustration and workload seen in the survey and physiological measures. For example, one pilot responded, "Unfortunately the controls were being a little sluggish, so that was…flying in the VBS, I was more flying outside. I really didn't pay attention to the inside instruments at all". This pilot's response evidences a focus on aspects of the game and frustration with control input latency experienced in VBS3. Only by adopting an interdisciplinary ASSET approach to TEE has this fuller picture of training effectiveness ben found.

DISCUSSION

The goal for the present study was to determine if the low physical fidelity VBS3 games-for-training system could provide comparable levels of cognitive fidelity as the medium physical fidelity RCTD for a set of aviation collective training tasks. Study results indicated no differences in performance between the RCTD and VBS3, or in the OFT based on the preceding simulated training environment. A traditional TEE might take these differences alone and conclude that VBS3 can be an effective supplement to current aviation collective training. However, the interdisciplinary approach taken in this study provides a more complete picture of training effectiveness through the inclusion of psychological and other variables relevant to the performance context. Through the lens of ASSET, the RCTD was found to correspond more closely to the test environment in three primary areas: perceived presence, frustration, and perceived performance, while VBS3 was found to correspond more closely to the test environment with respect to no simulation sickness. Differences in these measures, with pertinence to the question of cognitive fidelity, suggest a few considerations before attempting to augment collective aviation training with game-based training.

Cognitive Fidelity Considerations

As noted earlier, cognitive fidelity concerns how well a simulated environment stimulates the same psychological mechanisms as the performance context and is associated with positive transfer of training (Kozlowski & DeShon, 2004). Differences in psychological variables should be addressed concerning their relative impact on transfer of training. A primary difference in cognitive fidelity between RCTD and VBS3 was the level of perceived presence. Five of seven presence factors were significantly higher in the RCTD than VBS3 (realism, possibility to act, quality of interface, possibility to examine, and haptic) and four of seven factors corresponded more closely with perceived presence in the OFT test environment (realism, possibility to act, quality of interface, self-evaluation of performance). It is likely that the difference in presence resulted from differences in physical fidelity between the two simulated training environments. VBS3, while having similar graphics and mission rehearsal capabilities, did not include the helmet-mounted display used in the RCTD, nor did it replicate the actual UH-60A/L controls as did the RCTD. Differences in presence are important, but as long as performance is comparable, this difference alone may not significantly affect aviation collective training effectiveness.

At the same time, simulation sickness was higher in the RCTD and there was a closer correspondence in simulation sickness between VBS3 and the OFT than RCTD and the OFT. Pilot interviews and researcher observations suggest this is likely due to the helmet-mounted display with transmissive lenses in the RCTD. Some participants described the helmet as heavy and mentioned eye strain during the simulation. Therefore, while the HMD helps increase one's sense of presence, it can also induce unwanted simulation sickness symptoms (not present in the test environment or actual performance context).

VBS3 and RCTD environments were highly similar with respect to performance, subjective (NASA-TLX) and objective (HRV) mental workload, physical workload, temporal workload, and effort. With the high number of similarities between RCTD and VBS3 training environments as a whole, the higher frustration in VBS3 suggests the potential need to increase familiarization time with game-based systems before training. The high frustration in VBS3 may have been driven more by familiarization than effectiveness since lack of familiarization can negatively impact learning factors such as workload. An increased familiarization time for more complex training systems might help alleviate frustration such as that observed in VBS3 in the present study.

It is important to note that while there was a difference in how pilots perceived their performance in the training simulated environment compared to the OFT between RCTD and VBS3 conditions, there were no significant differences in objective task performance between VBS3 and RCTD. One possible explanation for this finding could be that the performance metrics lacked the sensitivity required to reveal any true differences in performance stemming from differences in the two training environments. However, a more plausible explanation may be that people simply overestimate the association between presence and performance. That is, people may assume that higher levels of presence are necessary for task performance than are truly required for a given task or set of tasks. Research has been mixed on the relationship between presence and task performance (Pallamin & Bossard, 2016; Youngblut & Huie, 2003; Nash, Edwards, Thompson, & Barfield, 2000; Stevens & Kincaid, 2015), and it likely depends on the nature of the individual tasks and training scenarios under study. For the set of collective training tasks employed in the present study, it appears that a higher sense of presence is not required for equivalent performance. Increased familiarization time, as mentioned above, may help to close the gap between performance and self-perception of performance.

Phenomenological interviews, a unique component of the ASSET approach, provided further information as to the differences between VBS3 and RCTD. When asked to describe what it was like performing the mission in the VBS3 environment, pilots expressed frustration with the controls and the way in which controls responded. VBS3 pilots also expressed a greater focus on physical control of the aircraft ("game-focus") throughout the mission compared to RCTD pilots, who more often could shift to a focus on mission-related aspects ("mission-focus") when flight control became more automatic. In particular, 42% (5/12) of VBS3 pilots maintained a game-focused control throughout the mission scenario compared to only 7% (1/15) of RCTD pilots. A focus on controls and aircraft manipulation takes away from the main mission tasks required. Yet overall, VBS3 and RCTD pilots expressed a high degree of similarity in their responses toward using the simulated environments for training. Pilots in both simulated training environments expressed concerns over the use of these systems for procedural and instrumental flight training, while expressing their benefits for mission rehearsal and analysis, crew communication, collective training, multi-ship operations, and situational awareness and decision making training for Air Mission Commander or Pilot in Command training.

Augmenting Training

Overall, the results of this study provide support for enhancing the role of VBS3 as a platform for collective mission training and mission rehearsal training tasks. The higher frustration in VBS3 was linked to inaccurate controls (noted in pilot interviews) that introduced control differences from an actual UH-60A/L and a lack of familiarity with the training environment. Future research needs to consider different types of mission tasks and scenarios and more sensitive performance metrics, but there is evidence for the effectiveness of game-based systems such as VBS3 to support Army aviation collective training. It is important to note that overall, perceived performance and workload are reduced in the OFT following experience in VBS3 and this is not the case with the RCTD. Therefore, the VBS3 might be a good training aid for inducing stress and teaching coping skills. It is recommended that reliance on the RCTD for collective training can be reduced to augment training with cost-effective game-based systems. However, at this time, the use of VBS3 should be restricted to the types of tasks that do not rely heavily on physical manipulation of controls until control inputs better match the flight characteristics of the UH-60A/L.

Design Improvements

Lastly, the results and pilot interview responses in this study suggest a few areas to consider in the future design and improvement of simulated training environments for collective aviation training. First, improvements in game-based systems should focus on decreased control input latency and improving the flight characteristics by modeling input responses based on the actual physics of a UH60A/L aircraft. Second, results suggest that an ideal simulated training environment for UH-60A/L collective training might combine the collective and cyclic controls from the RCTD with the desktop application of game-based systems. However, due to the proprietary nature of VBS3 software, it is unclear if different controls would be able to successfully integrate with the system. It is also possible that additional controls may provide no further benefit to training effectiveness. All-in-all, the present study provides initial support for the use of game-based systems in augmenting aviation collective mission training tasks.

ACKNOWLEDGEMENTS

This paper was prepared in conjunction with work sponsored by the U.S. Army Program Executive Office for Simulation, Training, and Instrumentation (CA# W911NF-14-2-001). The U.S. Army Program Executive Office for Simulation, Training, and Instrumentation nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product, or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the U.S. Army Program Executive Office for Simulation, Training, and Instrumentation.

REFERENCES

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction, 15*(2), 40-47.

Bockelman, P., Reinerman-Jones, L., & Gallagher, S. (2013). Methodological lessons in neurophenomenology: Review of a baseline study and recommendations for research approaches. *Frontiers in Human Neuroscience, 7*(608), 1-9. doi: 10.3389/fnhum.2013.00608

de Winter, J. C. F., Wieringa, P. A., Dankelman, J., Mulder, M., van Paassen, M. M., & de Groot, S. (2007). Driving simulator fidelity and training effectiveness. In *Proceedings of the 26th European annual conference on human decision making and manual control*, Lyngby, Denmark. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.4015&rep=rep1&type=pdf

Goodwin, M. S., Reinerman-Jones, L., Goldiez, B. F., & Crapanzano, R. A. (2017). *An interdisciplinary approach to evaluating U.S. Army aviation training*. International Symposium on Aviation Psychology, 2017.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183.

Hochmitz, I., & Yuviler-Gavish, N. (2011). Physical fidelity versus cognitive fidelity training in procedural skills acquisition. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 5395), 489-501. doi: 10.1177/0018720811412777

JASP Team (2016). JASP (Version 0.8.0.1) [Computer software].

Jorna, P. G. A. M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics, 36*(9), 1043-1054. doi: 10.1080/00140139308967976

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

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of American Society for Training and Development, 11*, 1-13.

Kirkpatrick, D. L. (1976). Evaluation of training. In R.L. Craig (Ed.), *Training and development handbook: A guide to human resource development*. New York: McGraw Hill.

Kirkpatrick, D. L. (1994), *Evaluating Training Programs: the Four Levels*. San Francisco, CA: Berrett-Koehler.

Kozlowski, S. W., & DeShon, R. P. (2004). A psychological fidelity approach to simulation-based training: Theory, research and principles. In E. Salas, L. R. Elliott, S. G. Schflett, & M. D. Coovert (Eds.), *Scaled worlds: Development, validation, and applications* (pp. 75-99). Burlington, VT: Ashgate Publishing.

L-3 Link Simulation & Training (2010, September 20). Flight School XXI: Enabling Army aviators to hone their tactical skills. Retrieved from https://www.link.com/media/datasheets/FSXXI_2011.pdf

L-3 Link Simulation & Training (2012). Rotary-wing gallery. Retrieved from https://www.link.com/media/galleries/pages/rotary-wing.aspx

Liu, D., Macchiarella, N. D., and Vincenzi, D. A. (2008). Simulation fidelity. In D. A. Vincenzi, J. A. Wise, M. Mouloua, & P. A. Hancock (Eds.), *Human Factors in Simulation and Training* (pp. 61-73). Boca Raton, FL: CRC Press.

Morris, K. (Photographer). (2015, December 21). U.S. Army Aviation Center of Excellence uses VBS3 [digital image]. Retrieved from http://dogsofwarvu.com/forum/index.php?topic=1555.0

Mulder, L. J. M., de Waard, D., & Brookhuis, K. A. (2004). Estimating mental effort using heart rate and heart rate variability. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendrick (Eds.), Handbook of Human Factors and Ergonomics Methods (pp. 201-208). Boca Raton, FL: CRC Press.

Nash, E. B., Edwards, G. W., Thompson, J. A., & Barfield, W. (2000). A review of presence and performance in virtual environments. *International Journal of Human-Computer Interaction, 12*(1), 1-41.

Pallamin, N., & Bossard, C. (2016). Presence, behavioural realism and performances in driving simulation. *IFAC-PapersOnLine, 49(*19), 408-413

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences, 5*(3-4), 229-269. doi: 10.007/s11097-006-9022-2

Rockwell Collins. (2017). SimEye SX50TII Helmet Mounted Display. Retrieved from https://www.rockwellcollins.com/Products_and_Services/Defense/Simulation_and_Training/Products_and_Services/HMD-Simulation-Helmets/SimEye_SX50TII_Helmet_Mounted_Display.aspx

Sotomayor, T. M., & Proctor, M. D. (2009). Assessing combat medic knowledge and transfer effects resulting from alternative training treatments. *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 6*(3), 121-134. doi: 1177/1548512909350170

Stevens, J. A., & Kincaid, J. P. (2015). The relationship between presence and performance in virutal simulation training. *Open Journal of Modelling and Simulation, 3,* 41-48. doi: 10.4236/ojmsi.2015.32005

Tony. (2013, April 10). L-3 link awarded contract to deliver UH-60M operational flight trainers to Taiwan army. Retrieved from http://www.militarysystems-tech.com/articles/l-3-link-awarded-contract-deliver-uh-60m-operational-flight-trainers-taiwan-army

U.S. Army Aviation Center of Excellence. (2016, January). Army Aviation Training Strategy. Fort Rucker, AL.

Virtual Battlespace 3. (2017). Retrieved from https://bisimulations.com/virtual-battlespace-3

Whitney, S. J., Temby, P., & Stephens, A. (2014). A review of the effectiveness of game-based training for dismounted soldiers. *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 11*(4), 319-328. doi: 10.1177/1548512912472773

Witmer, B. G., Singer, M. J. (1994). Measuring presence in virtual environments (Technical Report no. 1014). Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA286183

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*(3), 225-240.

Youngblut, C., & Huie, O. (2003). The relationship between presence and performance in virtual environments: Results of a VERTS study. In *Proceedings of the IEEE Virtual Reality*, 277-278.