Using Virtual Reality for Training Maintenance Procedures

Shannon K.T. Bailey[1], Cheryl I. Johnson[1], Bradford L. Schroeder[2], and Matthew D. Marraffino[1]

[1]Naval Air Warfare Center Training Systems Division, Orlando, FL
[2]StraCon Services Group, LLC, Orlando, FL

shannon.bailey@navy.mil, cheryl.i.johnson@navy.mil, bradford.schroeder.ctr@navy.mil, matthew.marraffino@navy.mil

ABSTRACT

In light of rapid technology advances and budget declines, the Navy is exploring innovative training solutions though initiatives such as Sailor 2025 and High Velocity Learning, which call for more hands-on, learner-centric training. Consistent with these initiatives, virtual reality (VR) offers a low-cost alternative to traditional methods of training by offering Sailors interactive and immersive 3-D simulation environments to train critical skills. Indeed, theoretical research predicts that such immersive training will result in better learning outcomes for training a procedural task than traditional computer-based training, yet there are few systematic experiments examining how and why VR may be effective for training. We conducted an experiment to: 1) test whether VR is as effective for training a military-based task as desktop-based training, and 2) compare two different input methods for interacting within the VR environment. Eighty-three participants were trained on maintenance procedures for the E-28 arresting gear, a system that hooks aircraft and rapidly decelerates them as they land. Participants were assigned randomly to one of three training conditions: Desktop-based simulation, Gesture-based VR, or Voice-based VR. A written recall test served as our measure of learning outcome. We analyzed the errors that trainees made during training and found differences between the conditions that suggest that Desktop training may be less efficient than VR training: The Desktop group committed more procedure-based errors, while the VR-Gesture group committed more gesture-related errors (indicating they understood the procedure but had issues with using the system). This experiment addresses a critical gap in VR research by examining characteristics that may contribute to VR training optimization. Furthermore, these results demonstrate the potential of VR to provide ready, relevant training to the Fleet.

ABOUT THE AUTHORS

Shannon K. T. Bailey is a doctoral candidate in Psychology at the University of Central Florida where she received a M.A. in Applied Experimental and Human Factors Psychology. She is a Research Psychologist at NAWCTSD in Orlando, FL. Her research experience includes investigating the effectiveness of training simulations in virtual reality, gesture-based technology interaction, and spatial ability.

Cheryl I. Johnson is a Senior Research Psychologist at NAWCTSD in Orlando, FL, performing research on emerging training technology and adaptive training systems. She earned her M.A. and Ph.D. in Cognition, Perception, and Cognitive Neuroscience from the University of California Santa Barbara. Dr. Johnson has over 12 years of experience performing technology-based training research. Her research interests include adaptive training, instructional strategies, virtual reality training applications, and multimedia learning.

Bradford L. Schroeder is a doctoral candidate at the University of Central Florida, where he received a M.A. in Applied Experimental and Human Factors Psychology. He works as a Research Psychologist at StraCon Services Group, LLC in support of NAWCTSD in Orlando, FL. His research experience includes examining individual differences in technology use, such as texting behavior, video game performance, and virtual reality-based learning.

Matthew D. Marraffino is a Research Psychologist in the Simulation and Training Research to Improve Knowledge and Effectiveness (STRIKE) Lab at NAWCTSD in Orlando, FL. He received his Ph.D. in Applied Experimental and Human Factors Psychology at the University of Central Florida. His research examines how best to use cognitive theories of learning (e.g., Cognitive Load Theory and Embodied Cognition) to guide the design of trainers using state-of-the-art technology including computer-based training, tablet computers, and virtual reality.

Using Virtual Reality for Training Maintenance Procedures

Shannon K.T. Bailey[1], Cheryl I. Johnson[1], Bradford L. Schroeder[2], and Matthew D. Marraffino[1]

[1]Naval Air Warfare Center Training Systems Division, Orlando, FL
[2]StraCon Services Group, LLC, Orlando, FL

shannon.bailey@navy.mil, cheryl.i.johnson@navy.mil, bradford.schroeder.ctr@navy.mil, matthew.marraffino@navy.mil

INTRODUCTION

In a time of rapid technology advances and training budget constraints, the Navy is exploring innovative training solutions to promote more hands-on training opportunities through initiatives such as Sailor 2025 and High Velocity Learning. The goal of these initiatives is to provide Sailors learner-centered training available at the point of need. Consistent with these initiatives, virtual reality (VR) offers a low-cost alternative to traditional methods of training by offering Sailors interactive and immersive 3-D simulation environments to train critical skills. Virtual reality (VR) systems are of interest to the training community because they have the potential to simulate complex tasks that are not feasible or practical to teach in real life due to safety or cost concerns. Although VR may offer a safer or more cost effective alternative, the effectiveness of VR should be evaluated to determine whether these systems offer an added value over other types of training (e.g., desktop-based computer training). Research in the last few decades has shown that VR is promising for procedural knowledge acquisition (Buttussi & Chittaro, 2017; Hamblin, 2005; Loftin et al., 1994), but previous VR studies have limitations that make generalizing results to other training systems tenuous, such as lack of control groups and random assignment to conditions (Ganier, Hoareau, & Tissequ, 2014). Additionally, the extent to which different features of VR (i.e., immersion, interactivity) are responsible for learning outcomes needs to be assessed to understand how to optimize VR for training. To address these gaps in the literature, we conducted an experiment to systematically compare VR and Desktop training to determine whether VR adds value over traditional computer-based training. Furthermore, the reason VR may or may not be beneficial for training should be investigated to elucidate which features of VR are useful to include in training systems.

Two potential reasons VR might be effective for training are that it is immersive and/or that it allows more physical interactions with the training system. First, a 3-D environment (i.e., VR) may be more immersive than a 2-D environment, such as a desktop system (Nash, Edwards, Thompson, & Barfield, 2000). Early theoretical work defined immersion as a characteristic of technology that produces a “vivid illusion of reality” (Slater & Wilbur, 1997). Slater and Wilbur described a four factor model characterizing immersive technology, including the technology’s ability to shut out the physical world (“inclusive”), utilize multiple sensory modalities (“extensive”), use panoramic field of view (“surrounding”), and contain a richness of content and resolution (“vivid”). VR systems may employ each of these four characteristics to a larger extent than traditional desktop-based training. For example, a VR system could occlude the physical world with a head-mounted display (HMD), include multiple types of sensory information (e.g., visual, audio, tactile), offer panoramic field-of-view HMDs, and provide more depth cues that add to content richness.

Subsequently, the immersive quality of VR may produce a sense of presence, or “being there” (Blade & Padgett, 2002), and increased presence may result in better training outcomes (Sheridan, 1992). Witmer and Singer (1998) explained that immersion leads to feeling present in a virtual environment because the virtual world is experienced directly and thus the experience is more meaningful. They then suggested that because the experience is more meaningful with a sense of presence, immersive training should be better for learning than traditional 2-D computer-based training. Although presence has been purported in theoretical articles to enhance learning outcomes from immersive environments, the empirical evidence supporting this claim is limited and more research is needed (Mikropoulos & Natsis, 2011; Stevens & Kincaid, 2015). To test whether immersion is important for learning a procedural task, we compared learning outcomes from training in VR to that of a desktop simulation. We hypothesized that VR would be more effective for training a procedural task than a desktop-based environment because VR is theorized to be more immersive.

Alternatively, it may not be the immersive quality of VR that is effective for training, but the type of interaction with the training system that helps learning. The second feature of VR training that may contribute to better learning outcomes is the interactivity between the learner and the system. Specifically, interactions with the training system that are closer to the physical task may facilitate better understanding of the learning material. For example, gesture-based commands may be better for learning a procedure because the participant is physically enacting the task. Gestures have been shown to aid in understanding new information, such as learning a math concept (Cook, Mitchell, & Goldin-Meadow, 2008). Additionally, physically enacting an action has been shown to improve recall for that action in a finding called the "enactment effect" (Engelkamp & Jahn, 2003). Research has shown that enacting is better for recall than hearing the action or seeing someone else perform the action, and memory for the enacted information is better even weeks later (Nilsson, Cohen, & Nyberg, 1989). To test whether the type of interactivity in VR results in better learning, we compared recall between two types of VR training: VR with gesture-based commands or VR with voice-based commands. We predicted that VR with gestural commands would be better for learning than voice-based commands because the interaction represents the physical action, leading to an enactment effect.

Because the conditions differed in the way participants interacted with the virtual environment, we also examined participants' ratings of system usability, or how easy it is to learn and use a technology. Systems that are harder to use may hinder trainees' ability to learn from the training system. Thus, evaluating the usability of novel training systems is essential to ensure that any differences in performance due to training quality are not a function of one condition being more user-friendly than another. The conditions were pilot-tested thoroughly to ensure participants could complete the training as intended, so we did not expect perceived usability issues across the conditions.

By comparing VR and Desktop training on learning a procedural maintenance task, we can determine whether it is the immersive or interactive features that lead to learning outcomes. Specifically, if both VR training conditions result in better learning outcomes (as measured by recall performance) than Desktop, but do not differ from each other, there is evidence that an immersive environment is effective for training regardless of interaction type; however, if the VR conditions differ from each other on recall performance, this suggests that interaction type within a virtual environment is indicative of learning differences. If neither VR condition differs from Desktop training, there is not supporting evidence that an immersive environment is necessary for training a procedural task. In addition to comparing training types on learning outcomes (i.e., recalling the procedure), we also investigated whether the performance within the training differed, which could affect learning outcomes. To differentiate the quality of training types, we considered performance during the training phases (e.g., time to complete training, kinds of errors made during training). Evidence from these analyses can be used in development of future simulation training in determining the most appropriate features of training systems for desired learning outcomes.

METHOD

Participants and Design

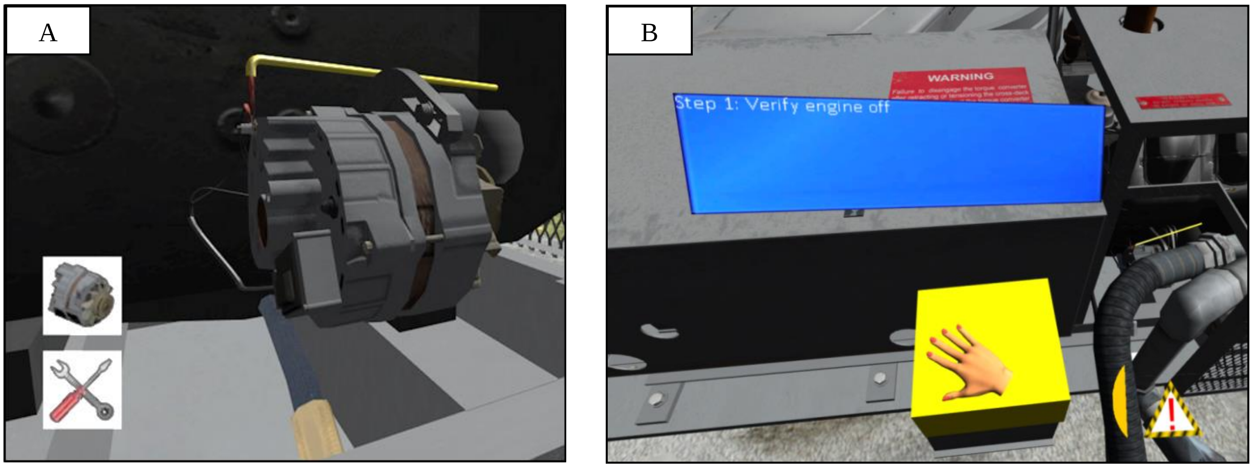
Participants were assigned randomly to one of three between-subjects training conditions: Desktop (control), VR Voice, or VR Gesture. Participants (n=83) were recruited from a university through the university's research participation system. Participants were excluded from participating if they indicated that they had mechanical experience. The ages ranged between 18-36 years old, and the average age of participants was 21.29 years (*SD*=3.47). There were 36 males and 47 females. Participants were paid $15 per hour for up to three hours of participation.

Training Task and Testbed

Three versions of the testbed were developed for the experiment. The testbed was developed using the Unity 3-D game engine and was presented either on a 2-D desktop monitor (Desktop condition) or 3-D VR stereoscopic head-mounted display (HMD; VR Gesture or VR Voice). For the VR conditions, the HMD was the Oculus Rift DK2, and head, body, and hand movements were tracked using the Microsoft Kinect V2 infrared motion capture.

In all three conditions, participants completed a practice task of removing and replacing the engine cage of the E-28 arresting gear, which is a land-based emergency gear for arresting hook-equipped aircraft, to become comfortable with the controls for interacting within the training environment. The practice task was followed by the main task for training the procedure for removing and replacing the E-28's alternator, and participants performed this task three times. In the training procedure (Figure 1**)**, the number of required attempts was 22 steps to remove and replace the alternator. Each step involved selecting the appropriate part and/or tool needed and performing the action command for that step of the procedure (i.e., gesture command, voice command, or mouse click). For example, to replace the bolt on the adjustment arm of the alternator, the participant would select the appropriate bolt from the parts bin, the common hand tools from the toolbox, and use the replace action on the location where they want to place the bolt. Figure 1A shows the participant's view of the alternator, and the icons that indicate the part and tools that are currently selected. For all three training phases, there were a combined total of 66 required steps. The three training phases became more difficult with time as less instructional scaffolding was given in each subsequent phase. In the first training phase, the participants received scaffolding in the form of narrated instructions, step text, and highlighting to indicate where the correct part was located (Figure 1B). In the second training phase, participants received narrated instructions and step text, but not highlighting. In the final training phase, participants had to perform the procedure from memory and were not given any narration, step text, or highlighting.

In addition to the display (2-D vs. 3-D), condition differed by how participants interacted within virtual environment. The Desktop condition used a mouse and keyboard to move around the virtual environment on a 2-D monitor, completing actions by clicking the mouse. The VR conditions both used five commands to complete the procedural actions: select, open, close, remove, and replace. For each of the five actions, the VR Voice condition said the action aloud to complete a step, and the VR Gesture condition performed a gesture. The gesture commands were full-body positions using the right arm that represented the physical action. To select a part or tool, the participant held their hand over the part/tool they wanted to select and made a closed-fist grasp. To open a part (e.g., open control panel), the participant lifted their arm up to the right side, making a 90 degree angle at the elbow. Closing a part was done by dropping the arm from chest height down past the hip. To remove a part, the participant put their right arm out to their right side so that the arm was parallel to the ground. Finally, to replace a part, the right hand was brought up to the left shoulder, crossing the chest. These gestures were relatively static in that the correct action triggered an animation of the step, and the command to computer response timing was not a 1:1 relationship.



Step 1: Verify engine off

Figure 1. A) Screenshot of the E-28 arresting gear alternator. The alternator was removed and replaced in the training procedure. The icons on the left indicated the part (e.g., replacement alternator) and tool (e.g., common hand tools) that were equipped. B) Screenshot from the first training phase (which included narrated instructions, step text, and highlighting object) of the VR system in which the participant moved their virtual hand to the highlighted object and used either verbal or gesture-based commands to complete an action. The blue box contained instructions for completing each step of the procedure (i.e., step text). The caution icon in the lower right indicated when a participant was near a physical wall in the room to avoid collisions.

Materials

Learning outcome was measured by a written recall measure asking participants to write the procedural steps for replacing an alternator in the correct order, listing all tools and parts associated with each step. Participants were given five minutes to complete this task. The 19-item Presence Questionnaire (PQ; Witmer & Singer, 1998) was used to report feelings of presence, or "sense of being-there" from the training environment (α=.76). Responses ranged on a scale of 1 ("Not at All") to 7 ("Completely"), indicating agreement with each item. The 10-item System Usability Scale (SUS; Brooke, 1996; α=.69) was used to assess how usable participants felt each training environment was by indicating agreement to statements on a scale of 1 ("Strongly disagree") to 5 ("Strongly agree").

Procedure

After reading an informed consent and agreeing to participate, participants completed individual difference measures. Participants then read a slideshow-style tutorial for approximately 10 minutes that explained the procedural task they would be learning in the training (i.e., removing and replacing the alternator of the E-28 arresting gear) and instructions for interacting with the virtual environment consistent with training condition to which they were assigned. After the tutorial, participants completed a brief knowledge check to confirm they had read the tutorial before proceeding. Participants in the VR conditions briefly practiced the voice- and gesture-based commands and were then calibrated in the VR system. Next, participants completed a short practice task of removing and replacing the engine cage (about 5 minutes) and were allowed to ask questions during the practice task. After the practice task, participants completed three training phases for the procedure of removing and replacing the alternator, with decreasing amount of scaffolding in each phase until the final phase was completed from recall (average time to complete training phases reported in Table 1). Once the training was complete, participants completed the PQ, SUS, and the written recall measure.

RESULTS

IBM SPSS 19 was used to conduct a series of one-way ANOVAs. First, we first analyzed how many steps participants could remember from the training procedure in a written recall test by condition. Written recall of the procedure did not differ overall across training conditions, $F(2, 80)=0.60$, $p=.55$, with conditions recalling an average of 15.81-17.31 out of 22 steps (Table 1). We then tested whether immersion during training, as measured by feelings of presence, was related to recall. We found that there were no differences in feelings of presence between the training conditions, $F(2, 80)=0.03$, $p=.97$. Average PQ ratings were around 5.2 for all conditions, which was slightly above the mid-point on a scale of 1-7. Although we hypothesized that more immersion would lead to better recall and more natural interactions would lead to better recall, we did not find any differences in recall as a function of immersion or interactivity. Finally, to determine whether type of interaction was better or worse for recall due to usability issues, we tested whether reports of usability differed by condition. Again, there were no differences in overall usability depending on condition, $F(2, 66.62)=0.78$, $p=.45$ (Levene's test was significant, $p=.02$, so the Brown-Forsythe correction was reported), although usability for all training conditions was good overall (Table 1; Bangor, Kortum, & Miller, 2008). Because there were no differences overall in how much a participant could recall from training or feelings of presence and usability, we conducted further analyses to investigate why these theorized relationships were not found. By analyzing the time spent in training and the kinds of errors made during training, we were able to explore how the learning processes may have differed across the three training conditions.

Time

The amount of time to complete the three training phases was measured in minutes for each training condition. The first two phases were longer due to the narration of instructions, and the last phase was the time to perform the procedure without narration (i.e., time to recall the task without scaffolding). Overall, the average time to complete the three training phases was 24:59 min. (*SD*=8:38 min.), and there was a large range in the amount of time the training took to complete (*Minimum*=13:05 min., *Maximum*=57:05 min.). Time to complete all three training phases was correlated negatively with written recall of the procedure, such that those who completed the training faster remembered more steps later, $r(83)=-.306$, $p=.005$.

To determine whether training times differed between conditions, we conducted a one-way ANOVA. There was a significant difference in the amount of time spent in training depending on condition, with Desktop participants completing the training faster than the VR Voice and VR Gesture conditions, $F(2, 80)=17.55$, $p<.001$, $\eta_p^2=.305$, Table 1. LSD post hoc comparisons indicated Desktop training was about six minutes faster than VR Voice ($d=1.029$) and ten minutes faster than VR Gesture ($d=1.556$). The difference between VR conditions was not significant at $\alpha=.05$, although it was trending in the direction of VR Voice being faster than VR Gesture ($d=0.513$). Results showed that Desktop training was faster than either VR training conditions, but time in training may not directly indicate quality of training. For example, does rushing through a training phase lead to more mistakes during training, and are certain kinds of errors worse than others for later recall? To determine the efficacy of each type of training, we next analyzed the number of attempts it took to complete the procedure and the subsequent kinds of errors made in each training condition.

Table 1. Written Recall, Presence and Usability Ratings, Training Time, and Number of Attempts to Complete the Procedural Task by Condition

	Training Condition					
	VR Gesture (*n*=26)		**VR Voice (*n*=25)**		**Desktop (*n*=32)**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Written Recall (Number of Steps)	17.31	4.61	16.16	4.91	15.81	6.08
Presence Ratings	5.21	0.79	5.24	0.71	5.20	0.73
Usability Ratings	71.92	16.54	76.60	8.47	75.39	15.19
Training Time	30:29 min.	7:54 min.	26:31 min.	7:33 min.	19:19 min.	6:32 min.
Attempts Total[a]	103.42	23.41	83.76	11.61	84.13	13.56
Attempts Phase 1	40.19	12.63	31.60	5.39	28.31	4.79
Attempts Phase 2	32.15	8.18	26.12	3.73	25.88	4.03
Attempts Phase 3	31.08	6.01	26.04	4.37	29.94	8.16

Attempts

The number of attempts to complete each step of the procedure was recorded during the training phases (Table 1). The average number of attempts to complete the 66 steps was 90.06 (*SD*=18.93), and the range was between 67 attempts to 169 attempts. A 3 (Condition) X 3 (Phase) repeated-measures ANOVA was conducted on the number of attempts needed in each training phase to complete the procedure. To take into account the dependence of observations within subjects, we tested the sphericity assumption of repeated-measures ANOVA, i.e., the assumption that the difference in variances between conditions is similar (Field, 2013). Mauchly's test of sphericity was significant, $\chi^2(2)=8.55$, $p=.014$, so degrees of freedom were adjusted by the Huynh-Feldt estimate ($\varepsilon=.95$). Results indicated there was a main effect for Phase, such that the overall number of attempts went down with more training, $F(1.9, 152)=27.78$, $p<.001$, $\eta_p^2=.258$. There was also a significant main effect for Condition, $F(2, 80)=11.95$, $p<.001$, $\eta_p^2=.230$. VR Gesture had more total attempts than either VR Voice ($d=1.057$) or Desktop ($d=1.036$) with no difference between VR Voice and Desktop training ($d=0.029$). There was a significant Phase by Condition interaction effect, showing that the number of attempts in each phase depended on training condition, $F(3.8, 152)=9.30$, $p<.001$, $\eta_p^2=.189$. Although the overall number of attempts for all three phases was greater for VR Gesture, the VR Gesture and Desktop conditions did not differ in the final training phase ($p>.05$, $d=0.157$).

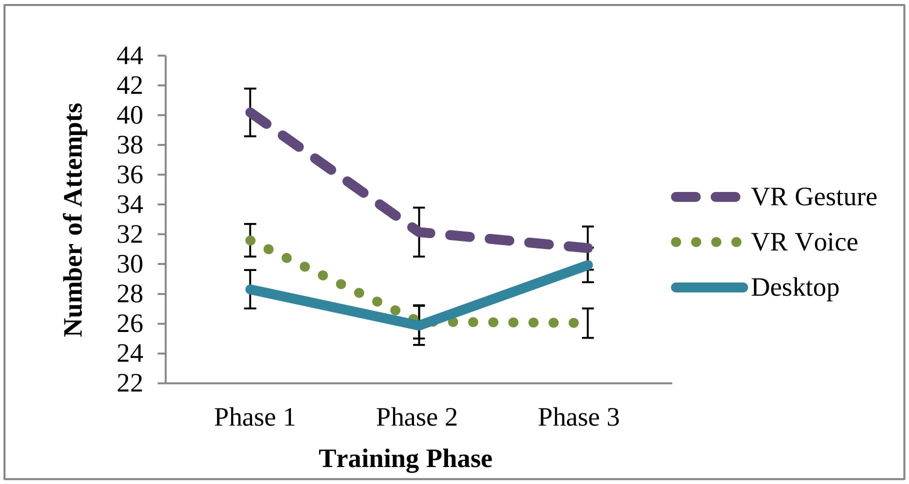


Figure 2. A 3 (Condition) X 3 (Phase) ANOVA was conducted on the number of attempts used to complete the 22 steps in each phase. The minimum number of steps needed to complete the procedure was 22. Both VR training groups improved in the number of attempts over time; however, the Desktop condition took more attempts in the final phase in which no help was given. Error bars represent standard error.

Figure 2 reveals that the Desktop condition started with fewer attempts, but sharply increased the number of attempts needed in the last phase (with no scaffolding). There was a significant simple effect for the Desktop condition showing that more attempts were made in the last training phase than the previous phase (p=.001, d=0.667). This result raised the question, why did the Desktop training condition become worse at performing the procedure when scaffolding was removed but the VR conditions did not? To answer this question, we investigated the kinds of errors made throughout training to determine whether the type of training led to different kinds of errors in training.

Errors

We focused the following analyses on the kinds of errors made during the last training phase in which the participant performed the procedure from recall without scaffolding (i.e., no narrated instructions, step text, or highlighting correct parts). There were three types of errors participants could perform: *wrong action*, in which a participant uses a gesture or command that is incorrect for the current step (e.g., enacting/saying "remove" instead of "replace"); *wrong part*, in which a participant tries to interact with the incorrect part for the current step (e.g., trying to replace the pivot bolt with the adjustment bolt); and *wrong tool*, in which a participant tries to interact with an object using the incorrect tool (e.g., trying to remove the alternator cable nut with the battery terminal puller). A greater quantity of wrong actions may reflect usability issues with the system, whereas more wrong tool and part errors would reflect errors in recalling the procedure. These errors were calculated as proportions of total actions committed in the last training phase (including correct actions), which we compared among the three training conditions. By analyzing errors in the last training phase, we were able to determine whether errors in each training condition were primarily usability related (i.e., wrong action) or a product of not recalling the procedure (i.e., wrong part or tool).

The proportion of wrong actions was inherently unequal between the VR and Desktop conditions, because the VR conditions used five gesture or voice commands to control actions, while the Desktop condition used only one possible action (i.e., click with a mouse), and thus could not commit wrong actions. Therefore, the Desktop condition was not compared to the VR conditions for wrong action errors. Instead, wrong action errors were indicative of differences in the VR conditions related to the system interaction. The proportion of wrong actions differed significantly depending on training condition, $t(49)$=3.045, p=.004, d=0.853. VR Gesture training had significantly more wrong action errors as a proportion of total attempts (M=19.13% of attempts were wrong action errors, SD=11.47%) than the VR Voice condition (M=9.91%, SD=10.07%). This analysis shows that the kind of interaction in VR training led to different errors, because the VR Gesture condition made proportionally more usability-related errors (i.e., using the wrong gesture action). Additionally, choosing the wrong action during the last phase of training was not related to later recall of the procedure, $r(83)$=-.039, p>.05, so there was no indication that more action-related errors during training hurts recall.

The proportion of wrong parts differed significantly depending on training condition, $F(2, 73.40)$=3.489, p=.036, η_p^2=.074. Levene's test was significant (p=.04), so the Brown-Forsythe correction was applied. The Desktop training group made significantly more wrong part errors as a proportion of total attempts in the final phase (M= 10.87% of attempts were wrong part errors, SD=11.35%) than the VR Voice condition (M=4.80%, SD=6.42%, d=0.638). The proportion of wrong part errors in the VR Gesture condition (M=7.45%, SD=7.93%) was between that of VR Voice (d=0.367) and Desktop (d=0.343), though it was not significantly different from either. Finally, part errors during the last training phase were related to lower recall in training overall (r[83]=-.412, p<.001), supporting the idea that choosing the wrong part during training reveals a lack of procedural understanding. Analyzing the proportion of wrong part errors gives insight into why the Desktop condition did not improve in the final training phase like the VR conditions; Desktop training led to wrong attempts that were proportionally more procedural errors.

Unlike wrong part and wrong action errors, the proportion of wrong tools selected to attempts made in the last training phase did not differ significantly depending on training condition, $F(2, 80)$=1.587, p=.211, η_p^2=.038. This lack of difference in wrong tool errors was likely a floor effect, because there were only three tools to choose from during the procedural training, whereas there were many different parts involved in every step of the procedure. Therefore, there were not enough instances of tool selection to detect meaningful differences among training conditions. However, making more wrong tool selections in the last training phase was related negatively to recall, $r(83)$=-.430, p<.001. Just as choosing the wrong part may indicate a lack of learning during training, making wrong tool selections also reflects difficulty in learning the procedure without instructional scaffolding.

Time, Procedural Errors, and Written Recall

The error analysis revealed that the Desktop condition performed worse on the last training phase and committed proportionally more procedural errors, whereas the VR Gesture condition generally improved over time and committed proportionally more usability errors. Additionally, time to complete the last phase was correlated negatively with written recall for the Desktop training, *r*(83)=-.531, *p*=.002. However, the question remains whether time in training or procedural errors were more predictive of subsequent procedural recall. A greater proportion of part errors could indicate lack of procedural understanding during training and lead to lower recall of the procedure, or it could be that time in training is more important for recall.

To better understand the relationships among time, errors, and recall, we conducted a post-hoc mediation analysis for the Desktop condition. Time was the predictor variable, part errors was the mediating variable, and recall was the outcome variable. The analysis indicated that the negative relationship between time and recall was fully mediated by part errors, such that participants in the Desktop group who spent a longer time in the final phase of the training tended to show lower recall because they were committing more part errors (the overall model was significant; R^2=.32, *p*=.004, $R^2_{mediation}$=.16, β =-.29 [95% CI -.69, -.10]; see Figure 3). In essence, committing more part errors is associated with poorer recall performance, regardless of how much time is spent in training.

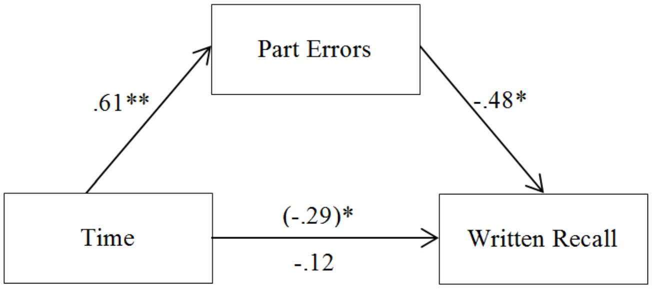


Figure 3. Conceptual diagram of mediation analysis for Desktop group performance, where Time is the predictor variable, Part Errors is the mediating variable, and Written Recall is the outcome variable. Standardized regression coefficients (β) are provided, with the indirect effect of time on recall in parentheses. *Note.* The indirect effect of time on recall was estimated using a bootstrapping procedure with 5,000 samples. *n*=32; **p*<.05, ***p*<.01

The results of the mediation analysis suggest that time should be ignored as a benchmark for performance and training quality. Simply completing training more quickly does not directly predict improved recall performance. Instead, it is more important to consider the kinds of errors participants make during training. The Desktop group committed proportionally more part errors. Not only do proportionally more part errors relate to lower recall in the Desktop condition as shown in the mediation analysis, part errors during the last training phase were related to lower recall in training overall. These part errors represent errors in learning during training (i.e., participants picked the wrong part because they failed to learn which part was to be used at what step in the procedure). In contrast, action errors during the final phase were not related to written recall for overall training, and do not necessarily represent errors in learning, but instead represent potential usability issues with the system. Thus, to judge the effectiveness of the training based on time and recall alone is insufficient – the kinds of errors participants make more strongly predict learning outcomes.

DISCUSSION

The objective of this experiment was to compare the effectiveness of VR training to Desktop training for learning a procedural task to determine whether learning outcomes were related to differences in level of immersion (i.e., 2-D or 3-D presentation) or amount of interactivity (i.e., gesture commands, voice commands, or clicking a mouse). We hypothesized that VR training overall would lead to more recall of the procedure because it is more immersive, and specifically that gesture-based commands in VR would be most beneficial because enacting the procedure would help participants to learn the task better; however, we did not find differences in number of steps recalled from the procedure depending on training condition. We did not find an enactment effect on procedural recall from type of interaction, nor did we see a direct relationship between level of immersion and recall. We may not have seen an enactment effect because the written recall measure may not have been sensitive enough to detect such differences and/or the gesture-based commands in our experiment did not match the learning material closely enough. For example, VR gestures were large, static gestures that represented each action, but these gestures were not exact pantomimes of the real-world actions they were intended to represent.

Feelings of presence and usability, while fairly high overall, also did not differ depending on training. The subjective measure of presence did not differ even though 3-D environments are theorized to be a more immersive than 2-D environments. This may be due to limitations in measuring presence subjectively (Schroeder, Bailey, Johnson, & Gonzalez-Holland, 2017), or that once a technology meets a threshold of usability, presence is felt regardless of 2-D or 3-D presentation. Usability ratings also did not differ depending on training groups, so perhaps usability is confounded with presence as discussed in Schroeder et al. (2017). Even though the subjective measure of presence did not differ by training type, there were differences between VR and Desktop groups in performance during training. Because the immersive environment mattered for training performance, it may be that our subjective measure of presence did not capture the psychological construct of immersion imposed by the differing technologies. To investigate why these theoretically hypothesized differences were not found, we conducted a systematic series of analyses to determine whether the training itself differed between groups.

We tested whether time to complete training differed among the groups. We found that Desktop training was faster overall than the two VR conditions by about eight minutes. The VR Voice group was about four minutes faster than the VR Gesture group, but this was not a significant improvement. Based on the fact that time to complete the Desktop training was fastest and there were no differences in recall between the training groups, an initial interpretation would be that Desktop is more efficient training than VR for learning a procedural task; however, a deeper look at the performance within the training phases showed that Desktop performance got worse in the last training phase when instructions were taken away. An analysis of the number of attempts it took to complete the procedural task showed that both VR groups improved with more training, while the Desktop group started out with fewer attempts in phases with instructional scaffolding, only to make more errors in later training when scaffolding was not given. This drop in performance at the end of training for the Desktop condition was associated with errors that negatively impacted the later recall of the procedural task.

Performance during training improved for VR conditions but not for Desktop training, indicating that participants in the immersive training conditions continued to improve when instructional scaffolding was removed, while those in the non-immersive condition did not. To determine why the Desktop condition needed more attempts to complete the procedure during the last training phase in which help was removed, we analyzed what kinds of errors were made during training and if those errors varied by training type. Specifically, we examined the proportion of errors to attempts made during the last training phase (no scaffolding) to answer whether errors in performance were primarily usability related (i.e., using the wrong gesture to complete a step) or were procedural errors (i.e., selecting the wrong part or tool for a step). Analyses showed that the VR Gesture condition made proportionally more wrong action errors because they were using the wrong gesture to complete a step, while the Desktop condition made proportionally more wrong part selection errors. Proportion of wrong tool errors was not different depending on training type because there were too few instances of tool selection during the procedure to see any meaningful differences. By analyzing the kinds of errors made during training, we were able to see that during the last training phase in which performance got worse for the Desktop condition, participants were making proportionally more errors associated with not knowing the procedure. Less-immersive training environments, therefore, may be less effective in that the kinds of errors made during training are more egregious for learning. One explanation could be that the Desktop condition was simply clicking around during training at the expense of processing the information, because clicking a mouse may not require much effort, unlike gestural commands. Another explanation could be that the 2-D desktop training was a more impoverished visual environment relative to the 3-D VR conditions (e.g., the “vivid” factor of immersive technology described by Slater & Wilbur, 1997), which may aid in acquiring more visual information (e.g., depth cues) about the learning material.

Although the Desktop condition initially seemed more efficient than VR training because participants completed the training faster, the final mediation analysis dispelled the idea that Desktop training was more effective. The Desktop condition made errors in the final stage of training that were related to lower recall of the procedure, and time was no longer predictive of better recall when these errors were taken into account. Completing the training more quickly did not directly relate to recall, but kinds of errors made during training did predict recall. Usability errors, such as using the wrong gesture command, did not indicate lower recall of the procedure. Although usability errors differed depending on interactivity with the training system, issues with system interactions did not seem to be related to learning in that recall was not associated with usability errors made during training. This suggests that making procedural errors during training represent worse learning in training and are detrimental for later recall, unlike usability errors.

Limitations and Future Research

A limitation of the current experiment was using a written measure of recall for the procedural task. The ideal recall measure would be for participants to perform the procedural task of replacing the alternator on a physical E-28 arresting gear. Due to practical constraints, having participants complete the recall task on the physical machinery was not feasible. Future research could include a more realistic procedural recall task, such as using 3-D-printed replicas of the machinery that are not readily available. Another limitation that may have diluted the interaction manipulation was that the gesture-based commands in the VR Gesture group were large and static. The gestures represented the procedural actions, but they were not an exact pantomime of the real-world action. Additionally, each gesture triggered an animation of the step, but the gesture to computer action timing was not a 1:1 mapping. Because the gestural interactions were not fully pantomimic, the naturalness of the interaction may have been diluted such that the gestures were more symbolic and harder to remember or perform. Future research could include more minute and dynamic gestures. Gestures that more closely represent the physical task may be a stronger manipulation, and results may show bigger effects of gesturing on learning as the enactment is closer to the learning material.

CONCLUSION

Overall, we did not see differences in recall of the procedure depending on training group, but analyses of performance during training gave insight into how VR and Desktop systems differ in terms of training efficacy. Less-immersive 2-D desktop training may be faster, but it may lead to more detrimental errors when instructional scaffolding is taken away. Additionally, 3-D VR training performance differed depending on type of interaction, with the VR Gesture group making more usability errors. Those interaction issues (usability errors), however, did not reduce later procedural recall. Furthermore, it may be possible to mitigate these usability errors related to VR interaction as technology improves user experience in VR. As gesture-based interactions in VR systems more closely mimic the physical world, usability may improve and participants may make fewer action errors (e.g., using the wrong gesture). Once usability issues are mitigated, the theoretically-predicted enactment effect may emerge.

This research addresses important gaps in the literature with regard to the efficacy of using VR for training by directly comparing its instructional effectiveness to an informationally-equivalent desktop-based training control. Furthermore, two different interaction methods within VR also were compared to test leading theories on why VR may be effective for training. Although there were no significant differences between conditions on learning gains, more systematic research like this is needed in this domain to better understand the relationships between VR training applications, learning outcomes, presence, usability, and other variables. Such an understanding is pivotal to begin to develop useful guidelines for developing effective VR-based training and determining whether VR-based training is actually a cost-effective training solution.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. James Sheehy who sponsored this work through the Section 219 Naval Innovative Science and Engineering Basic and Applied Research program. We would like to thank the Advanced Gaming Interactive Learning Environment (AGILE) team, and especially Courtney Mcnamara and Matthew Proetsch, who created the experimental testbed. The views expressed herein are those of the authors and do not reflect the official position of the organizations with which they are affiliated.

REFERENCES

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction, 24*(6), 574-594.

Blade, R. A., & Padgett, M. L. (2002). Virtual environments standards and terminology. In K. M. Stanney (Ed.), *Handbook of Virtual Environments* (pp. 15-27). Mahwah, NJ, USA: Lawrence Earlbaum Associates.

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

Buttussi, F., & Chittaro, L. (2017). Effects of different types of virtual reality display on presence and learning in a safety training scenario. *IEEE Transactions on Visualization and Computer Graphics*.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition, 106*(2), 1047-1058.

Engelkamp, J., & Jahn, P. (2003). Lexical, conceptual and motor information in memory for action phrases: A multi-system account. *Acta Psychologica*, *113*, 147-165.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Los Angeles: Sage.

Ganier, F., Hoareau, C., & Tisseau, J. (2014). Evaluation of procedural transfer from a virtual environment to a real situation: A case study on tank maintenance training. *Ergonomics, 57*(6), 828-843.

Hamblin, C. J. (2005). *Transfer of training from virtual reality environments* (Unpublished doctoral dissertation). Wichita State University, Wichita, KS.

Loftin, R. B., Kenney, P. J., Benedetti, R., Culbert, C., Engelberg, M., Jones, R., ... & Saito, T. (1994, November). Virtual environments in training: NASA's Hubble space telescope mission. In *Interservice/Industry Training Systems & Education Conference* (pp. 1-10).

Mikropoulos, T. & Natsis, A. (2011). Educational virtual environments: A ten-year review of empirical research (1999-2009). *Computers & Education, 56*, 769-780.

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.

Nilsson, L., Cohen, R. L., & Nyber, L. (1989). Recall of enacted and nonenacted instructions compared: Forgetting functions. *Psychological Research, 51*(4), 188-193.

Schroeder, B. L., Bailey, S. K. T., Johnson, C. I., & Gonzalez-Holland, E. E. (2017). Presence and usability do not directly predict procedural recall in virtual reality training. In Stephanidis C. (Ed.), *Communications in Computer and Information Science: Vol. 714. HCI International 2017 - Posters' Extended Abstracts*

Sheridan, T. B. (1992). Musings on telepresence and virtual presence. *Presence: Teleoperators & Virtual Environments, 1*(1), 120-126.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603-616.

Stevens, J. A. & Kincaid, J. P. (2015). The relationship between presence and performance in virtual simulation training. *Open Journal of Modelling and Simulation, 3*(2), 55270

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