Can Synthetic Coaching Using an Immersive Training Device Effectively Train Student Pilots? A Field Study

Sandro Scielzo
CAE USA
Arlington, TX
sandro.scielzo@caemilusa.com

Gary Eves
CAE Australia
Brisbane, Australia
gary.eves@cae.com

Beth M. Hartzler
CAE USA
Arlington, TX
beth.hartzler@caemilusa.com

ABSTRACT

Developing and validating innovative solutions to train student pilots as effectively as experienced Instructor Pilots (IP) is a priority for many defense and civilian aviator training programs around the world to increase student throughput, minimize impact of IP shortages, and reduce overall training costs. Innovative training paradigms target the development of low-footprint, immersive simulators that maximize training task coverage and training effectiveness in self-paced environments when aptly paired with digital training solutions that can mimic the behaviors and evaluation heuristics of expert IPs. The current study investigated the utility of training using a Virtual Reality (VR) simulation-based training device paired with a next-generation synthetic IP providing real-time coaching, feedback, and scoring along with immersive and gamified debrief capabilities aimed to maintain student motivation and engagement. Thirty cadets from a large Indo-Pacific Asian Air Force participated in an hour-long training event practicing basic maneuvers across time. Difficulty was manipulated by alternating time of day. Maneuver performance was assessed automatically against syllabus-based criteria. Cadet workload was assessed via both NASA-TLX and the biometric-based objective Cognitive Workload Classifier (CWC). Pre and post surveys were administered to gauge cadets' confidence and overall training system perceptions. Results show significant training effects across time, along with a decrease in cognitive workload trend. Results are further discussed in terms of cadets' perceptions by experience and performance levels. A key finding shows a strong motivational effect for cadets when using the training system with synthetic coaching and feedback. This study is unique as it was field-tested in a germane operational pilot training environment and proves the viability of core aspects of next generation training solutions. Study findings are discussed to address overall strengths and limitations of next-generation pilot training solutions, as well as important consideration for integrating such systems within new and existing training courses.

ABOUT THE AUTHORS

Sandro Scielzo is a Human Systems Technical Authority and Learning Science Fellow at CAE USA. Dr. Scielzo received his PhD in Applied Experimental Human Factors and M.S. in Modeling & Simulation from the University of Central Florida in 2008 and 2005 respectively. His research has concentrated on the validation and implementation of next generation training solutions for military and commercial applications. Over the course of his career, Dr. Scielzo oversaw a wide portfolio of DoD R&D applied research projects to enhance warfighter training and readiness.

Gary Eves is a Principal Technology Officer & CAE Learning Science Fellow. Gary has led technology development and marketing of simulation solutions for over 30 years. In that time, he has worked developing and commercializing simulations in flight, driving, manufacturing, energy, entertainment, healthcare and now defense. Having developed simulations using computer game engines for customers, he completed a PhD on the validity of using game technology in simulation for the development of technical and non-technical skills. He lectures at UNSW on the Masters of Simulation and Interaction Design.

Beth M. Hartzler is a Senior Research Scientist in the Defense and Security division of CAE USA. Her doctorate is in Experimental Psychology, with particular focus on cognition as well as judgment and decision making. Beth has 10 years of experience working in the defense industry and currently works in conjunction with the Operational Learning Sciences Branch of the 711 Human Performance Wing/Air Force Research Laboratory. Her research experience includes the evaluation of performance and training efficacy, training requirements for trauma care, and aviation stressors.

Can Synthetic Coaching Using an Immersive Training Device Effectively Train Student Pilots? A Field Study

Sandro Scielzo
CAE USA
Arlington, TX
sandro.scielzo@caemilusa.com

Gary Eves
CAE Australia
Brisbane, Australia
gary.eves@cae.com

Beth M. Hartzler
CAE USA
Arlington, TX
beth.hartzler@caemilusa.com

INTRODUCTION

Next generation pilot training technologies are quickly saturating the market with a variety of seemingly disconnected solutions and capabilities touting superior training benefits when compared to traditional instructor-led and simulation-based training. However, what are the tangible training benefits that such technology can offer? How can they support existing courses and training programs? To answer these questions, we must first identify the main factors driving next-generation pilot training technologies, mainly: immersive training devices, adaptive training using synthetic instruction, and use of biometric data. Second, we empirically verify the overall training impact of combining these technologies, both from a skill acquisition standpoint and from an acceptance standpoint. To do so, we review the theoretical benefits and key research questions associated with each of these technology factors. Then we introduce the present study, aimed at testing some of these important questions to determine the extent to which next generation training technologies can indeed provide measurable and significant training benefits.

Immersive Training Devices

Immersive Training Devices (ITDs) are typically defined as medium-to-high fidelity simulators characterized by a low footprint when compared to high-end devices (e.g., full flight simulators). These devices often use Virtual Reality (VR) to render the cockpit and the visual environment, and incorporate haptic feedback, ranging from simple Hands on Throttle and Stick (HOTAS) to full cockpit replication. ITDs are designed to cover training needs to bridge the gap between basic training devices and traditional Operational Flight Trainers (OFTs) and Weapons System Trainers (WSTs) used in undergraduate flight training enterprises. ITDs are rapidly garnering popularity by being low-cost alternatives to high-fidelity training devices, while being able to cover a significant portion of a student pilot curriculum (e.g., Eves, 2007). Some key questions about ITDs are: can pilot training tasks be effectively taught using ITDs? How effective is the level of haptic feedback? Is the visual environment realistic to provide a positive training experience? This study addresses these important questions within the context of military undergraduate pilot training.

Adaptive Training Using Synthetic Instruction

Adaptive training is a powerful approach to learning that tailors the delivery of training materials to the unique needs and abilities of individual learners. In recent years, adaptive training has gained increasing attention as a promising tool for enhancing learning outcomes in a variety of settings, including education, workforce training, and military training (e.g., Schatz et al., 2015; Spain et al., 2012). Acting as a main facilitator to adaptive training, synthetic instruction—also known as virtual coaching or Artificial Intelligence (AI) coaching—espouses the characteristics of a human Instructor Pilot (IP) to deliver real-time verbal instructions, prompts, directions, and feedback to students. Additionally, a synthetic IP is typically designed to automatically score a student against established syllabus criteria and provide targeted recommendations to students (e.g., Azevedo & Bernard, 1995; Deaton et al., 2007; Guevarra et al., 2022). Some important questions regarding the use of synthetic IPs are: how effective is a synthetic IP? Are students engaged and motivated when interfacing with a synthetic IP? Can synthetic IPs accurately mimic the behaviors and scoring capabilities of human IPs? This study aims to start answering some of these important questions.

Use of Biometrics for Targeted Training

Endeavouring to measure warfighters' psychophysiological states and cognitive functions that are diagnostic of performance has been a major aim for the military, industry, and academia alike to develop better training systems which enhance readiness, and effective decision support tools that save lives and improve mission outcomes (see

Scielzo et al., 2020). Recent advances in biometric sensor capabilities allow for the relatively unobtrusive collection of a multitude of physiological indicators (e.g., electrodermal activity, heart rate, pupil dilation) that may be indicative of internal states, such as workload, stress, and fatigue. These capabilities are leading to a variety of real-time measurement frameworks such as workload assessment via pupillometry data (e.g., Rafiqi et al., 2014; Wangwiwattana et al., 2018), or using multiple physiological indicators to diagnose stress, engagement, and fine motor control (Wilson, 2018). Another promising venue is the use of Machine Learning (ML) classifiers that can provide diagnostic information based on multiple and simultaneous stream of biometric data, such as the Cognitive Workload Classifier (CWC), which was used in this study (see Wilson et al., 2020). Some important questions regarding the use of real-time biometric-based measures include: are workload visualizations useful, and to what extent do workload measures correlate with performance?

Training Ecosystems

ITDs, adaptive training with synthetic instruction, and use of biometrics in training are all important components of next-generation pilot training; however, in isolation they cannot provide a cohesive experience to the student. Furthermore, an essential component of training is driven by Instructional System Design (ISD) analyses, such as a Training Need Analysis (TNA), that determines optimal pairing of training devices based on task proficiency requirements and sensory needs (i.e., visual, audio, and haptic fidelity). In addition, ISD is also responsible for designing adaptive training logic by which a student can progress along a curriculum. As a result, a key concept espoused by industry at large is that of a training ecosystem that coherently marries interrelated training technologies with learning science principles to maximize training benefits and provide an engaging training experience.

Present Study

The current study is unique as it represents an opportunity to validate next generation capabilities in the field, using a population of military student pilots in a controlled experiment. Specifically, the main goal of the current study was to field test a small-footprint ITD, paired with a synthetic IP and a gamified interface to support debriefs, and to track Cadets' workload across maneuvers and time of day using an unobtrusive wrist-worn biometric device. This study concentrated on undergraduate pilot training for takeoff and landing on a turboprop training airframe. Syllabus-based criteria were ingested to automate scoring and coaching feedback. Automated feedback was provided in terms of pre-maneuver instructions and verbal prompts based on maneuver execution performance. Specifically, prompts were provided verbally when students' performance was fair or below, or when students violated safety of flight criteria. The thesis of the current study answers the following question: can pilot skills significantly improve using a next-generation training platform? We posit that yes, next generation training technologies using an ITD with synthetic coaching can successfully train critical pilot skills based on syllabus criteria. Table 1 lists our main study hypotheses.

Table 1. Study Hypotheses

Category	Hypotheses
Perceptions	- High experience Cadets will have more maneuver confidence than low experience Cadets - Cadets will have positive subjective perception of the training experience
Performance	- Cadets will improve performance over a 1-hour training event - Cadets will improve performance across time for both daytime and nighttime maneuvers
Individual Differences	- Both high and low experience Cadets will improve performance across trials - Both high and low performing Cadets will improve performance across trials
Workload	- Cognitive workload will be significantly higher during nighttime maneuvers - Real-time cognitive workload will be diagnostic of performance

METHOD

Study Design

The current study is a mixed model design with scenario type (takeoff, landing) as the between-subjects independent variable, and time of day (day, night) as the within-subjects independent variable. Each participant conducted six training events (Trials 1 through 6) on either takeoff or landing, alternating time of day.

Apparatus

The next-generation training platform used in this study consisted of a small footprint ITD (simulating a Grob 120TP turboprop airplane) with VR headset, a biometric wearable device, and a synthetic coach with gamified user interface showing real-time scoring, feedback, 3D flight path, and cognitive workload (see Figure 1).

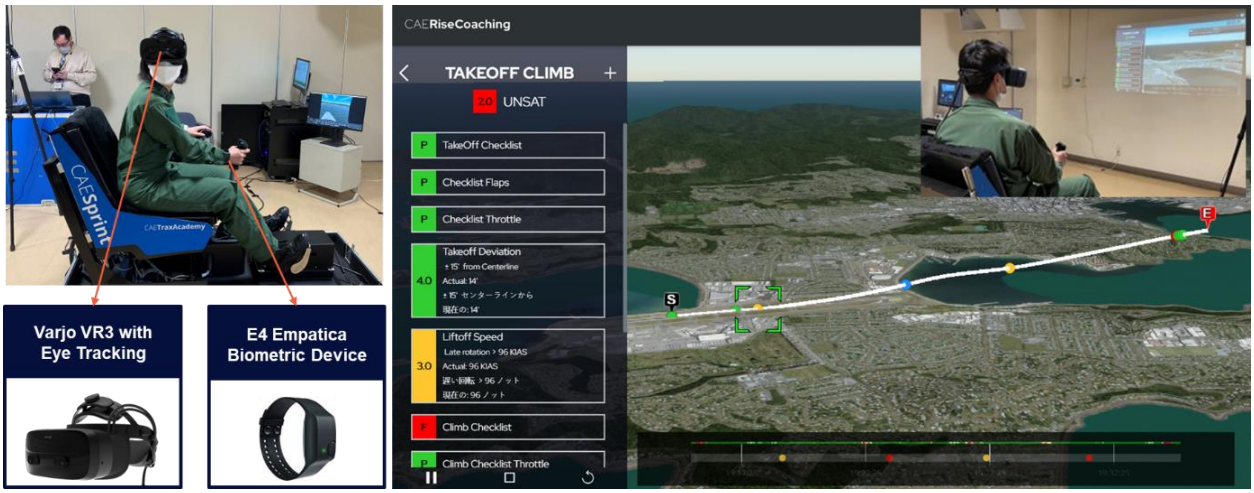

Figure 1. Small Footprint ITD with VR headsets, Wearables, and Real Time Interactive Scoring Interface

Measures

The current study utilized both objective and subjective measures to capture Cadet performance and perceptions. Objective measures include automated maneuver grading and scoring, and the CWC, which takes real-time multi-sensor data from a wearable device to assess mental workload. Subjective measures include training attitudes and post study surveys, the NASA Task Load Index (NASA-TLX), and Cadets' end-of-study comments. Each of these measures are reviewed, in turn.

Automated Maneuver Scoring

Automated grading was conducted based on a US Air Force Undergraduate Pilot Training (UPT) Maneuver Item File (MIF). MIF criteria for takeoff and landing were adapted for the Grob 120TP flight parameters by Subject Matter Expert (SME) IPs. The system used to automate scoring is based on the Adaptive Learning Environment (ALE), a proven performance assessment engine (see *Adaptive Learning Environment, 2023)*. Performance criteria were also used by the synthetic IP to provide coaching directions and prompts during a maneuver. Overall, each training event (takeoff or landing during daytime or nighttime) was automatically graded and visualized with feedback (score insights) on the interface. Figure 1 shows how each participant received an overall maneuver score (as demonstrated in Figure 2 in red, "2" = *Unsatisfactory* or "UNSAT") as well as various insights based on maneuver criteria (the overall maneuver failed because of a failed takeoff checklist). Thus, the overall score is not a simple aggregate, but based on MIF scoring logic.

Cognitive Workload Classifier

A real-time and objective biometric-based ML CWC provided continuous workload data (see Wilson et al., 2021). The CWC was used in this study to provide real-time student pilot workload data on the performance dashboard, above the scenario timeline (the thin color-coded line on top of the maneuver timeline in Figure 1). All participants wore a wrist-worn device to feed real-time biometric data to the workload classifier.

NASA Task Load Index

The NASA TLX was used as the gold standard measure for assessing subjective workload. NASA-TLX is a highly validated metric used for about half a century to provide accurate workload insights across several dimensions (see Hart, 2006). Specifically, NASA-TLX provides workload estimates for mental demand, physical demand, temporal demand, performance, frustration, and effort. Additionally, NASA-TLX provides an overall index of workload based on data aggregation across all dimensions.

Survey Questionnaires

A biographical survey gathered data on age, flying hours experience, and familiarity with VR systems. A pre-trial training attitudes survey addressed participants confidence in maneuvers and skill acquisition. Participants answered these survey items using a 5-point Likert-type interval scale, ranging from "*Strongly Disagree*" (1) to "*Strongly Agree*" (5). Example items include "*I am confident in my ability to perform the Takeoff maneuver*" and "*I am aware of how my skills are improving in my current training*." A post-trial survey was designed to capture participants impressions of their training experience. Participants answered survey items using the same 5-point Likert scale, and example survey items are "*The inflight AI instructions improved my ability to perform the Take Off maneuver in the VR simulator*" and "*The flight simulation analytics were helpful*." Finally, any post-trial comments and observations from participants were captured and coded. As an important note, all questions were professionally translated into Japanese to preserve content validity.

Participants & Procedures

Study participants were taken from the pool of student pilot Cadets undergoing basic flying training with a large Indo-Pacific Asian Air Force. Overall, 30 Cadets participated in the study. Each session consisted of a 1-hour block. Cadets were greeted and informed on the nature of the study and expectations for their participation, including an introduction to the device, followed by a trial conduct briefing. Participants then completed the biographical data and training attitudes surveys. Participants were then seated in the ITD, donning the VR headset and wrist-worn biometric device. A brief familiarization of the virtual environment and ancillary controls was provided equally to all participants, followed by up to 8 minutes of free play with the ITD to ensure familiarity. This introduction process took approximately 15 minutes to complete.

The six maneuver trials followed, with each trial including a post-trial live debrief mediated by a human IP to walk through the scoring of the trial and provide guidance for improvement on subsequent attempts. Specifically, at the beginning of each trial, participants listened to the synthetic IP's instruction for the maneuver to be performed. Once participants began the maneuver, the synthetic IP would provide verbal prompts based on participants' performance. At the completion of each trial, the synthetic IP would provide the participant with a prompt signaling that the maneuver was complete. All verbal prompts and instructions were in Japanese (using Google text-to-speech AI). NASA-TLX was then administered, followed by the post-trial debrief, mediated by a human IP using the performance dashboard (the scoring interface is shown in Figure 1). The next trial was then conducted, alternating time of day. All six trials took about 30 minutes to complete. After the last event, the post-trial survey was administered and any verbal feedback from the participants was recorded, which took about 5 minutes. The total experimental session did not exceed 60 minutes.

RESULTS

Survey Data

Pre-Training Survey Responses

In response to the pre-trial survey items, participants' self-ratings did indicate substantially greater confidence in their ability to complete the two maneuvers focal to this effort. Specifically, 51.7% indicated moderate to high confidence for their ability to complete the landing maneuver (M = 3.03), and 65.5% reported moderate to high confidence in their ability to perform the takeoff maneuver (M = 3.40). Using their reported number of flight hours, participants were categorized as having either low experience (range = 4 – 15 hours) or high experience (range = 20 – 65 hours). Using these categories, statistical analyses revealed a marginally significant difference between the groups such that those with more flight experience (M = 3.00) indicated greater confidence to perform aerobatic maneuvers than did those with less experience (M = 2.29; $t(26.35) = -2.00$, $p = .057$). Similar significant differences were also evident for their confidence for landing maneuvers ($t(25.74) = -2.48$,

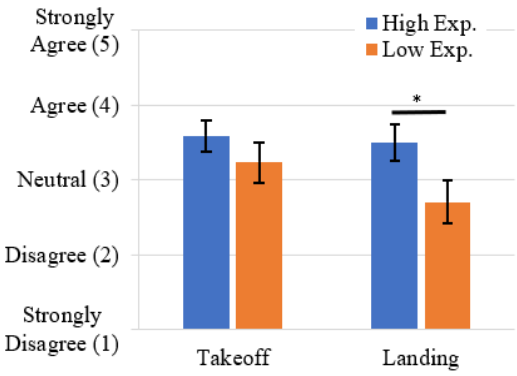

Figure 2. Maneuver Confidence by Experience Levels

$p < .05$), with high experience participants reporting greater levels of confidence ($M = 3.50$ and $M = 3.17$, respectively) compared to their peers with less experience ($M = 2.71$ and $M = 2.47$, respectively). A visual comparison of the significant and non-significant results is shown in Figure 2. For all figures, asterisks are shown to denote a significant mean difference (1 asterisk = $p < .05$, 2 asterisks = $p < .01$, 3 asterisks = $p < .001$).

Post-Training Survey Responses

Overall, the analysis of the post-training survey indicated that participants responded favorably across all survey items. Figure 3 shows these items, with green bars indicating items scoring above "4" on the 5-point Likert scale. Specifically, the quality of the AI instruction (i.e., synthetic coaching and feedback) and interface were highly rated. When asked about impact on performance, 76.7% agreed that the instruction would help them develop skills more rapidly ($M = 3.97$) and 80% agreed that it would improve their flight performance ($M = 4.13$). Likewise, most agreed that the AI instruction received during the training was clear (73.3%, $M = 3.93$) and helpful (86.7%, $M = 4.17$). Moreover, the trial analytics were beneficial in helping participants understand their performance (80.0%, $M = 4.27$) and monitor their skill progression (90.0%, $M = 4.20$). Most participants (93.3%, $M = 4.50$) also agreed that simulator graphics were exceptional. Participants were however less favorable in their ratings of the actual simulator, with only 43.3% agreeing that the controls were realistic ($M = 3.17$) and 60% who felt the performance was realistic ($M = 3.50$).

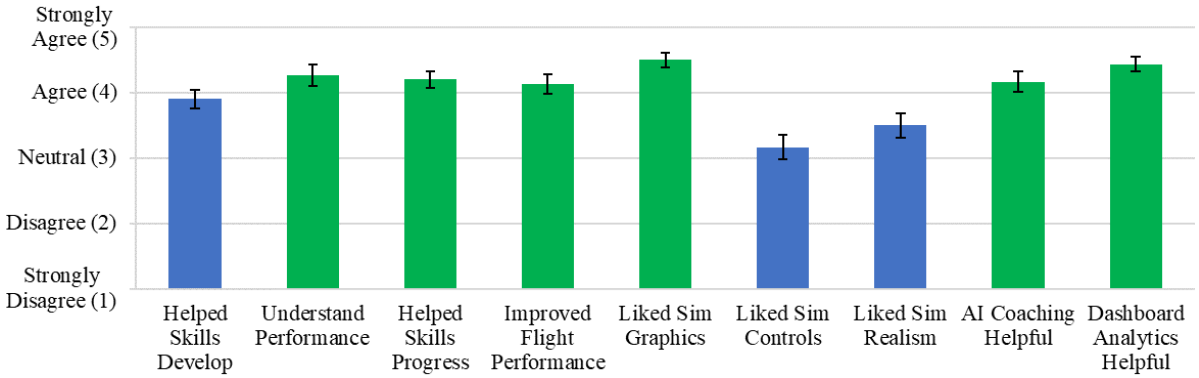

Figure 3. Cadet Post-Training Survey Responses Training Effects

Overall Training Effects

The next phase of analysis was to determine whether participants demonstrated any improvement on their performance over the course of the training experience (see Figure 4). Focusing first on the change in trial grades for daytime runs, participants' scores improved significantly from Trial 1 ($M = 2.90$) to Trial 5 ($M = 3.63$; $t(52.07) = -3.44$, $p < .01$).

A comparable improvement was evident for training runs conducted with the nighttime setting, increasing from a mean of 3.00 to 3.78 ($t(50.84) = -3.56$, $p < .001$). Separating the participants by their training maneuver condition revealed that this difference in scores was most evident among those who completed the takeoff and climb training, as their average scores improved from 2.73 to 3.69 ($t(24.92) = -4.12$, $p < .001$). Conversely, there was no significant improvement for those in the landing training condition during the daytime trial. Focusing instead on the nighttime trials did however reveal significant gains for both conditions. Between night Trials 2 and 6, mean grades for those in the takeoff condition increased from 3.07 to 3.86 ($t(26.89) = -2.57$, $p < .05$). Likewise, scores for participants completing the landing training improved by a similar margin (Ms = 2.93 and 3.69, respectively; $t(18.57) = -2.34$, $p < .05$). These data are presented in Figure 5.

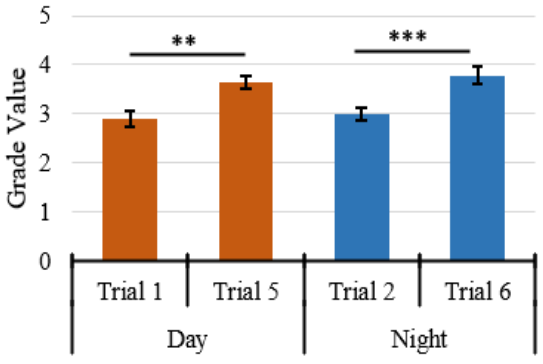

Figure 4. Overall Training Effects Across Trials During Daytime and Nighttime

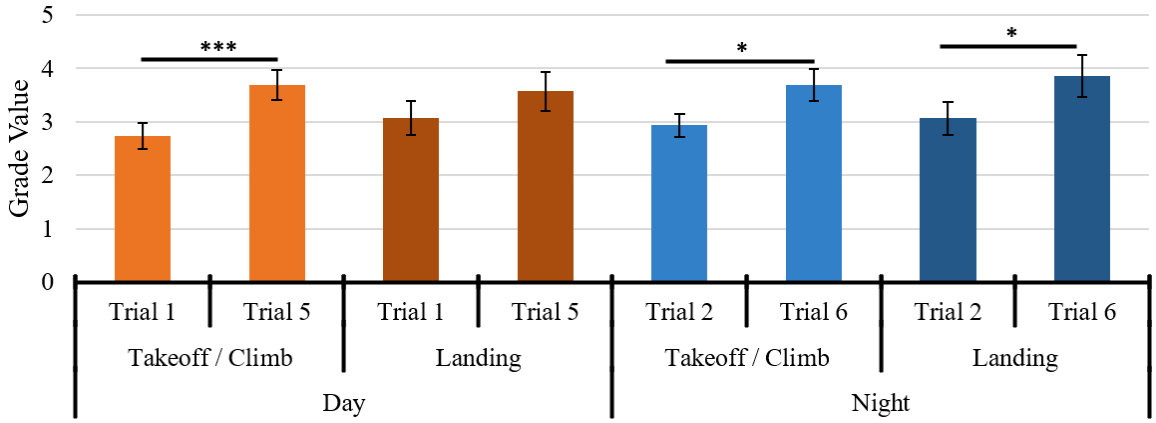

Figure 5. Grades for Day and Night Trials, by Training Condition

Learning Curves by Performance Level

Participants were categorized as either high performers (grade ≥ 4, "High Perf") or low performers (grade ≤ 3, "Low Perf") based on a median split. When looking at participants' score changes by performance levels, improvements were consistently most pronounced for low experience participants between their first and last trial. In particular, we looked at learning curves for specific maneuver subtasks. Figure 6 presents a subset of subtasks that are representative of overall results. Similarly, for the "climb check" takeoff subtask, low performing participants significantly improved their learning curves from their first trial (M = 1.00) to their last trial (M = 4.00; $t(8) = 1.86$, $p < .01$). A final comparison looks at the learning curve for low performers on the "liftoff speed" subtask, showing that low performing participants significantly improved their learning curves from their first trial (M = 2.00) to their last trial (M = 4.25; $t(5) = 2.01$, $p < .01$).

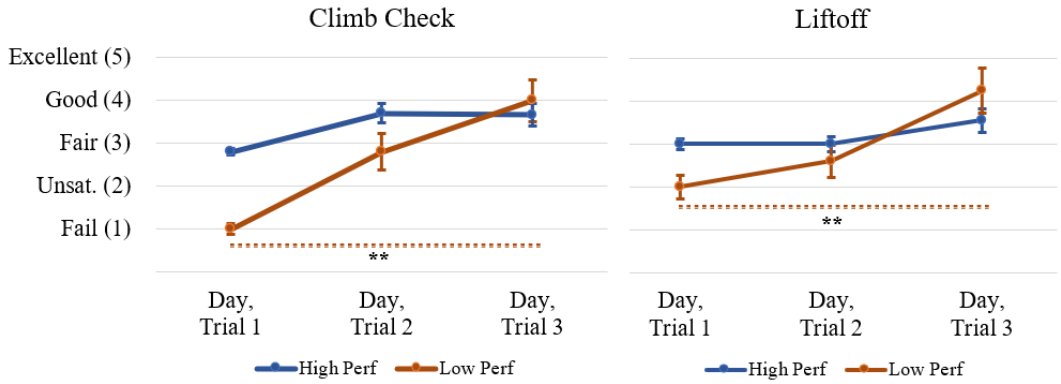

Figure 6. Learning Curves by Performance Level

Workload

NASA-TLX Survey Responses

Among participants' estimates of workload for the first and last daytime runs, only their responses for the physical demand and performance measures revealed any significant difference. Interestingly, participants rated Trial 5 as being significantly more physically demanding (M = 28.65) than what they had experienced during Trial 1 (M = 18.04; $t(49.42) = -2.15$, $p < .05$). Conversely, they rated their daytime performance as suffering significantly more during Trial 1 (M = 66.61) than was reported in Trial 5 (M = 46.35; $t(51.39) = 3.75$, $p < .001$). This same difference was further evident when comparing Trials 2 and 6 for the night-time runs, in that participants perceived their performance as suffering significantly more during Trial 2 (M = 60.93) than during Trial 6 (M = 42.04; $t(42.08) = 4.06$, $p < .001$).

Cognitive Workload Classifier

When correlating the CWC maximum and average workload values in a training event with performance across all maneuvers and conditions, as hypothesized CWC and performance were found to be significantly negatively correlated, but only for the maximum CWC values, $r(166) = -0.13$, $p = .05$. We then compared CWC maximum values

with performance across each trial, independent of maneuver type, showing statistical trends in the hypothesized direction for half the trials, while the other trials did not show any statistical difference. Additionally, we correlated the CWC with NASA-TLX to ascertain the level of convergent validity between the two cognitive load measures. We found that, although correlating in the correct direction, only one trial showed a marginally significant trend between the CWC and NASA-TLX (see Table 2).

Table 2. Correlations between CWC Maximum, Performance, and NASA-TLX

Measure	Performance		NASA-TLX	
	r	*p*	*r*	*p*
CWC max.				
Trial 1, day	.18	.17	.01	.97
Trial 2, night	-.24	.10(t)	.19	.17
Trial 3, day	-.30	.06(t)	.07	.35
Trial 4, night	-.26	.09(t)	-.01	.88
Trial 5, day	.01	.49	.26	.09(t)
Trial 6, night	.06	.38	.11	.29

(t) indicates statistical trend

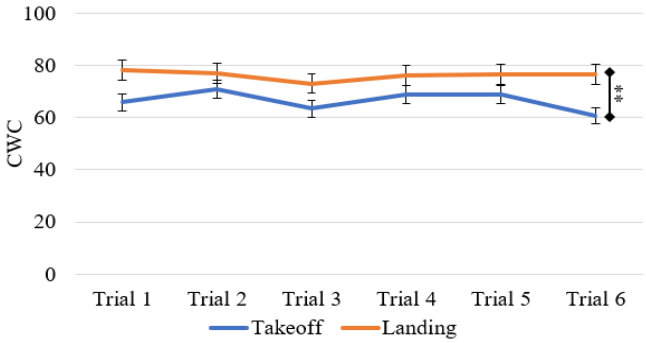

Figure 7. CWC Maximum Across Trials

When looking at CWC maximum values across maneuvers and trials, a clear pattern emerged indicating that participants consistently experienced higher workload during the landing when compared to takeoff (see Figure 7). This was not a hypothesized effect but is consistent with maneuver difficulty. Specifically, differences in objective workload were significant for Trial 6, with $t(25) = 2.22$, $p < .01$. Unfortunately, no significant results emerged when comparing workload across the time-of-day condition—as was initially hypothesized—indicating that the workload manipulation was not effective. However, referencing Figure 8, when comparing CWC maximum values across trials between high and low experienced participants, we did find a significant difference between the first trial ($M = 71.67$, $SE = 3.14$) and the last trial ($M = 61.89$, $SE = 3.14$), with $t(8) = 2.31$, $p < .01$. When further comparing between time of day, we found that although there were no differences in CWC scores between the first and last day trials, there was a significant difference—only for high experience participants—between the first night trial ($M = 72.78$, $SE = 5.07$) and the last night trial ($M = 61.89$, $SE = 5.07$), with $t(8) = 1.86$, $p < .05$.

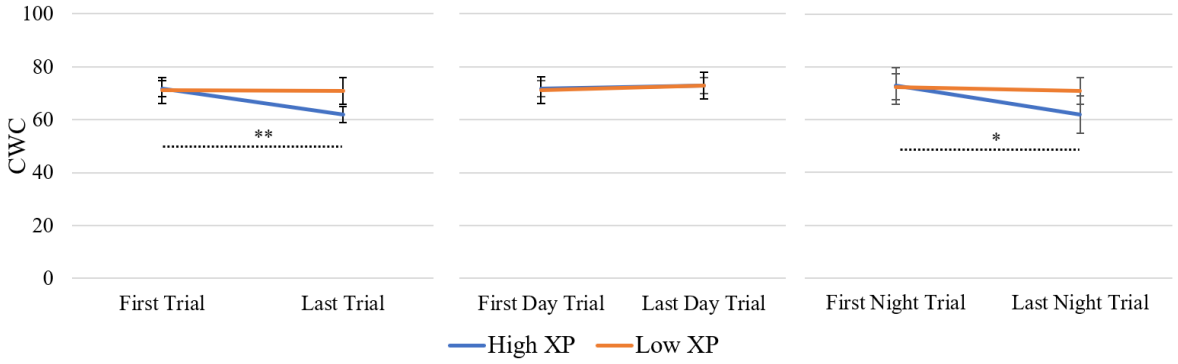

Figure 8. CWC Maximum for First and Last Trial by Participants' Experience ("XP") and Time of Day

Finally, when reviewing real-time CWC data within trials, we initially identified multiple instances of an increase in cognitive workload following a verbal cue from the synthetic coach and leading to a low score. Unfortunately, subsequent analyses determined that any effects were not sufficient to demonstrate statistical significance.

DISCUSSION

Summary of findings

This study verified many of our hypotheses and supported the overall thesis that using a small footprint VR training device with synthetic coaching can indeed successfully train pilot skills based on existing syllabus criteria (see Table 3). Overall, this study demonstrated—in a controlled, fielded experiment—that Cadets (a) exhibited positive

subjective perception of the training experience (across ITD, synthetic IP, and debriefing interface), and (b) these positive perceptions translated into significantly improved performance over a 1-hour training event.

Table 3. Hypotheses Testing Outcome

Category	Abridged Hypotheses	Outcome
Perceptions	- High experience Cadets will be more confident than low experience Cadets - Cadets will have positive subjective perception of the training experience	- Hypotheses met
Performance	- Cadets will improve performance over a 1-hour training event - Cadets will improve performance across daytime and nighttime maneuvers	- Hypotheses met
Individual Differences	- High and low experience Cadets will improve performance - High and low performing Cadets will improve performance	- Hypotheses partially met
Workload	- Cognitive workload will be higher during nighttime maneuvers - Real-time cognitive workload will be diagnostic of performance	- Hypotheses partially met

Cadets' Perceptions

In terms of Cadets' perceptions of the training experience, pre-training survey results confirmed expectations that cadets with less experience (in terms of flying hours) reported lower self-confidence in performing maneuvers than more experienced cadets. We posited that a simulation-based training capability with synthetic coaching could offer means to increase confidence through repeated practice without consuming the valuable time of a human IP. This effect was verified in two ways. First, post-training survey responses showed that Cadets agreed that the training platform allowed them to understand their performance, skill progression, and flight performance. More importantly, Cadets agreed that the synthetic coaching and performance dashboard were helpful. Second, we verified that these positive perceptions translated into significant training gains. Overall, it is appropriate to infer that next-generation training technologies that involve the use of synthetic coaching were well-received and conducive to significant training gains. This indicates the potential for support to the IP training resource allocation by automating elements of instruction and assessment. In fact, IP availability and resource allocation is a challenge for many training programs. Cadets' unprompted and voluntary self-report comments collected at the end of the experiment further reinforce the notion that the next-generation platform used in this study can alleviate the IP availability pain-point, such as "*I would use this in my own time*," "*love to practice with self-pace*," "*good visual and detail evaluation helped for good motivation* (sic)," and "*AI coaching is good*."

Performance and Individual Differences

When looking at performance gains over time, there were consistently positive results for training effect across multiple measures. Not only did we verify an overall significant training effect across all Cadets over the course of a 1-hour training event, but performance gains occurred across maneuver type and time of day. Simply said, Cadets improved by a full "letter grade" in just one training session. This finding is important because it illustrates a path towards a new form of self-paced training mediated by a synthetic coach. The main criticism of simulation-based, self-paced training is the risk for negative training, whereby student pilots acquire "bad habits" that are hard to break. The use of synthetic coaching—like the one is used in this study—provides similar safeguards against negative training as instructor-led training does. That is, by providing real-time verbal coaching and feedback as a student performs a maneuver, synthetic instruction can ensure proper skill acquisition. Additionally, by providing automated performance assessments and insights, Cadets can better understand their strengths and weaknesses, thus allowing them to focus on a personalized progression. This is the main benefit of adaptive training. That is, the ability to tailor the delivery of training materials to the unique needs and abilities of student pilots. An adaptive training solution needs to react to student pilots' idiosyncratic needs. This study showed that using synthetic coaching and feedback allowed for strong learning curves for low performing Cadets. The adaptivity in question centers on performance-based feedback, and post-maneuver insights. This is perhaps the most encouraging result we found as it proves that synthetic coaching can provide the same benefits to all student pilots, and particularly to lower aptitude ones. The implication is that such novel application of technology is well-suited for ab-initio and undergraduate pilot training programs, to reduce high-attrition levels that can occur for low performing individuals.

Workload Outcomes

Finally, this study investigated the usefulness of visualizing real-time cognitive workload on the debriefing interface. Here, the results are mixed. NASA-TLX was used as a gold-standard to provide convergence validity to the CWC. Unfortunately, we did not find significant relationships between the two workload measures. However, the relationship was in the correct direction and a statistical trend was observed for one of the six trials, which is

encouraging. Additionally, we were interested in using the real-time CWC as a diagnostic component of performance, which could be used in debriefing to highlight the impact of high cognitive workload on performance. Results showed that for the landing maneuver, which is more stressful than a takeoff, Cadets showed consistently higher cognitive workload with the objective CWC measure. Additionally, the CWC showed significant decrease in cognitive workload over time for high experience Cadets, but not for low experience ones. This may indicate that the CWC is not sensitive enough to capture changes in cognitive workload unless there is a salient effect. Finally, when looking at the CWC visualized on the debriefing interface, we did not find significant relationships between CWC and performance. Although, anecdotally, we saw many instances when an increase of CWC was observed right after a coaching prompt was given. Overall, we believe there is clear utility in visualizing real-time, objective cognitive workload to further support instruction. However, before such a measure can be truly useful, it needs to be much more sensitive and predictive of performance, at which point such real-time measure can be used by a synthetic coach to make the training experience even more adaptive to individual needs.

Limitations and Path Forward

The main limitation in this study was due to its nature: a field experiment without a control group that largely depended on Cadet availability. As a result, our controlled experiment only looked at performance gains over a 1-hour training event. However, we do believe that our sample-size was considerable, given the circumstances. Additionally, we only investigated training effects across two basic maneuvers, takeoff and landing. Future research should ascertain performance gains across several training events and across a larger subset of the student pilot curriculum.

When looking at Cadets' perceptions—although, as discussed, most perceptions were positive—a smaller subset of self-report comments indicated areas for improvement. Specifically, 23% of Cadets felt that the small-footprint ITD did not provide sufficient tactile or appropriate feedback. We believe this is mainly driven by the fact that the ITD used in the study did not have physical gears and flaps controls. These controls were virtually activated, and a recommendation for ITD developers is that any high-frequently used controls should be physical whenever possible. This falls into tradeoff considerations when developing VR-based ITDs that aim for a considerable lower cost point when compared to traditional simulation-based training devices. This study does provide strong justification for ensuring that basic controls do need to be physical in nature to avoid frustration, and even negative training. Additionally, 10% of Cadets also indicated that the automated grading system was harsh. Although this is a small percentage of Cadets, it is a good indication that there is always room for improving grading algorithms. The key point is that if we want synthetic instruction to be widely adopted, 10% of Cadets thinking the system scores too harshly is not acceptable. Future research needs to refine automated scoring algorithms and integrate scoring nuances that expert IPs often use.

Finally, part of this study was to verify real-time diagnostic power of the CWC. This ML classifier model was based on a population of fast jet pilots (see Wilson et al., 2021) and included accelerometer data to support cognitive load classification. It appears that the model might have not generalized well for this student pilot population, using a turbo prop as a training platform. This information points to another tradeoff when using biometric-based ML classifiers: generalizability vs, specificity. This study underlined the promise and utility of using a real time objective indicator of cognitive workload, but also suggest the importance of developing use-case specific models.

CONCLUSION

This study was unique as it represents a concrete first step in validating next generation training technologies. Specifically, we investigated the utility of using an immersive small-footprint VR simulation device paired with a digital solution providing both synthetic coaching and real-time performance dashboard. Moreover, this investigation of advanced training technologies was tested in the field with Cadets as participants. As a result, results from this investigation are applicable to any ab-initio and undergraduate pilot training military program.

An important takeaway is that the simulation and coaching environment achieved objectives for training by showing, face, content, and construct validity. That is, the overall training experience was able to reproduce flight manoeuvres that provide consistent performance data and was sensitive to cadet skill level. Rapidly maturing future training technologies supporting self-paced training with synthetic instruction represents a clear and present need due to IP shortages and increased student throughout requirements. This study provides compelling data supporting the use of

self-paced training using lower cost, small-footprint ITDs when paired with scientifically validated digital training solutions—mostly in the form of synthetic instruction that motivates student pilots and guaranties skill acquisition.

ACKNOWLEDGEMENTS

We want to acknowledge the Japan Air Self-Defense Force (JASDF) for providing this unique opportunity to test next generation training technologies. A personal Thanks to Col. Nomoto for his engagement and commitment to support the data collection effort and making cadets available for the study. We also would like to thank Col. Kitamura, Commander, Hofu-Kita Air Base and Col. Ishida, Commander Flight Training Group for their support in coordinating the study participants and use of the base facilities.

REFERENCES

Adaptive Learning Environment. (2023). CAE Defense & Security. https://www.cae.com/defense-security/what-we-do/training-systems/adaptive-learning-engine-ale

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111-127.

Deaton, J. E., Bell, B., Fowlkes, J., Bowers, C., Jentsch, F., & Bell, M. A. (2007). Enhancing team training and performance with automated performance assessment tools. *The International Journal of Aviation Psychology, 17*(4), 317-331.

Eves, G. (2007). Virtual Reality Based Training for Industry, a Cognitive Process Approach. *VR Solutions, Australia*.

Guevarra, M., Das, S., Wayllace, C., Epp, C. D., Taylor, M. E., & Tay, A. (2022). Augmenting Flight Training with AI to Efficiently Train Pilots. *arXiv preprint arXiv:2210.06683*.

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage publications.

Rafiqi, S., Nair, S., & Fernandez, E. (2014, May). Cognitive and context-aware applications. In *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 1-7).

Schatz, S., Fautua, D., Stodd, J., & Reitz, E. (2015, November). The changing face of military learning. In *Proceedings of the I/ITSEC*.

Scielzo, S., Wilson, J. C., & Larson, E. C. (2020, November). Towards the development of an automated, real-time, objective measure of situation awareness for pilots. In *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL*.

Spain, R. D., Priest, H. A., & Murphy, J. S. (2012). Current trends in adaptive training with military applications: An introduction. *Military Psychology, 24*(2), 87-95.

Walcutt, J. J., & Schatz, S. (2019). Modernizing Learning: Building the Future Learning Ecosystem. *Advanced Distributed Learning Initiative*.

Wangwiwattana, C., Ding, X., & Larson, E. C. (2018). Pupilnet, measuring task evoked pupillary response using commodity RGB tablet cameras: Comparison to mobile, infrared gaze trackers for inferring cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(4), 1-26

Wilson, J. C., Nair, S., Scielzo, S., & Larson, E. C. (2021). Objective measures of cognitive load using deep multi-modal learning: A use-case in aviation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5*(1), 1-35.