Evaluation of Wearable Simulation Interface for Military Training

Grant S. Taylor, University of Central Florida, Orlando Florida, and **John S. Barnett**, U.S. Army Research Institute, Orlando, Florida

Objective: This research evaluated the training effectiveness of a novel simulation interface, a wearable computer integrated into a soldier's load-bearing equipment.

Background: Military teams often use game-based simulators on desktop computers to train squad-level procedures. A wearable computer interface that mimics the soldier's equipment was expected to provide better training through increased realism and immersion.

Method: A heuristic usability evaluation and two experiments were conducted. Eight evaluators interacted with both wearable and desktop interfaces and completed a usability survey. The first experiment compared the training retention of the wearable interface with a desktop simulator and interactive training video. The second experiment compared the training transfer of the wearable and desktop simulators with a live training environment.

Results: Results indicated the wearable interface was more difficult to use and elicited stronger symptoms of simulator sickness. There was no significant difference in training retention between the wearable, desktop, or interactive video training methods. The live training used in the second experiment provided superior training transfer than the simulator conditions, with no difference between the desktop and wearable.

Conclusion: The wearable simulator interface did not provide better training than the desktop computer interface. It also had poorer usability and caused worse simulator sickness. Therefore, it was a less effective training tool.

Application: This research illustrates the importance of conducting empirical evaluations of novel training technologies. New and innovative technologies are always coveted by users, but new does not always guarantee improvement.

Keywords: simulator, training, computer interface, usability, training effectiveness, training transfer, wearable simulation interface

Address correspondence to Dr. John Barnett, U.S. Army Research Institute, ATTN: DAPE-ARI-IF, 12350 Research Parkway, Orlando FL 32826-3276; e-mail: john.barnett1@us.army.mil.

HUMAN FACTORS
Vol. 55, No. 3, June 2013, pp. 672-690
DOI:10.1177/0018720812466892

INTRODUCTION

There is considerable interest in the military training community related to the use of computer games as simulators for training. Modifications of game engines can replicate realistic environments with the user's avatar performing realistic tasks. The virtual environments and avatars can simulate the performance of certain tasks with enough realism that users can utilize them to learn, practice, and improve skills (Seymour, 2008; Witmer, Bailey, & Knerr, 1995).

Simulators have certain advantages over live training. Tasks that are normally trained in a live setting (range, field, or using actual equipment) can often be trained in simulators at reduced cost. Simulated environments do not have the same scheduling, safety, transportation, or logistic concerns that live training ranges have. Simulated environments can also be modified at far less cost than traditional training ranges and provide a setting to safely practice tasks that would be too dangerous for live training. Thus, although simulated virtual training environments cannot replace live training, they are sometimes more appropriate for certain training situations than live training.

A review conducted by Knerr (2007) analyzed the need for and expected benefits of dismounted soldier training using simulators in virtual environments. One of the recommendations of this review was to evaluate the effectiveness of fully immersive simulators compared to desktop interfaces for dismounted infantry training. This article discusses a series of experiments conducted for this purpose. The evaluated system was a wearable computer interface that simulates the weapon and load-bearing equipment an individual soldier would wear and use in the field. The wearable interface used body motion as input and allowed the soldier to interact with the virtual environment in a more natural way than using a desktop interface. These features have led previous users and training administrators to assume that the

wearable interface provided training transfer superior to that from an equivalent desktop interface (Knerr, Garrity, & Lampton, 2005). This assumption was based on support from the identical elements theory (Holding, 1965; Thorndike & Woodworth, 1901), which states that training transfer is based on the degree to which the stimuli and responses utilized in training match those of the final performance environment. The wearable interface did indeed provide a training environment that more closely matched the performance environment; however, its training effectiveness relative to desktop simulators or traditional classroom or field training had not been empirically evaluated.

To assess the effectiveness of the wearable interface, a usability evaluation and two training transfer experiments were conducted. The usability assessment was a heuristic evaluation to identify the relative usability of the wearable and desktop interfaces, as well as to identify any usability concerns that might detract significantly from the training utility of the interface. The two experiments compared the ability of the wearable interface to train military tasks with both a desktop interface and a nonsimulator control. The first experiment assessed the retention of declarative knowledge, while the second experiment focused on training transfer of procedural skills.

Background

Heuristic usability evaluations. A heuristic evaluation is a means of considering a product or design to determine if it follows standard usability criteria. Its purpose is to find the most salient human-system interaction discrepancies (between the design and accepted usability practices), either for the evaluation of a system prior to its implementation or to guide the development team throughout the iterative design process. The process is designed to be easy to use, quick, and inexpensive, unlike more in-depth usability analyses that can be complex, expensive, and time-consuming. Nielsen refers to it as "discount usability engineering" (Nielsen & Mack, 1994, p. 25).

In a heuristic evaluation, a number of evaluators interact with the product and evaluate it against a set of heuristic usability criteria, which serve as a framework for the evaluation. The heuristics were derived from a factor analysis of 249 usability problems (Nielsen & Mack, 1994). For a thorough description of the usability heuristics, see Nielsen (1993).

Simulation training. A key question in simulator training and the use of virtual environments (VE) is how realistic must practice be to improve performance (Dorsey, Russell, & White, 2009). Ideally, the procedures practiced in the simulator should be identical to those for the real environment. However, for practical reasons this is not always feasible. For example, when a game-based simulator is used with a desktop computer interface, the trainee uses a keyboard and mouse to perform his or her actions rather than the physical movements normally required for the skill. Is the student in this simulator still learning?

The answer to this question often depends on the type of skills to be learned. Motor skills involve bodily movement and fine muscle coordination. Cognitive skills involve remembering procedures required to perform a task and sometimes problem solving. Learning motor skills through simulation requires the simulation to be an accurate representation of the physical operation of the real-world system. On the other hand, learning cognitive skills requires the learner to remember and think through the correct procedures, while the exact physical movements are less important (Wickens, 1992).

While a trainee in a simulator may not be performing the motor tasks the skill requires, he or she is typically performing the cognitive procedural tasks and therefore may be improving their performance with the skill. A simulator that allows the trainee to use relevant motor movements in training may improve training transfer if these movements are relevant to the skills being learned.

Immersion in simulator training. A VE that has a greater sense of immersion should produce higher levels of presence, the subjective feeling of being in one environment when actually being in another (Knerr et al., 1998). While immersion is primarily a mental state, the physical analog is fidelity. Fidelity is composed of three subcategories: physical, functional, and psychological fidelity (Hays & Singer, 1989). Physical fidelity

describes the extent to which a simulator provides a sensory experience (e.g., visual displays, auditory signals, physical controls, etc.) for trainees that matches the intended environment. Functional fidelity is determined by the simulator's ability to react appropriately to actions triggered by trainees. Psychological fidelity describes the extent to which the simulator induces the appropriate psychological response (e.g., fear, stress, engagement, etc.) in trainees.

Although it is logical to believe that a simulator with high fidelity will train better than a lower fidelity system, research has shown that this is not always true (Wickens, 1992). In some cases, the added realism of high-fidelity simulators may not provide enough training improvement to justify the increased costs. In other cases, simulators with high fidelity but that are not an exact match to the simulated system can force users to learn simulator-unique actions that are incompatible with the real system. These simulator-unique behaviors can actually interfere with the learning of skills needed for the real system. Wickens (1992) suggests it is important to know which components of training have to be similar to the target task and which are less important to learning.

The use of wearable simulators for dismounted soldier training is a relatively recent development. Initial studies investigating their effectiveness found that although early systems did allow soldiers to perform basic infantry tasks, they were too bulky and lacked the fidelity in their visual and weapons systems necessary to be truly useful (Lockheed Martin, 1997; Pleban, Dyer, Salter, & Brown, 1998). Over the past decade, simulation technology has continued to advance, and researchers have continued to investigate their usefulness for the training of dismounted soldiers (see Knerr, 2007). However, this research has been limited, and the research that has been done has primarily revolved around subjective questionnaires to assess users' perceptions of the system, rather than objective measures of their training effectiveness (Knerr et al., 2005).

Research Goals

Although prior research has found subjective opinions to support the use of wearable simulator interfaces, an empirical assessment was necessary to definitively evaluate their effectiveness. A number of factors were considered when validating the training ability of this novel interface. Of course, training performance was the primary concern, but this can encompass multiple factors. For example, a system may offer no benefit for the training of basic declarative knowledge while significantly improving the training of procedural skills.

Beyond training performance, other secondary factors were also considered. Certain positive factors, such as motivation or presence, may make novel interfaces worthwhile even if they do not directly improve training. On the negative side, poor usability, simulator sickness, or excessive workload are all factors that can potentially negate improvements in training performance.

Cost, in terms of money or time, was also important. An interface that provides slightly improved training at a substantially greater cost will decrease long-term training efficiency (Wickens, 1992). Conversely, an interface that provides equivalent training at reduced cost would be preferred.

This series of evaluations sought to provide an empirical assessment of the use of wearable simulation interfaces for military training. The assessment began with a usability evaluation to determine the system's ease of use, which directly affects its utility as a training tool. Desktop computers have been commonplace for decades, even within the realm of simulated training environments, and therefore the designers of these systems have likely recognized and resolved any major usability concerns. Conversely, given the relatively recent development of the wearable interface, as well as its limited market, the designers of wearable interfaces have had less opportunity to recognize these usability problems, and therefore the wearable interface was predicted to exhibit more significant usability concerns than a standard desktop interface.

However, once identified, usability problems can often be overcome relatively easily. Therefore, two additional experiments were conducted to evaluate the larger issue of the system's training capabilities. The first experiment compared the wearable interface to other

current standards in their ability to train declarative knowledge, while the second focused on the transfer of procedural knowledge from training to a live environment. The two experiments also evaluated critical secondary factors, such as simulator sickness, motivation, presence, and workload. For both experiments, the increased physical fidelity of the wearable interface was expected to improve training effectiveness. The wearable interface was also expected to improve presence and motivation by providing a more immersive experience for trainees. However, the use of the wearable interface's head-mounted display was expected to increase feelings of simulator sickness.

STUDY 1: USABILITY EVALUATION

Method

Evaluators. Eight evaluators (seven male, one female; age: $M = 33.5$, $SD = 12.01$) analyzed both the desktop and wearable versions of the simulator. Five were graduate students familiar with usability principles; two were applied psychologists, also familiar with usability principles; and one was a U.S. Army officer. Six of the evaluators had used the GDIS (Game Distributed Interactive Simulation) desktop system before, and three had experience with the wearable version.

Software. The simulation software, GDIS, was an immersive virtual environment developed by the Research Network Institute as a modification of the Half-Life graphics engine developed by Valve (Figure 1). All human characters not controlled by research participants (e.g., enemy soldiers, civilians, etc.) were controlled automatically by the GDIS system. All activities in GDIS took place in a virtual representation of the McKenna Military Operations in Urban Terrain (MOUT) training site located in Fort Benning, GA.

Desktop simulator. The desktop simulator was a Dell XPS computer, with a 2.66 GHz Intel Core 2 Duo CPU, 4 GB of RAM, an NVIDIA GeForce 8800 GTX graphics card, and a 20" LCD monitor with a 16:10 aspect ratio. A standard keyboard and optical mouse were used for controls, and headphones were used to hear sounds from the simulated environment. The controls used for the simulation were typical of other PC-based first-person shooters, with the W, S, A, and D keyboard keys controlling the avatar's movement, and the mouse controlling their view and the aim of the weapon.

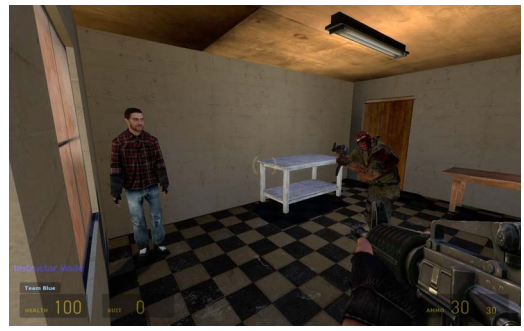

Figure 1. Participant's view within GDIS (Game Distributed Interactive Simulation).

Wearable simulator. The wearable simulator was an ExpeditionDI system developed by Quantum3D. The system consisted of a Thermite 1300 Tactical Visual Computer (1.4 GHz Intel Pentium processor, 1 GB RAM, ATI Mobility Radeon X300 graphics), which was worn on the back of a load-bearing vest. A helmet-mounted eMagin Z800 SVGA OLED visor provided two displays (one for each eye), each with 800 × 600 resolution with a 40° (diagonal) viewing angle. The displays were large enough to provide an immersive experience but still small enough not to completely occlude the wearer's view of their immediate surroundings, allowing them to maintain balance and avoid collisions through peripheral vision. The system was fully self-contained and the user was not tethered to any external equipment. The user's movements were tracked via three tri-axis motion sensors connected to the head (helmet), simulated weapon, and thigh. The user controlled their avatar through a combination of their own natural movements along with a small joystick and series of buttons on the front handgrip of the simulated M4A1 rifle (Figure 2). The user's head movements were used to control their view within the simulation, movement of the simulated weapon controlled the position and aim of the virtual weapon, and the leg tracker detected the user's posture (standing or

Figure 2. U.S. Army soldier wearing the ExpeditionDI wearable simulator interface (a). The front handgrip controls from the simulated M4A1 rifle (b).

kneeling) to adjust the avatar's position accordingly.

Procedure. Evaluators were welcomed by the experimenters and then given an overview of the evaluation with definitions of the 10 heuristic principles (Nielsen, 1993; see Table 1) and allowed time to become familiar with them. Next, they were introduced to either the desktop or wearable simulator (with the system order counterbalanced) and briefed on the controls. They were allowed to practice with the controls until they became familiar with them. As part of the control familiarization, the experimenter asked them to perform a list of actions and prompted them on which control to use if necessary.

Once the evaluators were ready, the experimenter guided them through a scenario by asking them to perform a series of tasks. When necessary, the experimenter would provide guidance on how to complete the action. The scenarios were designed to incorporate all of the functions the simulator could perform related to common military tasks such as movement, observation, target engagement, and communication.

When the evaluators completed the scenario, they were asked to record their evaluation on a questionnaire. The questionnaire asked them to rate the simulator on each of the 10 heuristic principles using a 5-point Likert scale. The questionnaire also had space to discuss specific usability concerns. Evaluators were encouraged to report all usability concerns that they encountered.

Evaluators then followed the same procedure with the other simulator, using a different scenario. The two scenarios included the same tasks but in a different order and followed a different path through the environment. The order of the two scenarios was also counterbalanced independently of simulator order. After completing the second scenario, the evaluator again responded to the same usability questionnaire.

After all evaluators completed this process, the experimenters analyzed the responses, aggregated similar comments, and identified 24 unique usability concerns. Due to the time

TABLE 1: Mean (and median, in parentheses) Reviewer Ratings of the Usability Heuristics Using a Scale From 1 to 5

Usability Heuristic	Desktop	Wearable	Wilcoxon *Z*	*p* (two-tailed)
Recognition rather than recall	4.14 (4)	2.50 (2)	1.98	.048*
Help and documentation	4.00 (4)	2.66 (2)	0.44	.655
Visibility of system status	4.42 (4)	3.25 (3)	2.23	.026*
Error prevention	3.75 (4)	2.62 (2.5)	1.56	.119
Help users recognize, diagnose, and recover from errors	3.50 (3)	2.75 (3)	2.00	.046*
Aesthetic and minimalist design	4.62 (5)	4.25 (4.5)	1.13	.257
Consistency and standards	4.50 (5)	4.14 (4)	0.73	.461
User control and freedom	3.85 (4)	3.50 (3)	0.81	.414
Flexibility and efficiency of use	3.83 (4.5)	3.57 (4)	1.73	.083
Match between system and the real world	3.75 (4)	4.12 (4)	0.79	.429

Note. Larger numbers indicate better system performance. Items are ranked by the difference between the means of each group, with cases in which the desktop ranked higher than the wearable interface at the top.
*$p < .05$

required to consolidate all of the evaluators' comments, a follow-up survey allowing the evaluators to rate the severity of each identified usability concern was conducted online. The time delay between an evaluator's initial system evaluation and completion of the subsequent survey ranged from 5 to 16 days. Evaluators rated each usability concern on a scale from 0 to 4 for its frequency (how often the problem occurred), impact (difficulty in overcoming the problem), and persistence (would the problem endure over time as the user gained experience with the system), following guidance from Nielsen and Mack (1994).

Results

Heuristic ratings. The results of the reviewers' ratings of the usability heuristics for both systems are presented in Table 1. Ratings were made on a 5-point Likert scale, with higher values indicating better system performance. A series of Shapiro-Wilk tests determined that 8 of the 20 variables significantly deviated from a normal distribution ($p < .05$ in each case) and so Wilcoxon Signed Ranks tests (the nonparametric equivalent of a *t* test) were conducted to determine group differences between the two simulators within each of the usability heuristics. These tests determined that the desktop system was rated significantly better on the visibility, recognition, and error recovery heuristics ($p < .05$ in each case). The desktop system also received better average ratings on all other usability heuristics, except for match, though these results were nonsignificant.

Specific usability concerns. A total of 23 unique usability concerns were identified from the evaluators' responses. Of these, 11 applied to both systems, 9 were specific only to the wearable system, and 3 were specific only to the desktop simulator (Table 2).

Although the evaluation rated two simulator versions, the results highlight that there are actually three systems being examined: the wearable interface, the desktop interface, and the GDIS environment that both interfaces display. In fact, many of the concerns related to both interfaces were actually concerns with the GDIS environment. However, because a user cannot use the GDIS environment without an interface, or an interface without a virtual environment, it is appropriate to consider the usability concerns of the virtual environment to affect both the desktop and wearable interfaces.

In addition to the total number of usability concerns, it is important to consider the magnitude

TABLE 2: List of Usability Concerns Determined for Each System

Usability Problem	System	Average Rating
Aiming is difficult due to problems calibrating the weapon and interference between the weapon and display preventing holding the weapon in a proper firing position.	Wearable	3.67
Actions requiring the use of the four handgrip buttons (especially those that require combinations of buttons) are difficult to remember and require additional training.	Wearable	3.54
The handgrip controls on the wearable system are difficult to use when pressing buttons in combination, requiring exact timing for combination presses.	Wearable	3.46
It is difficult to determine cardinal direction.	Both	3.13
The system causes sweating, nausea, claustrophobia, and headache.	Wearable	2.88
The thigh tracker requires too specific of an angle to cause the avatar to kneel.	Wearable	2.83
The system is prone to brief freezes, low frame rate, and lag.	Wearable	2.50
The fact that some of the controls are natural movements makes those that are unnatural (i.e., moving with the thumbstick) seem awkward and unnatural.	Wearable	2.42
The weapon selection process is difficult, especially the speed required to select a highlighted weapon in order to activate it.	Both	2.38
Additional instruction is needed for inexperienced users.	Both	2.29
Throwing grenades accurately is difficult.	Both	2.17
Some of the display information is confusing/unnecessary/unrealistic (team affiliation, "health" value, crosshair, unlabeled ammo numbers).	Both	2.11
It is easy to fire the weapon accidentally.	Both	2.04
Pressing the "Windows" key (between "Alt" and "Control") switches to the desktop. Switching back to the simulation does not always load properly, requiring a restart.	Desktop	2.00
The simulated weapon works differently than real weapons (unable to switch between semi/auto, charging handle unused, no separate trigger for firing rifle grenades).	Wearable	2.00
No help system is provided (though it is arguable whether one should exist outside of providing basic control information).	Both	1.88
Modifying the controls requires connecting to an external keyboard and editing a text file. Not all button combinations are available.	Wearable	1.74
Mistakes can be easy to make by accidentally pressing a button.	Both	1.67
Direction of movement is directly tied to the direction you are looking.	Both	1.48
A printed control sheet is necessary to remember all of the controls.	Desktop	1.33
The amount of weapons/ammunition that can be carried is potentially unrealistic.	Both	1.25
The researchers/instructors provided necessary help/instruction throughout the session.	Both	1.17
Some of the controls do not work as initially expected (e.g., "O" for map/compass).	Desktop	1.08

Note. Ratings are the average of all reviewers' responses for the item's frequency, impact, and persistence, with higher numbers indicating more severe problems. Items are listed by the magnitude of this average rating.

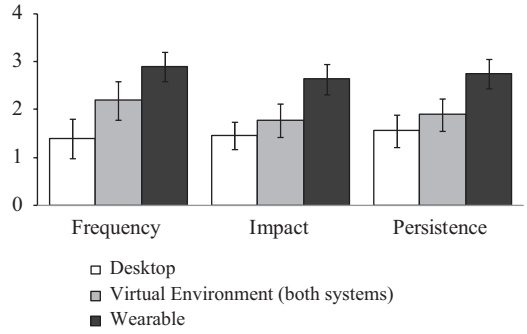

Figure 3. Ratings of the frequency, impact, and persistence of usability concerns related to each system. Higher values indicate more severe usability concerns.

of their ratings of frequency, impact, and persistence. These values were initially averaged to compute a single indication of the relative magnitude of each usability concern (Table 2). These values demonstrated that the usability concerns related to the wearable interface were not only the most prevalent, but also the most intrusive, with seven of the top eight concerns unique to the wearable interface.

Averaged ratings of frequency, impact, and persistence for the usability concerns of each system were also compared through repeated measures ANOVAs (Figure 3). A significant main effect was found for system type for frequency, $F(2, 14) = 27.653$, $p < .001$; impact, $F(2, 14) = 20.026$, $p < .001$; and persistence, $F(2, 14) = 15.281$, $p < .001$. Post hoc comparisons of frequency determined that the usability concerns for the wearable interface were rated significantly higher ($M = 2.90$, $SD = 0.414$) than were those from both the desktop interface ($M = 1.40$, $SD = 0.577$, $p < .001$, $d = 3.04$) and the GDIS software ($M = 2.20$, $SD = 0.637$, $p = .003$, $d = 1.33$), with the GDIS software also rated higher than the desktop interface ($p = .011$, $d = 1.33$). For impact, the usability concerns from the wearable interface were again rated significantly higher ($M = 2.68$, $SD = 0.502$) than those from both the desktop interface ($M = 1.50$, $SD = 0.563$, $p = .001$, $d = 2.21$) and the GDIS software ($M = 1.84$, $SD = 0.615$, $p < .001$, $d = 1.50$), with no significant difference between the desktop and GDIS ($p = .193$). The same pattern emerged from ratings of the persistence of the usability concerns, with the wearable interface again rated significantly higher ($M = 2.77$, $SD = 0.543$) than both the desktop interface ($M = 1.56$, $SD = 0.610$, $p = .003$, $d = 2.10$) and the GDIS software ($M = 1.95$, $SD = 0.587$, $p < .001$, $d = 1.46$), with no significant difference between the desktop and GDIS ($p = .155$).

For the sake of brevity, only the most poorly rated usability concerns of each system are presented. A more thorough discussion is provided by Barnett and Taylor (2010). For the wearable system, the most poorly rated usability concern described the difficulty users had aiming, resulting from inaccurate calibration and tracking of the weapon controller, as well as interference with the head-mounted display (HMD) when attempting to hold the weapon in a correct firing position. Evaluators also found the front handgrip controls cumbersome, with the button(s) required for each function difficult to remember given their arbitrary mapping.

The most poorly rated usability concern for the virtual environment was the difficulty determining cardinal direction. GDIS provides no dedicated compass; to find a heading the user must open an overhead map view, which has a small compass overlaid. The evaluators found it difficult to interpret their avatar's heading through this method, regardless of simulator interface.

The most problematic usability concern with the desktop interface was the result of accidentally pressing the "Windows" key (located between the "Control" and "Alt" keys on standard Windows-compatible keyboards). This key is easy to press inadvertently when using the "Shift" or "Control" key to run or crouch. Doing so closes GDIS to show the computer desktop and open the Windows Start menu, completely interrupting the virtual environment. Although an intrusive problem, this can be easily avoided by disabling the "Windows" key functionality in the system settings.

Discussion

Relative usability of the desktop and wearable simulators. As predicted, this evaluation demonstrated that the desktop interface was easier to use than the wearable system. When

the usability concerns of the desktop interface were combined with those of the GDIS environment, there were fewer and less severe concerns than for the wearable system combined with the GDIS environment. This suggests that soldiers who use the GDIS environment for training would find it easier with the desktop interface than with the wearable interface. Although this evaluation showed the wearable interface to be more difficult to use, this does not necessarily guarantee that it provides poorer training than a desktop computer. Studies 2 and 3 evaluated the training effectiveness of each system to determine exactly what benefit, if any, is achieved from the use of the wearable interface.

STUDY 2: RETENTION OF DECLARATIVE KNOWLEDGE

Method

Participants. Participants were university undergraduates who were compensated with course credit. A total of 98 students participated (66 males, 32 females; age: $M = 18.9$, $SD = 2.19$).

Procedure. Participants first completed a series of questionnaires, beginning with a simple demographics questionnaire used to collect their age, gender, dominant hand, and to ensure that they had normal sensory abilities and had no prior military experience. Participants then completed the Simulator Sickness Questionnaire (SSQ; Kennedy, Lane, Berbaum, & Lilienthal, 1993), short form of the Dundee Stress State Questionnaire (DSSQ; Matthews, Emo, & Funke, 2005; Matthews et al., 1999), Immersive Tendencies Questionnaire (ITQ; Witmer & Singer, 1998), and Game Experience Measure (GEM; Taylor, Singer, & Jerome, 2009). The GEM was modified to assess the participant's knowledge of first-person shooter games specifically, in addition to the original measure of general video game experience and knowledge. Following these questionnaires, the participants completed the Game-Based Performance Assessment Battery (GamePAB; Chertoff, Jerome, Martin, & Knerr, 2008; Taylor et al., 2009), a measure of their video game skill.

Participants were then trained on basic Army movement procedures, such as concealment techniques, firing positions, and correct grenade usage. These tasks were representative of the basic skills learned by all soldiers early in their training and were selected for use with the novice participants. This skill set included many tasks that required crouching, aiming, and shooting, all actions for which the wearable interface provided greater physical fidelity than the desktop interface.

Participants were randomly assigned to one of three training conditions: desktop, wearable, or Interactive Multimedia Instruction (IMI). Participants in the desktop and wearable conditions used the same systems described in the usability study. They were trained on the simulator controls and allowed approximately 5 minutes to practice on their own. They were then trained on the procedural tasks in the simulators. The training consisted of the participant's avatar following an avatar controlled by an experimental confederate (the trainer). The trainer explained and demonstrated each procedural task within the virtual environment and then prompted the participant to practice the task. The tasks were logically grouped into similar task groupings. The experimenter provided feedback on correct and incorrect performance throughout the training, which lasted roughly 20 minutes.

The IMI group used Interactive Multimedia Instruction videos that are currently in use by the Army to assist in the presentation of information from training publications. A total of three IMI videos were used: "Perform Movement Techniques During an Urban Environment," "Select Hasty Firing Positions During an Urban Environment," and "Employ Hand Grenades During an Urban Operation." Each contained a series of slides, which the trainee advanced through at their own pace using a mouse. Each slide presented information through text and recorded voice with animated images demonstrating the relevant principles.

Once the training was completed, all participants completed a series of posttest measures: SSQ, DSSQ, NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988), Presence Questionnaire (Witmer, Jerome, & Singer, 2005), and the interest/enjoyment and perceived

competence subscales of the Intrinsic Motivation Inventory (McAuley, Duncan, & Tammen, 1987).

Participants then completed the training retention test, in which they viewed a series of 14 videos, each 10 to 26 seconds, of an avatar performing the actions described in the training. These videos were created in OLIVE, a virtual environment similar to (but unique from) the training simulation environment. After each video, the participant was asked to describe both correct and incorrect procedural steps performed by the avatar by typing a response on the computer. In the assessment videos, most procedural steps were presented twice (once correctly and once incorrectly). The participant was graded on the number of correctly identified correct/incorrect actions performed in the video, out of a total possible score of 53. Errors, such as incorrectly identifying a procedural step as either correct or incorrect, were recorded and analyzed independently of the participant's correct responses.

After participants completed the training retention test, they were thanked for their time and allowed to leave. The entire experiment lasted 1.5 to 2 hours.

Results

Training retention. Training retention was measured as the total number of movement procedures correctly identified as either correct or incorrect from all of the videos. A reliability analysis was first conducted to ensure that this retention measure maintained an acceptable level of internal consistency. Cronbach's alpha was computed from the participants' scores on each of the 53 total movement procedures presented in the videos and was found to be sufficient ($\alpha = .742$). A one-way ANOVA found no significant differences between the training conditions ($p = .832$).

The frequency with which participants made errors in their responses was also evaluated. An error was defined as claiming that a specific aspect of the soldier's movement was correct when it was actually incorrect or claiming it was incorrect when it was actually correct. It was a relatively rare occurrence for a participant to commit an error, regardless of training condition, and again a one-way ANOVA revealed no significant group differences ($p = .116$).

Although no significant group differences were found, several significant correlations were found between training retention and other measured variables. Training retention was negatively correlated with the Disorientation subscale of simulator sickness, $r(96) = –.291$, $p = .005$, as well as both the Distress, $r(96) = –.248$, $p = .017$, and Worry, $r(96) = –.304$, $p = .003$, dimensions of the DSSQ. Training retention was also significantly correlated with GEM measures of Gaming Experience, $r(96) = .245$, $p = .018$, and First-Person Shooter Knowledge, $r(96) = .338$, $p = .001$. Significant correlations were also found with GamePAB measures of Time on Follow, which measures the participant's ability to follow a lead avatar within a narrowly defined distance, $r(96) = .528$, $p < .001$, and Posture Reaction Time, the latency of the participant's response to the lead avatar's posture changes, $r(96) = –.489$, $p < .001$ (note that the negative correlation implies faster reaction times and are related to greater Training Retention). Positive relationships were also found between Training Retention and the Interface Quality subscale of the Presence Questionnaire, $r(96) = .350$, $p = .001$, as well as the Perceived Competence subscale of the Intrinsic Motivation Inventory, $r(96) = .226$, $p = .029$.

Simulator sickness. A series of one-way ANOVAs reported significant group differences for each of the SSQ subscales (Figure 4; positive values indicate an increase from baseline): Nausea, $F(2, 91) = 7.146$, $p = .001$; Oculomotor, $F(2, 91) = 11.337$, $p < .001$; and Disorientation, $F(2, 91) = 6.815$, $p = .002$.

For the Oculomotor subscale, the wearable group reported a higher level ($M = 16.58$, $SD = 17.59$) than the desktop ($M = –0.245$, $SD = 14.04$, $p < .001$, $d = 1.06$) and IMI groups ($M = 5.62$, $SD = 9.97$, $p = .003$, $d = 0.795$), with no significant difference between the desktop and IMI groups ($p = .108$). For the Disorientation subscale, once again the wearable group reported a higher level ($M = 13.05$, $SD = 19.35$) than the desktop ($M = 0.898$, $SD = 12.42$, $p = .001$, $d = 0.765$) and IMI groups ($M = 3.14$,

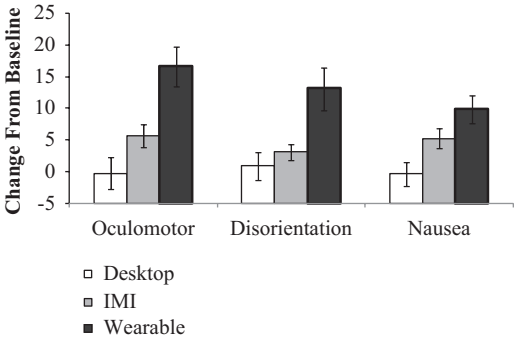

Figure 4. Simulator sickness results. IMI = Interactive Multimedia Instruction.

SD = 6.92, *p* = .006, *d* = 0.754), with no significant difference between the desktop and IMI groups (*p* = .527). The trend changed for the Nausea subscale, with the desktop group reporting lower values (*M* = –0.308, *SD* = 10.59) than the wearable (*M* = 9.84, *SD* = 12.23, *p* < .001, *d* = 0.889) and IMI groups (*M* = 5.23, *SD* = 8.83, *p* = .044, *d* = 0.570), with no significant difference between the wearable and IMI groups (*p* = .090).

Stress (DSSQ). A series of one-way ANOVAs found a significant effect for training condition on Task Engagement, *F*(2, 95) = 13.156, *p* < .001, with no significant effects on Distress or Worry. Post hoc tests determined that the IMI group reported significantly less Engagement (*M* = 3.97, *SD* = 5.27) than both the desktop (*M* = 10.25, *SD* = 4.20, *p* < .001, *d* = 1.33) and wearable groups (*M* = 8.37, *SD* = 5.38, *p* = .001, *d* = 0.827), with no differences between the desktop and wearable conditions.

Workload (NASA-TLX). One-way ANOVAs revealed a significant main effect of training condition for Physical Demand, *F*(2, 91) = 30.423, *p* < .001, with no effect on any of the other workload subscales. Post hoc comparisons showed the wearable group to have reported greater Physical Demand (*M* = 51.41, *SD* = 27.51) than the IMI (*M* = 9.20, *SD* = 12.09, *p* < .001, *d* = 2.13) and desktop groups (*M* = 17.53, *SD* = 24.93, *p* < .001, *d* = 1.29), with no difference between the IMI and desktop groups.

Presence. ANOVAs found significant group differences on each of the four presence subscales: Involvement, *F*(2, 92) = 28.505, *p* < .001; Sensory Fidelity, *F*(2, 92) = 13.716, *p* < .001; Adaptation Immersion, *F*(2, 92) = 6.971, *p* = .002; and Interface Quality, *F*(2, 92) = 5.285, *p* = .007.

Post hoc comparisons found the IMI group reported less Involvement (*M* = 43.71, *SD* = 12.01) than both the desktop (*M* = 61.72, *SD* = 10.18, *p* < .001, *d* = 1.62) and wearable groups (*M* = 62.22, *SD* = 10.94, *p* < .001, *d* = 1.61). The IMI group also experienced less Sensory Fidelity (*M* = 21.84, *SD* = 5.08) than did both the desktop (*M* = 28.91, *SD* = 6.88, *p* < .001, *d* = 1.18) and wearable groups (*M* = 28.31, *SD* = 5.64, *p* < .001, *d* = 1.21), as well as less Adaptation Immersion (*M* = 36.52, *SD* = 7.40) than both the desktop (*M* = 41.81, *SD* = 6.28, *p* = .005, *d* = 0.773) and wearable groups (*M* = 42.91, *SD* = 7.96, *p* = .001, *d* = 0.832). However, the IMI group did report higher levels of Interface Quality (*M* = 16.87, *SD* = 3.20) than the wearable group (*M* = 14.16, *SD* = 3.04, *p* = .002, *d* = 0.869), with neither group significantly different from the desktop group (*M* = 15.41, *SD* = 3.64).

Motivation. Only the Interest/Enjoyment and Perceived Competence subscales were included from the Intrinsic Motivation Inventory. Of these, one-way ANOVAs found significant group differences only in the Interest/Enjoyment subscale, *F*(2, 92) = 16.276, *p* < .001. Post hoc comparisons showed that the IMI group reported less Interest/Enjoyment (*M* = 3.659, *SD* = 1.29) than both the desktop (*M* = 5.147, *SD* = 1.51, *p* < .001, *d* = 1.06) and wearable groups (*M* = 5.563, *SD* = 5.56, *p* < .001, *d* = 0.556).

Discussion

Although no group differences were found for training retention, the correlations between training retention and the other measured variables, as well as the group differences found on several of the secondary variables, still serve to better understand the distinction between the simulators. For example, the wearable simulator evoked significantly higher levels of simulator sickness than the desktop interface on all of the subscales. Although the measured levels of simulator sickness were not excessively high, it

is important to consider that these levels were reached after only 20 minutes in the simulation. Beyond the impact on trainee well-being, the negative relationship between the Disorientation subscale of simulator sickness and training retention provides additional concern that the wearable simulator could potentially result in poorer training due to simulator sickness.

A positive relationship was also found between training retention and the Interface Quality subscale of the Presence Questionnaire. Given the relatively poorer usability of the wearable system (determined by the previous study), this is another issue that could negatively influence training with this system.

Although the wearable interface was found to elicit significantly higher levels of physical demand than both the desktop and IMI conditions, this is not necessarily detrimental to the wearable system. Given the challenging physical demands of most dismounted soldier tasks, it could be considered beneficial to have similar physical demands in the training environments to ensure soldiers maintain their conditioning. However, these physical demands could also limit the amount of time soldiers could spend in training scenarios before requiring rest.

Other group differences expose potential disadvantages of the traditional training methods used with the IMI group. The IMI group reported significantly less engagement and interest/enjoyment than both the desktop and the wearable groups, with no differences between the desktop and wearable groups. Although neither of these variables was found to be significantly related to training retention in this study, the concepts of engagement and interest have been considered important aspects in training for many years (Lepper, Woolverton, Mumme, & Gurtner, 1993; Matthews et al., 2007).

With the exception of the Interface Quality subscale, the IMI group consistently reported less presence than both the desktop and wearable groups, with no differences found between the desktop and wearable group. Presence is one aspect that has long been considered an important part of simulation-based training, but evidence of a direct link between presence and training retention from virtual environments has been weak (Jerome & Witmer, 2004; Mantovani & Castelnuovo, 2003). This still gives an additional advantage to both the desktop and wearable systems in that the increase in presence they provide should lead to an increase (albeit a minor one) in the knowledge retained by those who train with them.

Limitations. This study examined only the retention of declarative knowledge; that is, trainees memorized and recalled lists of correct and incorrect behaviors. The training did not address procedural knowledge, which is likely a more common use for simulated training environments. It is possible the IMI is better suited for training declarative knowledge, while the desktop and wearable interfaces provide better training of procedural skills.

Another limitation is the measure used to evaluate training. It consisted of students observing and evaluating the actions of others rather than performing the skills themselves. Performance on this written test may not necessarily be indicative of actual performance in the field. Study 3 resolved both of these limitations by evaluating each system's ability to train procedural skills that can be transferred to a realistic live setting.

STUDY 3: TRAINING TRANSFER OF PROCEDURAL SKILLS

Method

Participants. A total of 62 participants completed the study, with 20 in each of the desktop and wearable training conditions and 22 in the live condition. To match the Army's restrictions for soldiers conducting hostage rescue missions (the task to be trained), all participants were males between 18 to 30 years old (M = 20.27, SD = 2.128) and in good health. All participants were verified to have no prior military or ROTC experience to ensure they had no previous hostage rescue training. Participants were compensated with their choice of either course credit or $20 for their time.

Procedure. The study lasted for a total of 1.5 to 2 hours, with participants completing the procedures in groups of two. Upon arrival, both participants completed a series of initial questionnaires on a desktop computer. These questionnaires began with a standard demographics

Figure 5. Room used for all live scenarios (enemy/hostage targets and locations varied for each scenario). Pictured: hostage (left) and enemy targets (center and right).

Figure 6. A research assistant demonstrating the equipment used in the live condition, holding the replica M4 rifle and wearing vest with frag grenade (left) and flashbang (right).

form used to confirm that the participant's gender and age met the study requirements and that they had no prior hostage rescue experience. Participants then completed a baseline measure of the SSQ, as well as GEM and GamePAB.

The researcher then trained the participants on the proper military hostage rescue techniques for roughly 20 minutes within one of three randomly assigned training conditions (desktop, wearable, or live), with both participants working together as a team within the same environment. The desktop and wearable systems were identical to those used in Studies 1 and 2. Participants in both the desktop and wearable conditions received instruction within the GDIS environment from an avatar controlled by the researcher on a separate desktop computer, with all three avatars sharing the same virtual environment. Correct procedures were discussed verbally and also demonstrated by the researcher avatar. Those assigned to the live condition were trained in real rooms with life-size cardboard cutouts as enemies and hostages (Figure 5). They were provided with a replica M4 rifle as well as replica frag (fragmentation) grenades and flashbangs (stun grenades). Participants wore an ammo vest to carry the grenades and a helmet and goggles for safety (Figure 6). As with the other conditions, correct procedures were demonstrated by the researcher and explained verbally.

The trained techniques, a total of 24 individual steps, described the proper way to enter a potentially hostile room, the paths to take once inside the room, and how to respond to enemy targets. The missions required the participants to work together as a team. Most task steps were consistent for both team members, but each team member did have some specific responsibilities. Each participant was randomly assigned to one team role (1 or 2) before training began and maintained this role throughout training and testing.

Regardless of condition, the training consisted of four practice missions. For the first mission, the researcher walked the participants through each step of the mission, explaining the important task components along the way. For the remaining three training missions, the researcher observed as the participant team completed the missions on their own. Following each mission, the researcher provided feedback describing the correct and incorrect steps taken by the team.

After completing the training missions, the participants completed the SSQ again, as well as the Interest/Enjoyment and Perceived Competence scales of the Intrinsic Motivation Inventory. After the questionnaires, participants completed a testing phase in which they conducted four missions in live rooms under the same conditions as described for the live practice scenarios, though with no instruction or assistance from the researcher. The same room was used for each mission, with the location,

number, and type (enemy or civilian) of targets varying between missions, as well as the participants' direction of entry. The presentation order of the various mission conditions was counterbalanced across all participants. Performance was videotaped for later scoring of the participants' ability to correctly execute the procedures covered in the training. Following this testing phase, the study was complete and the participants were allowed to leave.

Results

Performance. The three training conditions were initially compared in terms of performance on the test scenarios. Performance was calculated as the percentage of task steps performed correctly for each scenario. The total number of task steps ranged from 20 to 24 for each mission due to the fact that not all steps were relevant for all mission configurations (e.g., the step "don't shoot civilians" was not relevant for missions in which no civilians were present). Seven task steps were identified for which the wearable interface provided a closer match to the live environment than the desktop interface. These tasks included the appropriate use of the weapon (e.g., "fire in a controlled pair") or physical movements (e.g., "crouch to move under windows"). This group of seven task steps (identified in the following as *matched tasks*) was analyzed separately from the remaining steps (*mismatched tasks*) to provide a more detailed investigation based on the identical elements theory of training transfer. In addition to the percentage of actions performed correctly, scenario completion time was also used as a dependent variable due to the critical importance of speed in the hostage rescue missions. The analyses were conducted using mixed-model ANOVAs with training condition (between subjects: desktop, wearable, or live) and scenario number (within subjects: first, second, third, or fourth) as independent variables.

Training condition was found to have a significant main effect on percentage correct for the mismatched tasks, $F(2, 59) = 4.950$, $p = .010$, but not for the matched tasks ($p = .244$; Figure 7). Pairwise comparisons determined that the live training condition performed significantly better on mismatched tasks

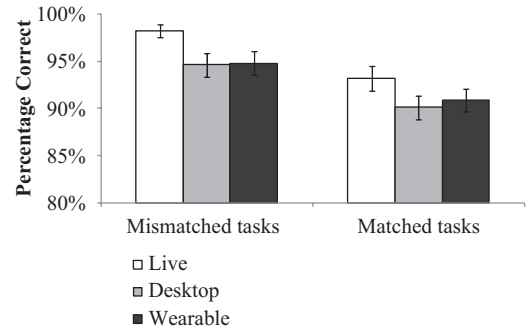

Figure 7. Percentage of actions performed correctly in the test scenarios. The number of mismatched tasks (steps for which the wearable interface did not provide an identical environment to the testing environment) ranged from 13 to 17, depending on the mission configuration. The total number of matched tasks (steps for which the wearable interface provided an identical environment) was 7.

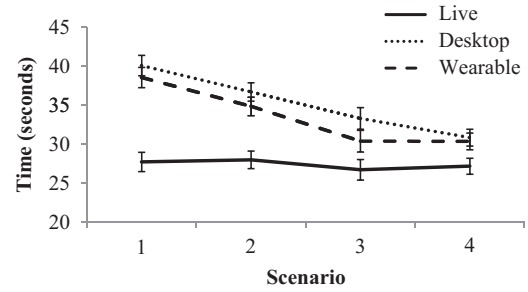

Figure 8. Scenario completion time for the four test scenarios by training condition.

($M = 98.2\%$, $SD = 3.27$) than both desktop ($M = 94.6\%$, $SD = 5.68$, $p = .008$, $d = 0.804$) and wearable ($M = 94.8\%$, $SD = 6.77$, $p = .011$, $d = 0.677$) training conditions, with no significant difference between the desktop and wearable conditions ($p = .902$). The main effect for scenario number as well as the Training Condition × Scenario Number interaction were not found to be statistically significant for either the matched or mismatched tasks ($p > .05$ in each case).

Training condition also had a significant main effect on scenario completion time, $F(2, 69) = 25.056$, $p < .001$ (Figure 8). Pairwise comparisons found the live training condition to perform the scenarios significantly faster

(M = 27.41 seconds, SD = 3.48) than both the desktop (M = 35.24 seconds, SD = 4.74, p < .001, d = 1.91) and wearable (M = 33.54 seconds, SD = 2.96, p < .001, d = 1.90) training conditions, with no significant difference between the desktop and wearable conditions (p = .161). The interaction between training condition and scenario number was also statistically significant, $F(6, 177) = 4.319$, p < .001. Subsequent one-way ANOVAs evaluated the effect of training condition on completion time of each scenario individually. These analyses found the live training condition to perform significantly faster than both the desktop and wearable conditions across all four scenarios (p < .05 in each case), but the strength of this effect diminished over time (Scenario 1: R^2 = .492; Scenario 2: R^2 = .352; Scenario 3: R^2 = .168; Scenario 4: R^2 = .110), with the live training condition's performance times remaining consistent as the desktop and wearable training conditions' performance times improved over time.

Questionnaires. The effect of training condition was also evaluated on the subjective ratings of simulator sickness and intrinsic motivation. For simulator sickness, each of the three subscales provided by the SSQ was obtained both before and after training. Pretraining baseline values were subtracted from posttesting values to calculate change scores for each subscale independently. These change scores were used as the dependent variables in a series of one-way ANOVAs, with training condition as the independent variable (Figure 9). A significant main effect for training condition was found for the Nausea subscale, $F(2, 59) = 7.640$, p = .001, with the wearable condition reporting significantly higher values (M = 18.60, SD = 30.71) than both the desktop (M = –0.477, SD = 3.76, p = .001, d = 1.11) and live training conditions (M = 0.000, SD = 4.16, p = .001, d = 1.07), with no significant difference between the desktop and live conditions (p = .931). The same trend was found for the Oculomotor subscale, $F(2, 59)$ = 13.192, p < .001, with the wearable condition reporting significantly higher values (M = 23.50, SD = 29.30) than both the desktop (M = 0.379, SD = 1.69, p < .001, d = 1.49) and live training conditions (M = 0.000, SD = 2.34, p < .001, d = 1.49), with no significant difference between the desktop and live conditions (p = .942). This trend was also present for the Disorientation subscale, $F(2, 59) = 4.144$, p = .021, with the wearable condition reporting significantly higher values (M = 21.58, SD = 49.78) than both the desktop (M = 0.000, SD = 0.000, p = .020, d = 0.867) and live training conditions (M = –1.27, SD = 5.94, p = .012, d = 0.820), with no significant difference between the desktop and live conditions (p = .886).

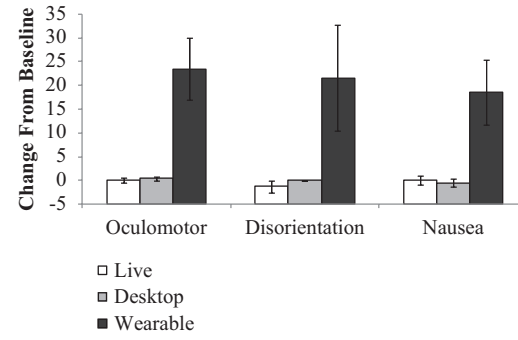

Figure 9. Simulator sickness values reported from each training condition. Values are reported as change from the baseline data collected prior to training, with positive values indicating an increase.

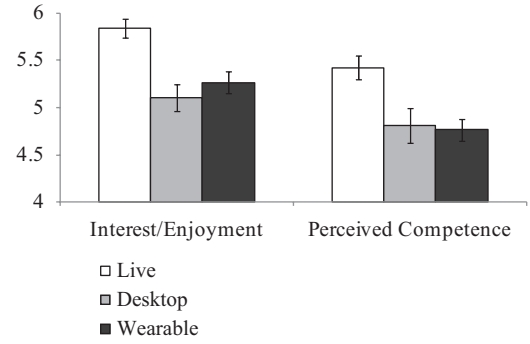

Figure 10. The Interest/Enjoyment and Perceived Competence subscales of the Intrinsic Motivation Inventory as reported from each training condition.

The effect of training condition was also evaluated on both the Interest/Enjoyment and Perceived Competence scales of the Intrinsic Motivation Inventory (Figure 10). A significant main effect of training condition was found for Interest/Enjoyment, $F(2, 59) = 11.021$, p < .001.

Post hoc comparisons determined that the live condition reported significantly higher values (M = 5.84, SD = 0.463) than both the desktop (M = 5.10, SD = 0.632, $p < .001$, d = 1.35) and wearable conditions (M = 5.26, SD = 0.510, p = .001, d = 1.19), with no significant difference between the desktop and wearable conditions (p = .338). Training condition was also found to have a significant effect on the Perceived Competence Scale, $F(2, 59) = 6.657$, p = .002. Post hoc comparisons again found that the live condition reported significantly higher values (M = 5.42, SD = 0.593) than both the desktop (M = 4.81, SD = 0.831, p = .004, d = 0.857) and wearable conditions (M = 4.77, SD = 0.517, p = .002, d = 1.17), with no significant difference between the desktop and wearable conditions (p = .842).

The influence of video game experience and skill (measured by GEM and GamePAB, respectively) on mission performance was also evaluated using standard Pearson correlations. A significant relationship was found between video game experience and scenario completion time, $r(60) = -.332$, p = .008, with those higher in experience performing the missions faster. A regression determined that this relationship did not vary as a function of training condition (p = .743). No significant relationship was found between video game experience and percentage correct or between the measures of video game skill with either percentage correct or scenario completion time ($p > .05$ in each case).

Discussion

Live training. One not particularly surprising finding is that live training is superior to virtual simulations for the learning of procedural skills. The results for both the percentage of actions performed correctly and the time to complete the scenarios showed live training to be superior to both simulation interfaces.

However, one confounding variable is that the live training condition trained in the same environment (only slightly modified) in which their performance was tested. The live training group had the advantage of not having to transfer their knowledge to a new environment during the testing phase. Therefore, they were more familiar with the surroundings, which likely improved both their speed and performance accuracy. As participants trained in the desktop and wearable simulators completed the four test missions in the live environment, their times improved, whereas the live control group's time scores about the same (see Figure 8). This suggests that as they became familiar with the live testing environment, the simulator groups were able to perform more quickly, though performance accuracy remained consistent.

Simulator training. The results also demonstrated there to be no significant differences between wearable and desktop interfaces, with the exception of simulator sickness symptoms. As in Study 2, participants who used the wearable interface reported significantly stronger symptoms of simulator sickness than either the desktop or live training condition. Although neither simulator condition trained as well as the live condition, both simulator conditions trained the procedural skills equally well. However, simulator training in general does seem to provide adequate training for procedural skills. The performance accuracy was high across all training conditions, averaging 93% to 97% depending on condition, indicating that all conditions provided acceptable training. The trend in time-to-complete for both simulator interfaces showed participants took less time to complete the live scenario each time it was performed. Although speculative, all groups would have had equivalent completion times on the fifth scenario if the trend had continued.

Gaming experience. Although prior video game experience improved trainees' speed, this effect did not vary as a function of training condition. Participants seemed to learn the simulator interfaces quickly and were able to operate them well enough to learn the target skills, regardless of video game experience.

GENERAL DISCUSSION

Summary of Findings

Contrary to the predicted hypotheses, the wearable interface did not provide significant advantages over the other types of training. The results of three independent evaluations found the wearable interface to perform no better than a desktop interface for simulation training. Participants who used the wearable system in

Study 2 performed equivalently to those who used the desktop interface and the IMI videos. In Study 3, those trained in the wearable interface again performed equivalently to those trained on desktop computers, though both groups were bested (by a slight margin) by those who received live training.

Other problems with the wearable interface would likely have a negative impact on training as well. For example, the usability of the wearable interface was not as good as that of the desktop interface, thus it would likely take more time to use the wearable for exercises than the desktop, as some training time would be committed to users having to overcome usability concerns. This additional time requirement, in conjunction with the higher cost of the system itself, results in a dramatic increase in the operational cost of the wearable interface relative to a desktop computer. The poorer usability of the wearable system may contribute to user frustration as well, resulting in a decreased long-term effectiveness by reducing trainees' motivation to utilize the system.

As predicted, simulator sickness scores were consistently greater for the wearable interface than all others tested. Workload measures also found users of the wearable system to report higher levels of physical demand. Although feelings of presence were generally greater for both the wearable and desktop than for the IMI videos, there was no difference between the wearable and desktop simulators. The wearable and desktop interfaces were rated equivalently in terms of motivation, with both rating higher than the IMI videos but lower than the live training.

Identical elements theory. Although the results of these studies may appear to contradict the foundation of identical elements theory, this is not necessarily the case. The simulated environment provided by the wearable interface more closely matched the real-world testing environment, and so identical elements theory would suggest that it should improve training. However, it is clear that there are additional differences between the wearable interface and the other interfaces evaluated beyond their fidelities. Most substantial of these differences is the fact that the wearable interface suffers from poorer usability and induces higher levels of simulator sickness. Both of these problems are issues that can be detrimental to training effectiveness and could offset any advantages provided by the system's improved fidelity. Considering these mediating factors, the results should not be considered to dispute identical elements theory.

CONCLUSION

Despite previous subjective reports predicting their benefits, the wearable simulator interface used failed to improve upon more traditional interfaces for any aspect of training across three separate evaluations. It should be noted that the wearable interfaces used were older models (current as of the beginning of the evaluation) and that newer models have since been released with enhanced capabilities that may mitigate some of the usability issues. As these interfaces evolve, it is possible that their usability and training utility may improve. Additionally, the use of novice participants limited the training content in both Studies 2 and 3 to relatively simple procedures, and so it is possible that the wearable interface could still prove beneficial when training more advanced tactics. However, the multiple shortcomings demonstrated through this series of evaluations clearly show that significant improvements are necessary before the wearable interface could even be considered to be equivalent to a desktop computer. More generally, the present research illustrates that although innovative technologies are often appealing, they are not necessarily an improvement over current standards. Therefore, any assumed advantages must be evaluated through empirical research prior to full fielding.

KEY POINTS

- A heuristic usability analysis rated the wearable interface poorer than a desktop computer for its option visibility (recognition rather than recall), error recovery, and visibility of system status.
- A greater number of specific usability concerns were found for the wearable interface than the desktop computer. The usability concerns related to the wearable interface were also rated as having greater frequency, impact, and persistence.

- The wearable interface was found to provide equivalent training retention to a desktop computer as well as a control condition using current standard training materials. The wearable interface elicited greater simulator sickness and physical demand than both of the alternative conditions.
- The wearable interface provided training transfer to a live environment equivalent to a desktop computer interface. Both simulated training environments were slightly inferior to live training.

REFERENCES

Barnett, J. S., & Taylor, G. S. (2010). *Usability of wearable and desktop game-based simulations: A heuristic evaluation* (ARI Study Note 2010-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Chertoff, D., Jerome, C., Martin, G., & Knerr, B. (2008). GamePAB: A game-based performance assessment battery application. *Proceedings from the 52nd Annual Meeting of the Human Factors and Ergonomics Society Meeting* (pp. 1570–1573). Santa Monica, CA: Human Factors and Ergonomics Society.

Dorsey, D., Russell, S., & White, S. (2009). Identical elements theory: Extensions and implications for training and transfer. In J. Cohn, D. Nicholson, & D. Schmorrow (Eds.), *The PSI handbook of virtual environments for training and education* (pp. 196–205). Westport, CT: Praeger Security International.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 159–185). Amsterdam: North-Holland.

Hays, R. T., & Singer, M. J. (1989). *Simulation fidelity in training system design*. New York, NY: Springer-Verlag.

Holding, D. H. (1965). *Principles of training*. New York, NY: Pergamon Press.

Jerome, C., & Witmer, B. (2004). Human performance in virtual environments: Effects of presence, immersive tendency, and simulator sickness. *Proceedings from the 48th Annual Meeting of the Human Factors and Ergonomics Society Meeting* (pp. 2613–2617). Santa Monica, CA: Human Factors and Ergonomics Society.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). A simulator sickness questionnaire (SSQ): A new method for quantifying simulator sickness. *International Journal of Aviation Psychology*, *3*, 203–220.

Knerr, B. W. (2007). *Immersive simulation training for the dismounted Soldier* (Study Report 2007-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knerr, B. W., Garrity, P. J., & Lampton, D. R. (2005, December). *Embedded training for future force warriors: An assessment of wearable virtual simulators*. Paper presented at the 24th Annual Army Science Conference, Orlando, FL.

Knerr, B. W., Lampton, D. R., Singer, M. J., Witmer, B. G., Goldberg, S. L., Parsons, K. J., & Parsons, J. (1998). *Virtual environments for dismounted soldier training and performance: Results, recommendations, and issues* (ARI Technical Report 1098). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools: Technology in education* (pp. 75–105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lockheed Martin Corporation. (1997). *Dismounted warrior network front end analysis experiments* (Advanced Distributed Simulation Technology II, Dismounted Warrior Network DO No. 0020, CDRL AB06, ADST-II-CDRL-DWN-9700392A). Orlando, FL: US Army Simulation, Training and Instrumentation Command.

Mantovani, F., & Castelnuovo, G. (2003). Sense of presence in virtual training: Enhancing skills acquisition and transfer of knowledge through learning experience in virtual environments. In G. Riva, F. Davide, & W. A. IJsselsteijn (Eds.), *Being there: Concepts, effects and measurement of user presence in synthetic environments* (pp. 167–181). Amsterdam, the Netherlands: Ios Press.

Matthews, D. E., VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., Y Rosa, A. C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*(1), 3–62.

Matthews, G., Emo, A. K., & Funke, G. J. (2005, July). *A short version of the Dundee Stress State Questionnaire*. Paper presented at the 12th meeting of the International Society for the Study of Individual Differences, Adelaide, Australia.

Matthews, G., Joyner, L., Gilliland, K., Campbell, S., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire: Towards a state "big three"? In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7., pp. 335–350). Tilburg, the Netherlands: Tilburg University Press.

McAuley, E., Duncan, T., & Tammen, V. V. (1987). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*, 48–58.

Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Morgan Kaufman.

Nielsen, J., & Mack, R. L. (1994). *Usability inspection methods*. New York, NY: John Wiley & Sons.

Pleban, R. J., Dyer, J. L., Salter, M. S., & Brown, J. B. (1998). *Functional capabilities of four virtual individual combatant (VIC) simulator technologies: An independent assessment* (Technical Report No. 1078). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Seymour, N. E. (2008). VR to OR: A review of the evidence that virtual reality simulation improves operating room performance. *World Journal of Surgery*, *32*, 182–188.

Taylor, G. S., Singer, M., & Jerome, C. (2009). Development and evaluation of the Game-Based Performance Assessment Battery (GamePAB) and the Game Experience Measure (GEM). *Proceedings from the 53rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 2014–2018). Santa Monica, CA: Human Factors and Ergonomics Society.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental functioning up the efficiency of other functions. *Psychological Review*, *8*, 247–261.

Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd ed.). New York, NY: Harper Collins.

Witmer, B. G., Bailey, J. H., & Knerr, B. W. (1995). *Training dismounted soldiers in virtual environments: Route learning and transfer* (Technical Report No. 1022). Alexandria, VA:

U.S. Army Research Institute for the Behavioral and Social Sciences.

Witmer, B. G., Jerome, C. J., & Singer, M. J. (2005). The factor structure of the Presence Questionnaire. *Presence: Teleoperators & Virtual Environments*, *14*, 298–312.

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

Grant S. Taylor is a research associate with the Applied Cognition & Training in Immersive Virtual Environments (ACTIVE) Lab at the University of Central Florida Institute for Simulation and Training. Grant obtained a PhD in applied experimental and human factors psychology from the University of Central Florida in 2012.

John S. Barnett is a research psychologist with the U.S. Army Research Institute for the Behavioral and Social Sciences, Technology-Based Training Unit in Orlando, FL. He holds a PhD in applied experimental and human factors psychology from the University of Central Florida awarded in 2000.

Date received: January 26, 2012
Date accepted: September 24, 2012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* June 2013	**2. REPORT TYPE** Legacy	**3. DATES COVERED** *(From - To)* January 2012 – December 2012
4. TITLE AND SUBTITLE Evaluation of wearable simulation interface for military training		**5a. CONTRACT NUMBER** N/A
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER N/A
6. AUTHORS Taylor, Grant S.; Barnett, John S.		**5d. PROJECT NUMBER** N/A
		5e. TASK NUMBER N/A
		5f. WORK UNIT NUMBER N/A
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 6000 6TH Street (Bldg. 1464 / Mail Stop 5610) Fort Belvoir, VA 22060-5610		**8. PERFORMING ORGANIZATION REPORT NUMBER** N/A
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 6000 6TH Street (Bldg. 1464 / Mail Stop 5610) Fort Belvoir, VA 22060-5610		**10. SPONSOR/MONITOR'S ACRONYM(S)** ARI
		11. SPONSOR/MONITOR'S REPORT NUMBER(S) N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT: Distribution Statement A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES Published journal article in Human Factors, Vol 55(3), June, 2013 pp. 672-690. DOI: http://dx.doi.org/10.1177/0018720812466892

14. ABSTRACT

This research evaluated the training effectiveness of a novel simulation interface, a wearable computer integrated into a soldier's load-bearing equipment. Background: Military teams often use game-based simulators on desktop computers to train squad-level procedures. A wearable computer interface that mimics the soldier's equipment was expected to provide better training through increased realism and immersion. Method: A heuristic usability evaluation and two experiments were conducted. Eight evaluators interacted with both wearable and desktop interfaces and completed a usability survey. The first experiment compared the training retention of the wearable interface with a desktop simulator and interactive training video. The second experiment compared the training transfer of the wearable and desktop simulators with a live training environment. Results: Results indicated the wearable interface was more difficult to use and elicited stronger symptoms of simulator sickness. There was no significant difference in training retention between the wearable, desktop, or interactive video training methods. The live training used in the second experiment provided superior training transfer than the simulator conditions, with no difference between the desktop and wearable. Conclusion: The wearable simulator interface did not provide better training than the desktop computer interface. It also had poorer usability and caused worse simulator sickness. Therefore, it was a less effective training tool. Application: This research illustrates the importance of conducting empirical evaluations of novel training technologies. New and innovative technologies are always coveted by users, but new does not always guarantee improvement.

15. SUBJECT TERMS

Evaluations; Heuristics; Microcomputers; Military Training; Mobile Devices; Military Personnel; Simulation; Transfer of training; Retention

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Dorothy Young
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	Unlimited Unclassified	19	**19b. TELEPHONE NUMBER**
Unclassified	Unclassified	Unclassified			703-545-2316