Operational Assessment of a CV-22 Virtual Maintenance Training Solution

Beth M. Hartzler, PhD
Defence & Security, CAE USA
Arlington, TX
Beth.Hartzler@caemilusa.com

Winston "Wink" Bennett, PhD
711 Human Performance Wing, AFRL
Dayton, OH
Winston.Bennett@us.af.mil

ABSTRACT

The CV-22 Osprey fills an important need for increasing the United States Air Force's (USAF) reach into remote and contested environments, but its unique design and capability profile leads to substantial operation and maintenance costs. Moreover, such a complex platform requires highly-skilled specialists to ensure optimal support of operational tempo, but limited fleet size translates to significantly constrained resources for training maintainers. Maintainers' opportunities for hands-on training is further complicated by varying aircraft availability, task demands, and operational tempo. To supplement traditional training opportunities, Link developed the Immersive Maintenance Guide (IMG) as an adjunct training tool that promotes knowledge and skill development among new maintainers. The IMG allows 3-Level maintainers to virtually and independently "walk" through the steps involved in a broad selection of maintenance tasks. A collaboration between the Rapid Sustainment Office (RSO) and the Air Force Research Lab (AFRL) evaluated the IMG using both objective and subjective measures to quantify the holistic benefits of the system over a 5-month evaluation period in comparison to maintainers who used only the traditional training resources. A comparison of the two groups revealed substantial improvements for both Experimental and Control participants, though advances over baseline were highest among those who used the IMG. This was evident for participants' knowledge of the task steps, such that Experimental participants' demonstrated a 42% greater improvement over their peers, and for their self-efficacy to lead completion of the task. Moreover, these gains were significant for both simple and routine tasks and those less commonly encountered. As a demonstration of training benefits, the current assessment revealed the IMG's promise as a valuable tool for improved task knowledge and an increased sense of readiness for task completion, gains which in turn likely contribute to increased combat power.

ABOUT THE AUTHORS

Beth M. Hartzler is a Senior Research Scientist in the Defence 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 nearly 10 years of experience working in the defense industry, first as an Aeromedical Research Psychologist for the US Navy and now in conjunction with the Continuous Learning Branch of the 711 Human Performance Wing/Air Force Research Laboratory. Beth's research experience is diverse, including the design and administration of measures to evaluate performance and training efficacy, training requirements for combat casualty care, spatial disorientation, and physiologic stressors such as fatigue and hypoxia in military and civil aviation.

Winston "Wink" Bennett, Jr received his Ph.D. in Industrial Organizational Psychology from Texas A&M University in 1995. He is currently the Readiness Product Line Lead for the Airman Systems Directorate located at Wright Patterson AFB Ohio. He is a Fellow of the Society for Industrial and Organizational Psychology, the Association for Psychological Science, and Air Force Research Laboratory Research. He has also been involved in I/ITSEC committee and program work for a number of years. He leads S&T supporting the Combat Air Forces migration to proficiency-based training and is conducting research related to the integration of live and virtual training and performance environments to improve mission readiness and job proficiency. He maintains an active presence in the international research and practice community through his work on various professional committees and his contributions in professional journals and forums including I/ITSEC. His involvement with the larger psychological communities of interest ensures that communication amongst international military, industry and academic researchers remains consistent and of the highest quality.

Operational Assessment of a CV-22 Virtual Maintenance Training Solution

Beth M. Hartzler, PhD
Defence & Security, CAE USA
Arlington, TX
Beth.Hartzler@caemilusa.com

Winston "Wink" Bennett, PhD
711 Human Performance Wing, AFRL
Dayton, OH
Winston.Bennett@us.af.mil

CV-22 AS A PRECISION INSTRUMENT

The CV-22 Osprey a tilt-rotor aircraft capable of vertical takeoff and landing, often used to conduct penetration of hostile territory to support special operations and unconventional warfare forces. Equipped for airspeeds twice those of the HH-60G (Losacker, 2017) to maximize the periods of darkness, the Osprey has served to extend Warfighter reach and capability in contested environments for nearly 15 years. Commensurate with this utility however, the CV-22 is also disproportionately expensive to operate. The cost is largely driven by the maintenance required to repair and sustain the complex tilt-rotor and hydraulic systems (Bolkcom, 2007), due in part to the absence of the extensive maintainer training programs such as those afforded to maintainers of more widely used aircraft, such as the AC-130. Consequently, new Osprey maintainers typically leave the schoolhouse with little more than familiarization experience, receiving only didactic training and minimal part-task trainer (PTT) experience. Formal training is expanded through the maintenance quality training program (MQTP) at their assigned unit, but their primary learning comes through on-the-job training (OJT). Though critical for the development of operational skills and knowledge, an analysis conducted by the RAND Corporation among USAF maintainers indicated that costs associated with OJT are tremendous, both direct (e.g., manning considerations) and indirect (e.g., errors attributed to distraction and inexperience; Manacapilli et al., 2007).

Additionally, new maintainers often rely heavily on external materials, such as electronic Technical Orders (eTOs) which include step-by-step guidance on completion of maintenance tasks. However the language used in these guides may be unclear, the source data may be out of date, and the explanatory pictures are of such low fidelity as to offer little in the way of a useful training aid. Consequently, novice CV-22 maintainers are generally reliant on shadowing opportunities to advance their knowledge and skills. To address this training gap, the Air Force Rapid Sustainment Office (RSO) provided the L3 Harris Link Immersive Maintenance Guide (IMG) Virtual Maintenance Training solution capability. The IMG was tailored to the CV-22 aircraft, equipping maintainers with the ability to walk through Technical Order Tasks in a high-fidelity, virtual environment. The IMG allows maintainers to train independently while providing a more visually representative reference tool through the Crawl, Walk, Run approach to learning (Goldberg et al., 2017). In other words, the IMG serves as a basic training tool to introduce new maintainers to tasks ("crawl"), includes an evaluation and testing environment to determine knowledge progress ("walk"), and an in-field reference of task steps ("run").

METHODS AND PROCEDURES

A longitudinal, operational assessment of the IMG was conducted to identify gains in knowledge, task proficiency, and self-efficacy associated with its use as a training adjunct. It was hypothesized that participants who used the IMG would demonstrate greater proficiency in completing hands-on tasks, decreased time-on-task, improved knowledge of task steps, and increased confidence about their readiness to lead completion of the task. This study employed a mixed-factorial design, in which novice CV-22 maintainers were recruited to participate either during the Control or Experimental phases, and were tested three times over the course of their 5 month-participation in the study. A sample of 22 participants were recruited and enrolled in this study, with 19 ultimately completing the entire study. Three participants withdrew from the study, one from the Control condition and the other two from the Experimental condition; their data are excluded here. All data were collected electronically, were password-protected, and the responses and results from individual participants were only available to the authors.

All participants were active-duty CV-22 maintainers assigned to the 58th Air Force Maintenance Squadron (AMXS) located at Kirtland AFB. Maintainers were eligible to participate if they were a 3-Level and had been with the 58th

AMXS for less than one year. Participants completed the informed consent process over the phone with the researcher, and were given the opportunity to ask any questions or decline participation.

Ages ranged from 18 to 26 years (M = 20.47, SD = 2.22) and current ranks were either E-2 (n = 4) or E-3 (n = 15). Most maintainers in the sample were Crew Chiefs (2A5x2D, n = 12), while the remainder specialized in either Integrated Communication/Navigation/Mission Systems (2A2x1a, n = 6) or Electrical and Environmental Systems (2A6x6, n = 1). Most reported having served less than a year prior to enrolling in the study (M = 325.16, SD = 91.02). On average, maintainers in this sample had graduated from the School House approximately four months prior to enrolling in the study (M = 122.26 days, SD = 96.45) and had been with the 58th AMXS for roughly 3.5 months (M = 109.95 days, SD = 95.03). Only two-thirds of the participants reported having any hands-on experience with one or more of the 10 tasks included in this evaluation. Preliminary analyses confirmed the Experimental and Control groups did not differ with regard to age (p = .091), or the number of days since they had enlisted (p = .60), graduated from the School House (p = .39), or joined the 58th AMXS (p = .31). Further, chi-square analyses confirmed there were no meaningful differences for the number of times participants in either group had participated in any of the 10 tasks (all p's > .30).

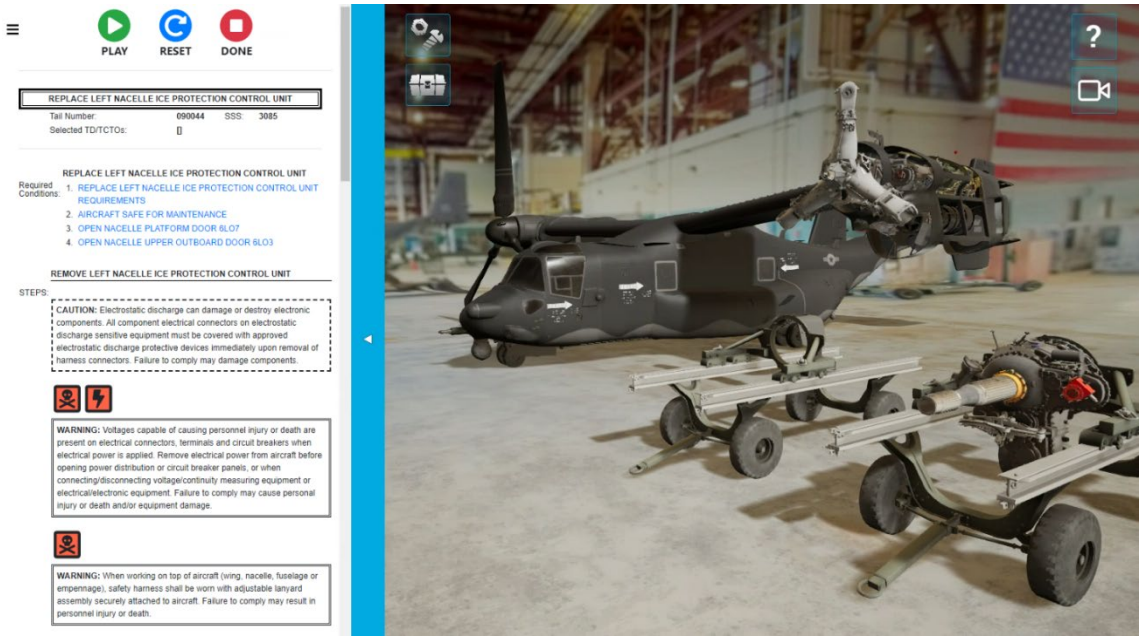

Figure 1. Screenshot of the IMG for the NIPCU Task in Guided Training Mode.

Measures

The IMG is a fully-integrated maintenance solution adapted for use by CV-22 maintainers, both as a training adjunct for those learning new skills and as a reference for hands-on task completion. The IMG may be used in three different modes:

1. Crawl – demonstration mode; student observes task completion, step-by-step with no interaction
2. Walk – training mode (Figure 1); student performs the steps, as directed by the tech manual, and using the virtual hardware, with remediation that highlights incorrect actions and guides corrective actions
3. Run – testing mode; student completes task without remediation except for safety errors
    a. If safety error is noted, the Learning Management System (LMS) notifies the instructor
    b. Instructor then determines whether to terminate the exercise or to allow the student to continue

Background data recorded by the LMS include the task attempted, completion status of each step, and the time necessary to complete the steps. The IMG also addresses several key training concepts, which may be difficult to achieve in traditional didactic learning environments: 1) centralized training control; 2) student-paced instruction; 3) immediate learning feedback; 4) instructor-training control, and 5) opportunity for demonstrated mastery.

In the current assessment, 10 maintenance tasks for the left nacelle were identified by senior leadership within the 58th AMXS. The selected tasks included a range of complexity and frequency, such as those that are considered straight-forward and are completed routinely (e.g, pitch link adjustment, removal and replacement of the nacelle interface unit, NIU) to those that are seldom performed and require multiple shifts to complete (removal and replacement of the prop-rotor gearbox). The complete list of tasks is shown in Table 1, with the number of steps involved in the task as written and the number of virtual steps included in the IMG. The number of steps differs due to some of the work being external to the nacelle, such as setting up the installation sling for larger components or disconnecting fixtures in other areas. Unlike the other tasks, the removal and replacement phases of the variable frequency generator (VFG) are detailed across two separate eTOs, and this same division is used for the IMG, for a total of 11 tasks on the system. In all other contexts, removal and replacement of the VFG is described as one task.

Table 1. Summary of included maintenance tasks with details regarding participants' prior experience.

Task Name	Activity	Total Task Steps	IMG Task Steps	Participants with Prior Task Experience			
				Control Phase		Experimental Phase	
Heat exchanger assembly	Remove / Replace	121	60	2	20.0%	1	11.1%
Nacelle ice protection control unit (NIPCU)	Remove / Replace	36	27	2	20.0%	4	44.4%
Nacelle interface unit (NIU), No. 1	Remove / Replace	28	13	1	10.0%	2	22.2%
Pitch link	Adjustment	25	8	3	30.0%	1	11.1%
Proprotor gearbox (PRGB) assembly	Remove / Replace	502	198	2	20.0%	1	11.1%
Pylon driveshaft assembly, L6	Remove / Replace	53	17	4	40.0%	3	33.3%
Slipring standpipe assembly	Remove / Replace	91	58	1	10.0%	2	22.2%
Tilt axis gearbox (TAGB) assembly	Remove / Replace	199	27	2	20.0%	1	11.1%
Variable frequency generator (VFG), No. 3	Remove	38	36	4	20.0%	4	44.4%
	Replace	47	40				
Wiring integration assembly WA57	Remove / Replace	52	22	4	40.0%	4	44.4%

**Source TOs for the VFG treat removal and replacement as separate tasks due to the different procedures between blocks; same distinction is represented in the IMG.*

Knowledge Evaluation. To measure changes in participants' knowledge of the steps for each task, research personnel developed a 40-item, multiple choice quiz. Questions included several items concerning general maintenance practice (e.g., "*Positive and safe clearance can be defined as having enough___.*"), and two to five questions on each of the 10 selected tasks (e.g., "*Performing maintenance on the NIPCU should include use of ___ protective devices to guard against potentially harmful ___ discharge.*"). All items were reviewed and approved by 7- or 9-Level maintainers to ensure each item was appropriate and correct. Participants were instructed to complete the measure independently, and to not use other resources (e.g., peers, eTOs, etc.) to answer the questions, and to do as best as they could.

Self-Efficacy. Researchers further hypothesized that the opportunity to complete each task virtually would contribute to changes in self-efficacy, or the confidence to complete a particular action. Available evidence indicates that an individual's sense of self-efficacy plays a pivotal role in his or her attention to task, response accuracy, and resilience to setbacks (Themanson & Rosen, 2015). After the knowledge evaluation, participants were asked to review a list of 12 items and for each indicate their agreement with the statement "*I am confident in my ability to lead the completion of this task.*" The list included the 10 tasks plus two questions on general maintenance practices, and responses were made using a 6-point Likert scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (6).

Task Proficiency and Time-on-Task. The research protocol also included measures to assess participants' proficiency during hands-on completion of the 10 target maintenance tasks. Performance was assessed by a senior maintainer

familiar with the task, and measured as a "Go" or "No-go" rating for each step to indicate whether the participant completed it satisfactorily. The senior maintainer also rated completion of each step on: 1) readiness to complete the step independently; 2) need for assistance; 3) errors of omission; 4) errors of commission; and 5) whether appropriate safety procedures were followed. An automatic timer recorded the time a step was rated, capturing how long it took for participants to complete each step as well as the total time-on-task. However, due to the inconsistent nature of aircraft maintenance and frequency with which participants could have hands-on experience, the selected tasks were not completed often enough to justify this as a reliable measure. Consequently, the data are not included here.

Procedures

Participants were recruited through flyer postings and word-of-mouth. Interested maintainers were invited to contact the researcher who completed the informed consent process over the phone. This included answering all questions, and ensuring participants understood the expectations of the study and that they were participating voluntarily. After consenting to join the study, participants were instructed to complete the initial knowledge evaluation form and self-efficacy questionnaire. Throughout the 5 months of the data collection phase, participants carried on with their usual duty assignments, completing all expected OJT. Ten weeks after enrolling in the study and completing the initial measures, participants were instructed to complete the mid-point knowledge evaluation and self-efficacy questionnaires. These same measures were completed a third time 10 weeks later at the end of their evaluation period.

The only difference between the Control and Experimental phases is that the latter had access to the IMG during their 5-month evaluation period. After completing the initial knowledge evaluation and self-efficacy measures, participants in the Experimental group completed each of the 10 tasks on the IMG using the Guided Training mode. They were further encouraged to use the IMG as a review tool as preparation before working on any of the 10 selected tasks. By the study mid-point, Experimental phase participants had completed each task at least once on the IMG. Following the mid-point evaluation, participants were asked to complete each task at least once more in Evaluation mode before the final evaluation and end of the study.

RESULTS

The primary focus of these analyses was the difference between the Control and Experimental groups for the change in their knowledge evaluation scores and self-efficacy ratings over the course of the 5-month evaluation period. Data from the knowledge evaluation were analyzed both as the participants' total percent correct for a given assessment and as the percent correct for each task or topic area. For the self-efficacy measure, participants' categorical responses were coded ordinally and the numeric equivalents were used for analysis, with summaries represented as numeric means within the body of the paper. Within-subjects differences were evaluated by group through Friedman tests, followed by Nemenyi test for post-hoc comparisons where appropriate, while differences between groups were analyzed using the Wilcoxon/Mann-Whitney test. To identify potential interactions between study phase and condition, the final type of analysis utilized the F1-LD-F1 model in the R package "*nparLD*", reported as a Wald statistic and developed as a non-parametric alternative to the two-way repeated measures ANOVA (Noguchi et al., 2012). To control for family-wise errors, Holm-Bonferoni corrections were calculated where appropriate. However, owing to both the small sample size and a preference to avoid Type II errors, all statistically significant results are presented here, though additional notes (e.g., HB-corrected *p*) indicate when a difference attained significance using the modified criteria.

Knowledge Evaluation Scores

For the total percent correct on the knowledge evaluation, the two groups started with equivalent scores (Experimental $M = 48.61\%$, $SD = 0.12$; Control $M = 48.25\%$, $SD = 0.11$; Figure 2), and over the course of their respective evaluation periods both improved significantly from the initial evaluation. Specifically, total scores for the Experimental group increased by 22.86% from the initial to the mid-point evaluation, with an average of 59.72% ($SD = 0.07$) 10 weeks into the study. Moreover, these participants had a total gain of 53.71% over the course of the study, scoring an average 74.72% ($SD = 0.07$) at the final evaluation ($Q(2, N = 9) = 17.00$, HB-corrected $p < .01$), and average scores for the individual topic areas increased by 19 to 120% over the study period. Post-hoc Nemenyi test results further indicated significant differences between the initial and final evaluation time points ($p < .001$). Conversely, total scores for the Control group increased by a smaller margin, gaining 12.95% between the initial and mid-point evaluations ($M = 54.50\%$, $SD = 0.09$), and 37.82% over the entire study for a final average of 66.50% ($SD = 0.09$; $Q(2, N = 10) = 13.28$,

HB-corrected $p < .01$). Additionally, post-hoc comparisons did reveal a significant difference between the initial and final measures ($p < .01$). Among participants in the Control group, the average changes in total scores across tasks ranged from -42.9% to +170%.

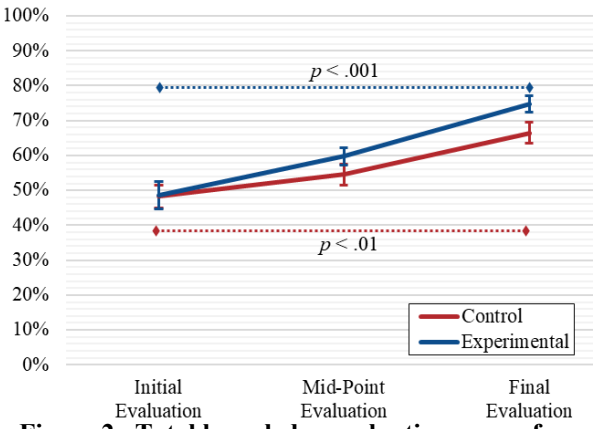

Figure 2. Total knowledge evaluation scores for Experimental and Control groups.

Proprotor Gearbox Assembly (PRGB), Figure 3B. Analysis of the knowledge evaluation scores for items related to the PRGB removal and replacement indicated a significant change for participants in the Experimental group ($Q(2, N = 9) = 7.66$, $p < .05$). Specifically, their average scores increased by 113% over the course of the study, from 22.22% ($SD = 0.20$) at the initial evaluation to 47.22% ($SD = 0.20$) at the final measure. This result is contrasted by the 30% decrease in knowledge scores among Control participants, whose average score deteriorated from 25.00% ($SD = 0.20$) initially to 17.50% ($SD = 0.17$) by the end of the study. The difference in performance trajectories further contributed to a significant time/group interaction (Wald $\chi^2(2, N = 19) = 6.67$, $p < .05$), as well as a significant difference between the two groups for their scores on the PRGB items on the final knowledge assessment ($W = 12.5$, $p < .01$).

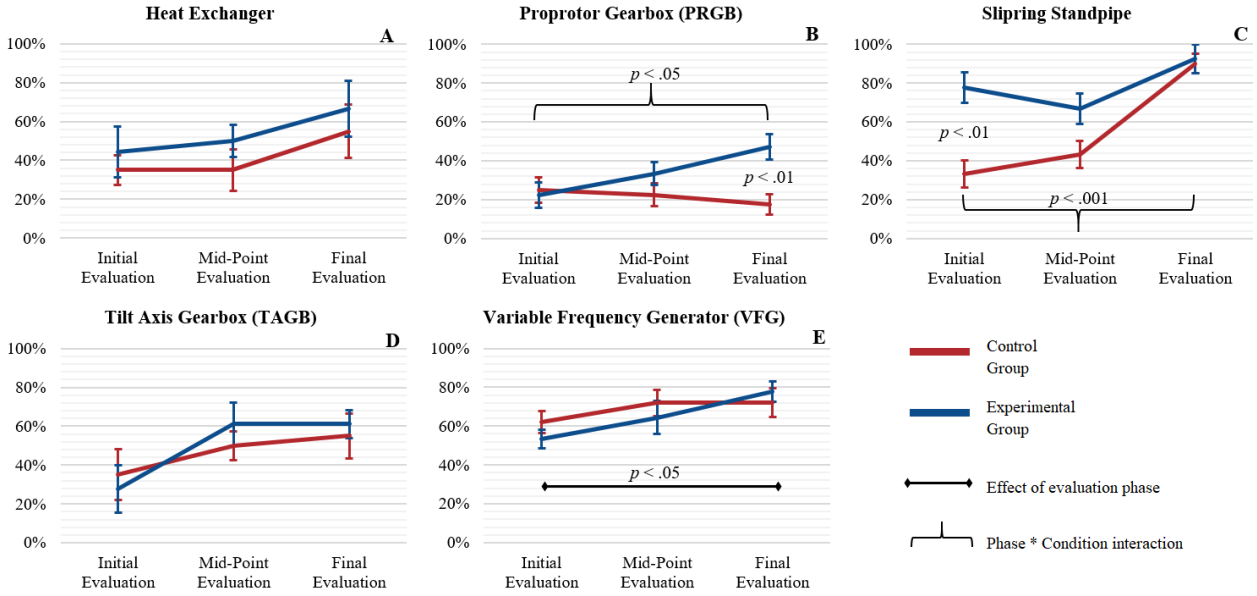

Figure 3. Knowledge evaluation scores for: A) heat exchanger; B) proprotor gearbox; C) slipring standpipe; D) tilt axis gearbox; and E) variable frequency generator.

Slipring Standpipe, Figure 3C. Participants in both groups demonstrated significant gains in knowledge for the slipring standpipe task, though the improvements were highest for participants in the Control group. In particular, scores at the initial assessment were significantly greater for Experimental participants compared to Control participants ($W = 71.5$, $p < .05$), though this difference was eliminated by the final evaluation. Moreover, the Experimental group demonstrated a 19% improvement in scores over the course of the study, increasing from 77.78% ($SD = 0.24$) to 92.59% ($SD = 0.22$; $Q(2, N = 9) = 6.09$, $p < .05$), though the improvement among Control participants from 33.33% ($SD = 0.22$) to 90.00% ($SD = 0.16$; $Q(2, N = 10) = 14.97$, HB-corrected $p < .01$) was substantially greater (Wald $\chi^2(2, N = 19) = 18.32$, $p < .001$).

Variable Frequency Generator (VFG), Figure 3E. Among participants in the Experimental group, significant gains were evident for their knowledge of the VFG maintenance task, such that average scores increased 45.83% between

the initial and final evaluation, from 53.33% (*SD* = 014) to 77.78% (*SD* = 0.16; Q(2, N = 9) = 6.07, $p < .05$). A comparable improvement was not indicated for Control participants, whose mean scores for VFG items changed by 16.13%, from 62.00% (*SD* = 0.18) to 72.00% (*SD* = 0.23). No post-hoc comparisons among Experimental participants achieved statistical significance, nor was there a significant interaction between group and study phase.

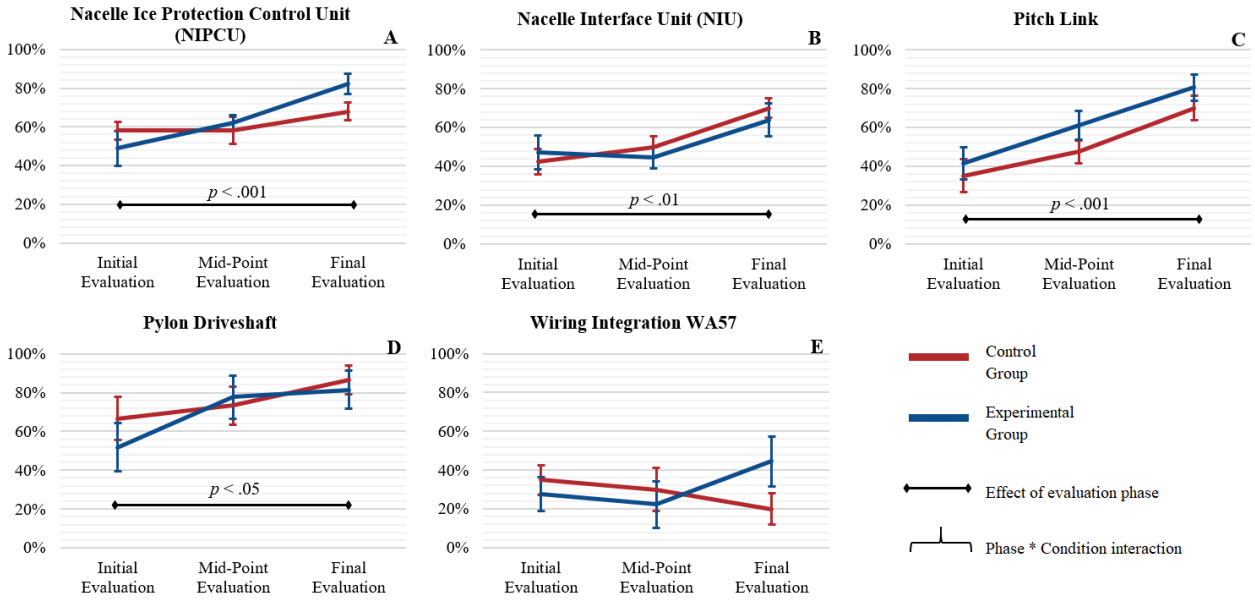

Figure 4. Knowledge evaluation scores for: A) nacelle ice protection control unit; B) nacelle interface unit; C) pitch link; D) pylon driveshaft; and E) wiring integration.

Nacelle Ice Protection Control Unit (NIPCU), Figure 4A. Experimental phase participants demonstrated significant gains in their knowledge of NICPU removal and replacement such that their average score increased from 48.89% at the initial measure (*SD* = 0.27) to 82.22% (*SD* = 0.16) for the final evaluation (Q(2, N = 9) = 10.13, HB-corrected $p < .05$). Post-hoc comparisons confirmed that the 68.18% improvement over baseline was significant ($p < .01$). No meaningful improvement was evident for Control participants.

Nacelle Interface Unit, No. 1 (NIU), Figure 4B. Average knowledge evaluation scores for the NIU task increased substantially among participants in the Experimental group, from 47.22% (*SD* = 0.26) to 63.89% (*SD* = 0.25), though this difference over time achieved only marginal significance (Q(2, N = 9) = 5.64, $p = .059$). Conversely, the improvement for participants in the Control condition was statistically significant, with their average score for the NIU steps increasing from 42.50% (*SD* = 0.21) to 70.00% (*SD* = 0.16; Q(2, N = 10) = 8.58, HB-corrected $p < .05$), and a significant difference between the first and final measures ($p < .05$).

Pitch Link, Figure 4C. For both Experimental and Control participants, the effect of time on knowledge evaluation scores was also significant for the pitch link task. In particular, average scores for Experimental participants nearly doubled, increasing from 41.67% (*SD* = 0.25) to 80.56% (*SD* = 0.21; (Q(2, N = 9) = 12.65, HB-corrected $p < .05$), while post-hoc analyses confirmed significant between-phase difference for the initial and mid-point assessments and the initial and final assessments (p's $< .01$). Comparable improvement was found among Control participants for change over the course of the study whose average scores increased from 35.00% to 70.00% (Q(2, N = 10) = 10.69, HB-corrected $p < .05$), with a significant difference between the first and final measures ($p < .05$).

Finally, analysis of the general knowledge evaluation items revealed significantly improved performance for both groups. In particular, the average score increased 38.89% for the Experimental group, from 66.67% to 92.59% (Q(2, N = 9) = 17.0, $p < .001$), and 31.71% for the Control group (68.33% to 90.00%, Q(2, N = 10) = 13.3, $p < .01$). Post-hoc comparisons for both groups revealed no significant differences between individual assessment phases (p's $> .06$). Further, there were no significant changes in knowledge evaluation scores for the heat exchanger (Figure 3A), tilt axis gearbox (TAGB, Figure 3D), the pylon driveshaft (Figure 4D), or wiring integration assembly (Figure 4E) tasks.

Self-Efficacy Responses

Heat Exchanger, Figure 5A. Initial self-efficacy ratings for leading completion of the heat exchanger task were marginally lower among Experimental participants (M = 2.89, SD = 1.27) compared to those in the Control group (M = 4.00, SD = 0.82; W = 68.5, p = .053). This difference was no longer evident by the mid-point evaluation, and by the final evaluation Experimental participants' self-efficacy for the task had increased significantly to a numeric mean of 4.44 (SD = 0.53; Q(2, N = 9) = 9.85, HB-corrected $p < .05$), increasing more than a full rating level. Moreover, the interaction between the group and study phase variables was significant (Wald $\chi^2(2, N = 19) = 8.51, p < .05$), indicating that the effect of study phase was greater for Experimental participants over Controls. There was however no significant change among Control participants for their self-efficacy on this task.

Slipring Standpipe, Figure 5C. The reverse pattern was evident for changes in self-efficacy to lead completion of the slipring standpipe task. On the initial measurement, self-reported confidence was significantly greater among Control participants (M = 3.60, SD = 0.70) compared to those in the Experimental group (M = 2.67, SD = 1.00; W = 71.5, $p <$ .05), though the difference was eliminated at the final measurement. Additionally, participants in the Experimental group reported significant gains in self-efficacy with numeric ratings increasing to 4.33 (SD = 0.87) at the final measurement (Q(2, N = 9) = 8.08, HB-corrected $p < .05$). Finally, results of the mixed-factor analyses indicated a significant interaction between the study phase and group factors (Wald $\chi^2(2, N = 19) = 10.07, p < .01$) with the greatest effect of study phase being evident for Experimental participants.

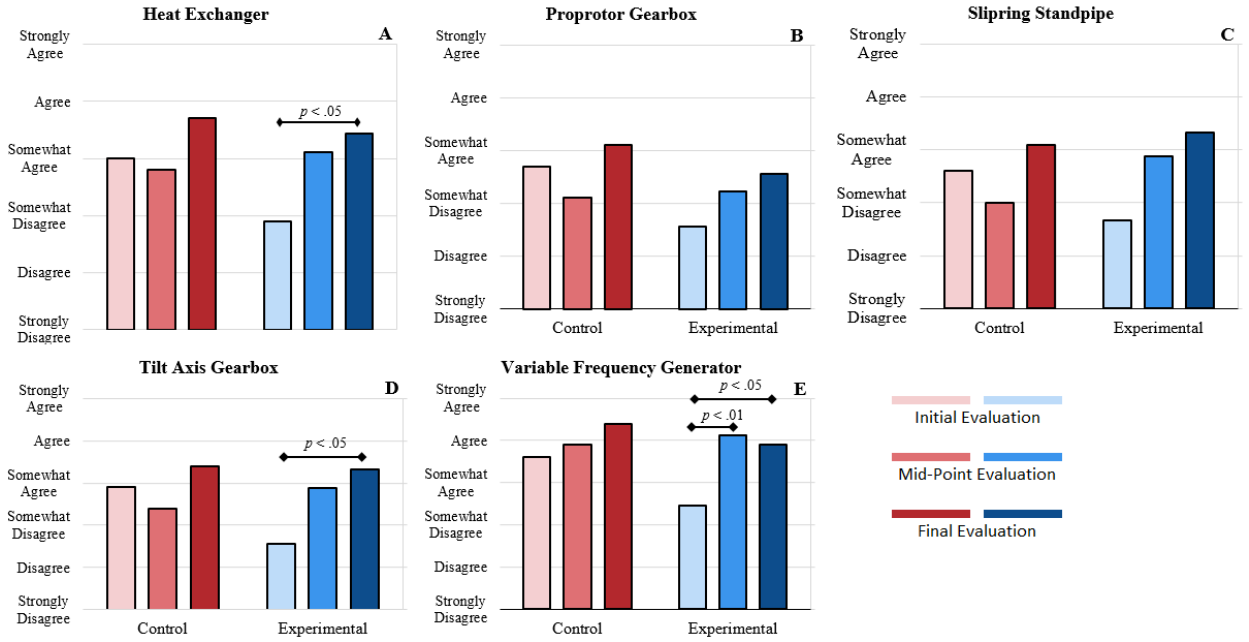

Figure 5. Self-efficacy responses for: A) heat exchanger; B) proprotor gearbox; C) slipring standpipe; D) tilt axis gearbox; and E) variable frequency generator.

TAGB, Figure 5D. As observed for the slipring standpipe task, participants' self-efficacy for leading completion of the TAGB increased significantly over the course of the evaluation period. In particular, self-efficacy for the Experimental group (M = 2.56, SD = 1.01) was significantly lower than that of the Control group (M = 3.90, SD = 0.74) at the initial evaluation (W = 76, $p < .01$), and demonstrated the greatest improvement. The numeric average of Experimental participants' self-ratings increased nearly two rating levels to 4.33 (SD = 1.22) at the final assessment (Q(2, N = 9) = 9.80, HB-corrected $p < .05$). This difference was significantly greater than the half-level improvement evident for Control group (final M = 4.40, SD = 1.26; Q(2, N = 10) = 6.07, $p < .05$). Further, post-hoc comparisons revealed a significant difference between initial and final measures among Experimental participants ($p < .05$), though no difference was evident among maintainers in the Control group (p = .085).

VFG, Figure 5E. Likewise, Experimental participants' self-efficacy on the VFG task increased significantly, with the numeric average rising 1.5 rating levels from 3.44 (*SD* = 1.13) initially to 5.11 (*SD* = 1.05) at the mid-point, though there was a slight decrease at the final measurement (*M* = 4.89, *SD* = 0.605; Q(2, N = 9) = 11.67, HB-corrected $p < .01$). Post-hoc comparisons revealed significant paired comparisons between the initial and mid-point evaluation ($p < .01$) and initial and final evaluation ($p < .05$). Significant improvements were also found for participants in the Control group, with average numeric self-efficacy ratings increasing nearly one level, from 4.60 (*SD* = 1.07) initially to 5.40 at the final assessment (*SD* = 0.84; Q(2, N = 10) = 6.32, $p < .05$), though post-hoc comparisons were not significant. Additionally, there was a significant difference between the two groups for their initial self-efficacy ratings (W = 69, $p < .05$), though this distinction was not evident at either the mid-point or final evaluation.

There were no significant changes for participants' self-efficacy in relation to the PRGB task.

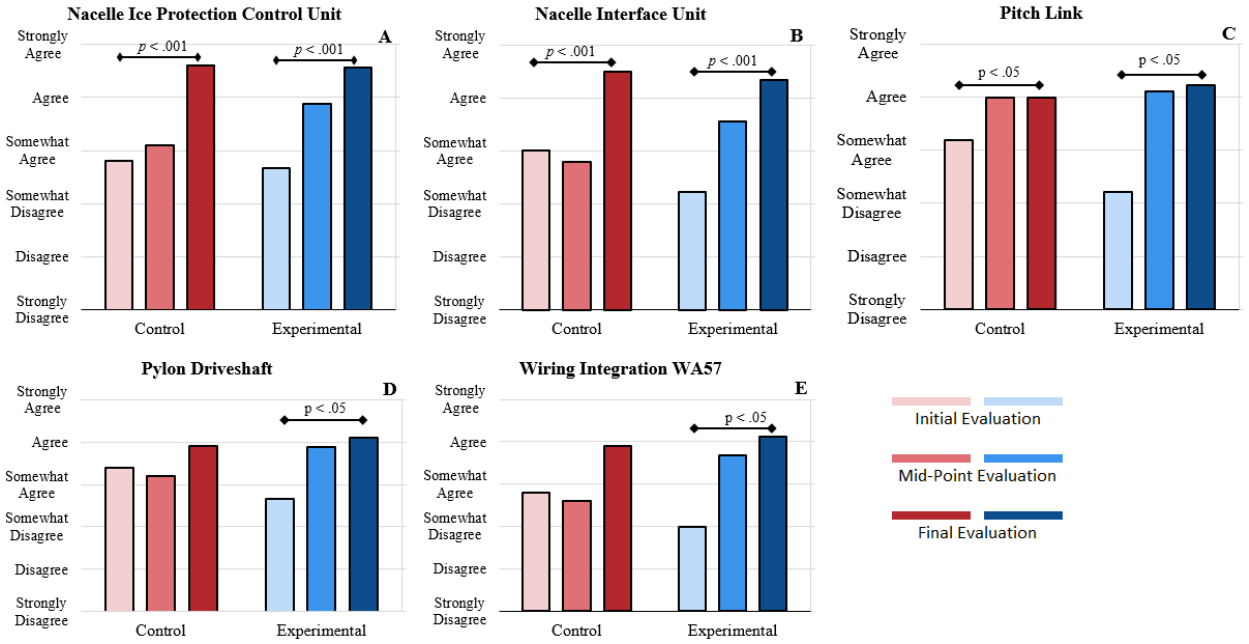

Figure 6. Self-efficacy responses for: A) nacelle ice protection control unit; B) nacelle interface unit; C) pitch link; D) pylon driveshaft; and E) wiring integration.

NIPCU, Figure 6A. Experimental participants reported confidence for task completion increased by nearly two full step ratings, from an initial numeric average of 3.67 (*SD* = 1.32) to 5.56 (*SD* = 0.53; Q(2, N = 9) = 14.00, HB-corrected $p < .05$). As demonstrated in Figure 6A, post-hoc Nemenyi tests further revealed that the difference between the initial and final measurements was significant ($p < .01$). Similar increases were evident in the self-efficacy ratings within the Control group, with average numeric ratings growing from 3.80 (*SD* = 0.92) to 5.60 (*SD* = 0.70; Q(2, N = 10) = 14.97, HB-corrected $p < .01$). Moreover, post-hoc comparison confirmed a statistically significant difference was present between the initial and final measures ($p < .01$).

NIU, Figure 6B. As seen for the NIPCU, participants in both conditions demonstrated significantly increased self-efficacy ratings for leading completion of the NIU task. The gain was greatest among Experimental participants, whose rating increased two rating levels from an average of 3.22 (*SD* = 0.83) at the initial measurement to 5.33 (*SD* = 0.71) at the final measurement (Q(2, N = 9) = 13.03, HB-corrected $p < .01$). A significant, though somewhat smaller, improvement was also seen among participants in the Control condition, with an initial average rating of 4.00 (*SD* = 1.15) which then increased to 5.50 (*SD* = 0.71; Q(2, N = 10) = 15.49, HB-corrected $p < .01$). Post-hoc comparisons for both groups indicated the difference between the initial and final rating was significant (p's $< .01$). Additionally, a significant time/group interaction was found for this measure (Wald $\chi^2(2, N = 19) = 12.6$, $p < .01$), again indicating that the effect of study phase was highest for Experimental participants.

Pitch Link. At their initial evaluations, participants in the Control group indicated significantly higher confidence in their ability to lead completion of the pitch link task (M = 4.20, SD = 1.14) in comparison to the Experimental participants (M = 3.22, SD = 0.83; W = 68.5, $p < .05$). This difference was eliminated at the mid-point evaluation and participants in the Experimental condition indicated a significant increase in self-efficacy over the course of the study (Q(2, N = 9) = 13.24, HB-corrected $p < .01$), increasing two full rating levels to 5.22 (SD = 0.83). Moreover, post-hoc comparisons indicated significant differences between both the initial and mid-point as well as the initial and final measures (p's < .05). The improvement among Control participants also achieved statistical significance, with final average ratings of 5.00 (SD = 0.82; Q(2, N = 10) = 6.28, $p < .05$), though post-hoc comparisons failed to reach significance.

Pylon Driveshaft, Figure 6D. Experimental participants' self-efficacy for leading removal and replacement of the pylon driveshaft assembly improved nearly 1.5 rating levels, growing from an initial average of 3.67 to 5.11 (SD's = 1.66 and 0.60, respectively) at the final measure (Q(2, N = 9) = 7.91, HB-corrected p < .05). Conversely, the change among Control participants did not achieve significance, only increasing from 4.40 to 4.90 (SD's = 1.26 and 0.99, respectively) and degrading somewhat at the mid-point (M = 4.20, SD = 1.69). Additionally, the interaction between the phase and condition factors was significant (Wald $\chi^2(2, N = 19) = 7.35$, $p < .05$) revealing a greater effect for study phase among Experimental participants compared to Control.

Wiring integration WA57, Figure 6E. The increase in self-efficacy reported by Control participants achieved only marginal significance (p = .058), though a significant improvement was evident among Experimental participants (Q(2, N = 9) = 12.00, HB-corrected $p < .01$). The initial numeric rating for the Experimental group was 3.00, which then increased two rating levels to 5.11 at the final measurement, in comparison to the increase from 3.80 (SD = 1.03) to 4.90 (SD = 1.52) among Control participants. Additionally, post-hoc comparisons confirmed the difference between the initial and final ratings for the Experimental group was significant ($p < .05$).

DISCUSSION

As hypothesized, these results provide compelling evidence for both objective and subjective improvements among 3-Level CV-22 maintainers who completed virtual training using the IMG. Average improvements on the knowledge evaluation were substantially greater among Experimental participants compared to those in the Control group at each time point, such that participants in the Experimental group improved 22.86% during the first 10 weeks and 53.71% overall, compared to 12.95% and 37.82% gains respectively for the Control participants. Further, comparison of total scores at the final evaluation revealed a marginally significant difference between the two groups, with average scores for participants who used the IMG being 12.36% greater than those of their peers. Moreover, it was demonstrated that use of the IMG was associated with significantly increased knowledge of three tasks (NIPCU, PRGB, and VFG) for which no meaningful gains were observed among Control participants.

Even more impressive gains were evident for the measure of participants' confidence to lead completion of the focal tasks. Compared to maintainers in the Control group, Experimental participants reported significantly lower confidence of self-efficacy at the initial evaluation for four tasks (pitch link, slipring standpipe, TAGB, and VFG), but for each the difference was eliminated by the mid-point (second) assessment. Additionally, the reported self-efficacy for members of the Experimental group increased significantly for nine of the 10 tasks (excluding the PRGB), whereas ratings increased for only five tasks among those in the Control group.

Additionally, the current study revealed that maintainers who used the IMG demonstrated similar gains on both routine and comparatively simpler maintenance tasks (e.g., pitch link adjustment, NIU replacement) as well as less common and more labor intensive tasks (e.g., replacement of heat exchanger, PRGB). This latter finding demonstrates one of the greatest possible benefits of using the IMG to supplement learning; namely that participants are able to virtually experience the complete progression of a task. Several of the tasks included in this initial application, such as the PRGB, TAGB, and heat exchanger, require several teams of maintainers working over the course of multiple shifts and days to complete. Alternatively, the removal of a given component might be completed, but the subsequent replacement does not take place until days or weeks later. Consequently, this process restricts maintainers from the opportunity to view the entire task from start to finish, and substantially limits their ability to properly understand the full procedure. Thus, as an adjunct training tool and resource, the IMG could be used to fill in these experiential gaps and allow maintainers of any level to review the task as a whole.

Initial knowledge evaluation scores and self-efficacy responses were generally lower among members of the Experimental group, but in each case rose to meet or exceed that of their peers in the Control group. This finding suggests that use of the IMG was in many cases associated with more rapid gains in both knowledge and self-efficacy than were apparent among maintainers who did not use the technology. Though the IMG does not represent game-based learning (GBL) or "serious games" in the traditional sense as there are no gaming elements, it does take advantage of the visual cues and active learning techniques common in GBL environments. Thus it is likely that experiential learning is happening through IMG use, and that knowledge and skills gained on the virtual task would translate to improved hands-on performance (Ahmed & Sutton, 2017). Moreover, the results of this assessment, in conjunction with evidence from Manacapilli et al. (2007), indicate even greater benefits might be evident if the IMG were made available while Airmen are still in the schoolhouse and initial skills training phase. This is further supported by the finding that the most significant difference in final knowledge evaluation scores between the two groups was evident for PRGB replacement, by far the most complex task included in the current evaluation.

In light of these encouraging findings, it is important to note some of the limitations of this operational assessment, foremost being the small sample size. Smaller samples are naturally less powerful, a problem exacerbated by the use of non-parametric analyses. However, effect sizes for many measures were greater than 0.5, indicating good concordance as well as a strong likelihood similar results would be obtained with a larger sample. Consequently, it is possible that greater effects might be evident if a larger sample could be made available for the assessment. Moreover, generalizability of the findings is somewhat limited by the concurrence of data collection with the global COVID-19 pandemic. Health and safety protocols required substantial changes to manning and duty schedules, and participating maintainers were often quartered for up to a week as a precaution following suspected contact. Finally, the infrequent nature of the 10 selected tasks eliminated the possibility of obtaining meaningful hands-on proficiency data, and consequently no directly operational relevant data could be included in the current assessment. Although both objective measures of knowledge for a task as well as self-assessed readiness to perform that task are psychometrically sound, neither is able to adequately substitute a demonstration of actual skill.

Collectively, the current study revealed a number of objective and subjective improvements associated with maintainers' use of the IMG as a training support tool. Use of the IMG was associated with final knowledge evaluation scores 12.36% greater than those of their peers. Additionally, the change in participants' ratings of confidence in their readiness to lead completion of the work increased significantly for nine of the 10 tasks, compared to five among Control group maintainers. Though it is not possible to directly quantify these changes in terms of improved skill, the findings clearly demonstrate that use of the IMG as a training adjunct further improved maintainers' foundational knowledge of and sense of readiness for maintenance tasks. Since greater knowledge and self-efficacy are critical components of improved proficiency and performance, it is further likely that use of the IMG would be associated with higher efficiency and a reduction in rework, which in turn results in substantially decreased aircraft downtime.

REFERENCES

Ahmed, A., & Sutton, M. J. (2017). Gamification, serious games, simulations, and immersive learning environments in knowledge management initiatives. *World Journal of Science, Technology and Sustainable Development, 14*(2/3), 78-83.

Bolkcom, C. (2004, April). *V-22 Osprey Tilt-rotor Aircraft*. Library of Congress, Washington DC: CRS. https://digital.library.unt.edu/ark:/67531/metacrs7265/m1/1/high_res_d/RL31384_2005Aug04.pdf

Goldberg, B., Davis, F., Riley, J. M., & Boyce, M. W. (2017, July). Adaptive training across simulations in support of a crawl-walk-run model of interaction. In I*nternational Conference on Augmented Cognition* (pp. 116-130). Springer, Cham.

Losacker, B. T. (2019). *Combat search and rescue: Restoring promise to a sacred assurance*. Air University, Air Command and Staff College.

Manacapilli, T., Bailey, A., Beighley, C., Bennett, B., & Bower, A. (2007). *Finding the balance between schoolhouse and on-the-job training*. Pittsburgh, PA: RAND Corporation.

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12), 1-23.

Themanson, J. R. & Rosen, P. J. (2015). Examining the relationships between self-efficacy, task-relevant attentional control, and task performance: Evidence from event-related brain potentials. *British Journal of Psychology, 106*(2), 253 – 271.