Timing of Feedback Delivery in Game-Based Training

Cheryl I. Johnson, Heather A. Priest
U.S. Army Research Institute
Orlando, FL
{cheryl.i.johnson; heather.priest}@us.army.mil

David R. Glerum, Jr., Stephen R. Serge
University of Central Florida
Orlando, FL
{glerumd; sserge}@knights.ucf.edu

ABSTRACT

The Army Learning Model calls for a shift to a more learner-centric environment that provides relevant and engaging training available anytime, anywhere. As a result, instructors have taken a blended learning approach that incorporates more serious games and simulations for training. While many purport the effectiveness of games for training, there is little research concerning what instructional features implemented within a game lead to better learning. It is generally accepted that feedback is important for improving performance and enhancing learning (Moreno, 2004), but what is less clear is when to deliver feedback during training. According to the temporal contiguity hypothesis, providing feedback immediately after a mistake would be most effective in that one could correct errors right away (Anderson et al., 1995). On the other hand, receiving feedback in the middle of the game may interrupt attention to the task, be distracting to the trainee, and hinder learning. Therefore, the goal of the present research was to examine the timing of feedback delivery in a game-based environment for novice trainees learning a procedural task. Participants performed a search and report task and received feedback on their errors immediately after the mistake (immediate), during a logical breaking point in the scenario (chunked), at the end of the scenario (delayed), or received no feedback (control). Performance during the three training missions (in which the trainees received feedback), during a performance mission (in which trainees did not receive feedback), and scores on a retention test were the main dependent variables of interest. The results show that providing feedback improves performance compared to not receiving any feedback, but the timing of feedback did not affect performance beyond the first mission. Implications of these results for future research are discussed.

ABOUT THE AUTHORS

Cheryl I. Johnson is a Research Psychologist at the U.S. Army Research Institute (ARI), Technology-Based Training Research Unit in Orlando. She received her Ph.D. and M.A. in Psychology from the University of California Santa Barbara and specialized in Cognition, Perception, and Cognitive Neuroscience. Dr. Johnson has over 10 years of experience in technology-based training research. Her research interests include adaptive training systems, instructional strategies, design principles for game-based training, and multimedia learning.

Heather A. Priest is a Research Psychologist at the U.S. Army Research Institute (ARI), Technology-Based Training Research Unit in Orlando. She received her Ph.D. in Human Factors Psychology from the University of Central Florida and her M.S in Experimental Psychology from Mississippi State University. At ARI, her research involves intelligent tutoring systems and adaptive training, automated feedback in game-based training, instructional design principles for training technology, and synthetic teammates in game-based training. Along with numerous publications, Dr. Priest also served as a co-editor on the Adaptive Training Special Issue of *Military Psychology (2012).*

David R. Glerum Jr. is a Ph.D. student and a graduate of the Master's program in Industrial/Organizational Psychology at the University of Central Florida. He is currently working as a Graduate Consortium Research Fellow at U.S. Army Research Institute (ARI), Technology-based Training Research Unit in Orlando, FL. He works closely with Army personnel and other graduate students on research related to technology-based training, adaptive training, and simulation.

Stephen R. Serge is a Graduate Consortium Research Fellow working with the U.S. Army Research Institute (ARI), Technology-based Training Research Unit in Orlando, FL. He received his M.A., and is currently a Doctoral

Candidate, in the Applied Experimental and Human Factors Psychology program at the University of Central Florida. Current work includes simulation and game-based training, with a focus on the development of effective embedded instructional guidance strategies, design guidelines for such systems, and usability.

Timing of Feedback Delivery in Game-Based Training

Cheryl I. Johnson, Heather A. Priest
U.S. Army Research Institute
Orlando, FL
{cheryl.i.johnson; heather.priest}@us.army.mil

David R. Glerum, Jr., Stephen R. Serge
University of Central Florida
Orlando, FL
{glerumd; sserge}@knights.ucf.edu

INTRODUCTION

The Army Learning Concept for 2015 (TRADOC Pam 525-8-2) calls for a shift from the current instructor-centered, lecture-based methods to a learner-centered, experiential approach in order to promote adaptable qualities in Soldiers and Leaders so that they may operate efficiently in uncertain and complex situations. This new Army Learning Model (ALM) requires training that is relevant and engaging through context-based, collaborative, and problem-centered instruction, which is tailored to an individual learner. Another requirement of the ALM is that training shall be made available at the point of need -- anytime, anywhere. As a result, instructors have taken a blended learning approach that incorporates technology-based training solutions, such as serious games and simulations for training. While many purport the effectiveness of games for training, there is little research concerning what instructional features implemented within a game lead to better learning (Hannafin & Vermillion, 2008; Hays, 2005; O'Neil & Perez, 2008). One instructional feature that is generally accepted as being important for improving performance and enhancing learning is providing feedback (Moreno, 2004). Yet there is no clear guidance in the research literature on *when* to provide feedback during a game-based training exercise. Therefore, the goal of the present research was to examine the timing of feedback delivery in a game-based environment for novice trainees learning a procedural task.

Feedback Timing Debate

Providing feedback can contribute to learning by allowing students to evaluate their responses or behaviors, identify a discrepancy in their knowledge, and potentially repair faulty knowledge. There have been hundreds of studies examining the effects of feedback on learning and performance and a number of reviews and meta-analyses summarizing the findings (e.g., Bangert-Drowns et al., 1991; Kluger & DeNisi, 1998; Kulik & Kulik, 1988; Mory, 2004; Shute, 2008). Throughout the expansive literature on feedback, there are conflicting results making it difficult to draw any strong conclusions on how best to provide feedback in any given situation. In addition, it is unclear how feedback in new technology-based training contexts, such as serious games and simulations, should be implemented to maximize effectiveness. One such area of debate in the feedback literature is the timing of feedback. Is it more effective to provide immediate feedback after the student makes a mistake? Or is it best to give the feedback after some sort of delay (and how long of a delay should that be)?

The Case for Immediate Feedback. According to the temporal contiguity hypothesis, providing feedback immediately after a mistake would be most effective in that one could correct errors right away and be prevented from encoding incorrect information into memory (Anderson, Corbett, Koedinger, & Pelletier, 1995; Azevedo & Bernard, 1995; Bangert-Drowns et al., 1991). In a study by Corbett and Anderson (2001), college students learned LISP programming and completed problems with an intelligent tutor. Students who received immediate feedback on their answers completed the lessons more efficiently than the other feedback conditions. Also, proponents of immediate feedback argue that it can promote mindful behavior and motivate the learner to practice (Hoska, 1993; Narciss & Huth, 2004; Shute, 2008).

*The Case for Delayed Feedback***.** On the other hand, according to the distraction hypothesis, immediate feedback may interrupt attention to the task and have a negative impact on performance. That is, feedback presented immediately may be distracting to the trainee and hinder learning, especially within serious games and simulations (Schmidt & Wulf, 1997), when paying attention to the task is highly important. Proponents for delayed feedback also argue that immediate feedback may cause the learner to rely on feedback as a crutch and have difficulty performing when it is taken away (Schmidt, 1991; Shute 2008); that is, immediate feedback may not promote active learning. In one often cited, classic study on delayed feedback, Schmidt and colleagues (1989) tested feedback intervals on a ballistic timing task over 90 trials; participants were asked to maintain a certain performance time and

provided feedback after every trial, every 5 trials, every 10 trials, or every 15 trials. After training, participants performed 25 more retention trials where they received no feedback and then completed 25 more trials two days later (i.e., as a delayed retention task). They found that during practice, performance decreased as the time between intervals increased. That is, during practice trials, participants who received feedback after every trial performed best, while participants who received feedback after every 15 performed the worst. However, during the retention and delayed retention trials, the opposite was found to be true, so that participants who received immediate feedback after every trial performed the worst on retention, and the participants who only received feedback after every 15 trials performed the best. These findings lend support to the criticism that immediate feedback promotes shallow, short-term learning, while delayed feedback requires more active learning and, therefore, improves retention.

The Present Experiment

One goal of the present experiment was to weigh in on the feedback timing debate by directly testing what method of providing feedback is more effective—should feedback be presented immediately after a mistake or at the end of the training exercise? Specifically, participants performed a search and report task in a game-based training environment and received feedback on their errors, which was manipulated as a between-subjects variable. The Immediate condition received feedback immediately after committing an error; the Chunked condition received feedback on errors during a logical breaking point in the scenario (i.e., an intermediate timing group between immediate and delayed); the Delayed condition received feedback on errors at the end of the scenario; and the Control condition did not receive feedback on errors. A second goal of this research was to provide guidelines for training and instructional designers interested in using game-based environments for their curricula. The results of this study provide some guidance on how to deliver feedback effectively in a game-based training context. In this case, the participants were novices learning a procedural task in a game-based training environment.

METHOD

Participants and Design

Participants were randomly assigned using a block randomization procedure to one of the four conditions: control, immediate, chunked, or delayed feedback. The experiment was a mixed design with Feedback Condition as a between-subjects variable (Immediate, Chunked, Delayed, and Control) and Mission as a within-subjects variable (Missions 1-3 were training missions and Mission 4 was a performance mission). There were 111 participants (58 males; 53 females) with a mean age of 23. There were 26 participants in the Immediate condition, 26 in the Chunked condition, 30 in the Delayed condition, and 29 in the Control condition. Participants were recruited from a large university in the southeastern U.S. and the surrounding areas by advertisements on web-based recruitment websites and flyers posted on campus. Not all participants were college students necessarily but all reported that they had at least a high school diploma, and all were inexperienced with the search and report task. They received payment for their participation at a rate of $10 per hour.

Materials

Apparatus. Participants used two separate desktop computer systems during the experiment. One was used for the Game Distributed Interactive Simulation (GDIS), a first person shooter video game developed from a modified version of the retail game Half Life 2 ®. The GDIS scenario simulated a Military Operations in Urban Terrain (MOUT) site that consisted of a small town with two main roadways and 18 buildings. Participants used a standard keyboard and two button mouse to control their avatar and navigate the environment and could open and close doors and explore buildings. The second computer, situated adjacent to the first, was used to send and receive text messages from headquarters (i.e., the experimenter) and also to display feedback messages on the participant's performance. Participants could type messages with a keyboard and use the mouse to close feedback windows once they were finished reading them. Similarly, an experimenter used two separate desktop systems during the experiment. The first was used to monitor the participants as they navigated the GDIS environment. The second computer was used to send and receive text messages from the participant, assess participants' performance, and send performance feedback using the Semi-Automated Feedback System.

Semi-Automated Feedback System. A scoring protocol was developed by the experimenters prior to data collection in order to assess participants' performance of the trained task in the GDIS environment. The Semi-Automated Feedback System (AFS) allowed the experimenters to set up a computer-based scoring checklist by which to assess performance of a particular behavior associated with the trained search and report procedures. The AFS also enabled experimenters to push feedback to the participants at the time appropriate to their experimental condition.

Training manual. The training manual consisted of 16 instructional slides that contained information about the participant's role in the scenario and detailed information about the proper procedures for the search and report task. There were three terminal learning objectives that included proper procedures for entering and exiting buildings, searching buildings, and communicating with headquarters. Each terminal learning objective included 3-4 enabling learning objectives. An example of an enabling learning objective for searching buildings is "Use the right turn rule when deciding the order to search rooms."

Training and performance missions. The main dependent variable of this experiment was the performance scores on the four missions in GDIS, demonstrating participants' knowledge of the search and report procedures. Participants received a mission briefing sheet prior to each mission detailing the buildings to be searched and the target items to be reported to headquarters. The first three missions were training missions, in which participants were to search three different buildings with a 10-min limit. During the training missions, participants received feedback on their performance; the timing of this feedback varied by condition. The last mission was considered a performance mission due to the fact that there were four buildings to search under limited light conditions with the same 10-min time limit, and participants did not receive any feedback on their performance at any point during this mission.

Knowledge tests. Participants were given a pre-test ($\alpha = .57$) and a post-test ($\alpha = .96$) to assess comprehension of the training materials used in the experiment. Both tests were composed of ten multiple choice items that asked participants about the search and report procedures; while they covered the same material, the pre- and post-tests included different questions. The pre-test was administered prior to the participants receiving any information about the task and served as a baseline measure. The reliability of the pre-test was notably low, but this is expected as the participants were not expected to be aware of the emergency search procedures at the outset of the study. The post-test was administered after all the missions had been completed.

Spatial abilities measures. Spatial ability has been shown to affect how individuals learn from multimedia presentations (e.g., Mayer & Sims, 1994) and how well they can navigate in virtual environments (Diaz & Sims, 2003). Participants completed three paper-based measures of spatial ability, the Paper Folding Test (PFT; $\alpha = .92$), the Santa Barbara Sense of Direction Scale (SBSOD; $\alpha = .89$), and the Perspective Taking/Spatial Orientation Test (PTSOT; $\alpha = .89$); these particular measures were selected to assess different facets of spatial ability. The PFT (Ekstrom, French, & Harman, 1976) consists of two parts, each with ten items, and participants are given a 3-min limit to complete each part. For this test, participants are shown series of diagrams of a piece of paper being folded several times along with a hole being punched through the folded paper. The participants are to judge which one of the five figures to the right is the figure that shows the correct positioning of the holes when the paper is unfolded. The SBSOD (Hegarty et al., 2002) is a 15-item questionnaire in which participants rate their agreement with statements about their general spatial and navigational abilities and preferences on a 7-point scale. An example item is "I am very good at giving directions" with 1 "strongly disagree" to 7 "strongly disagree." The PTSOT (Hegarty & Waller, 2004) tests participants' ability to imagine a scene from different viewpoints and has 12 items with a 5-min time limit. In this test, participants are shown an array of objects and are instructed to imagine they are standing at one object while facing another object. Their task is to draw a line from the origin of a circle to indicate the angle at which they would be pointing at the object while facing the second object. Their total score is the average deviation from the correct angle across all items.

Cognitive load and workload questionnaires. Cognitive load was measured using the single item Cognitive Load Questionnaire (CLQ; Paas, 1992). The CLQ asks participants to rate their level of mental effort on a 9-point scale ranging from 1 "very, very low mental effort" to 9 "very, very high mental effort." Workload was measured using the paper-version of the NASA-RTLX (average α across missions = .79; Hart & Staveland, 1988), in which participants rate their perceived mental demand, physical demand, temporal demand, perceived performance, effort level, and frustration by making a tic-mark on a scale with 21 gradients. Participants who marked their responses between the tic-marks for an item were scored on each item by adding a half point to the lower tick-mark. These

measures were used to examine whether the various feedback timing conditions led differences in the level of perceived cognitive load or workload.

Demographics and Video Game Experience. A paper-based demographics questionnaire was used to solicit basic information about participants such as their age, sex, computer use, and video game experience (VGE; $\alpha = .74$). According to prior research, VGE has been shown to improve performance on virtual tasks (Richardson, Powers, & Bousquet, 2011). As such, four individual demographic items were used as indicators of VGE, on a 1 to 5 that consisted of self-reports of how experienced one was with video games, how often one played generally, how confident one was with video games, and how often one played first-person shooter games, specifically. Anchors varied depending upon the item content of the question (i.e., frequency of video game play, degree of confidence, etc.). Results revealed that 94.6% reported daily use of a computer, 69.4% reported owning a video game system, 49.5% reported intermediate video game skills, and 75.7% of participants reported that they play video games between 0 to 9 hours a week.

Procedure

Participants were run individually, and the experiment took approximately 2.5 hours to complete. Following the informed consent procedure, participants completed a demographics questionnaire that included questions about the participant's background and experience with video games. Next, they received paper-based training on how to operate the avatar within the computer game-based environment and were given time to practice in GDIS. Once they completed the practice scenario, they read through a 16-slide paper-based manual that described the procedures for the search and report task. They then applied these procedures in three training missions, and a fourth, more complex performance mission. Participants received different types of feedback following the three training missions, depending on their condition, but no feedback was provided during the transfer mission.

All feedback was based on in-game performance demonstrating knowledge of the proper search and report procedures. An experimenter scored performance in real-time using the AFS digital checklist system that consisted of behaviors representing the procedural learning objectives, and the experimenter marked whether the participant performed the behavior correctly or incorrectly. When a participant performed an incorrect behavior (or failed to perform a behavior), the system delivered a feedback message. The timing of these messages was based on the participant's condition. In the Immediate condition, the participant was sent a feedback message immediately after an error was committed. In the Chunked condition, a feedback message was sent at a logical breaking point within the scenario; in this case, feedback messages were held until the participant completed search of a building. In the Delayed condition, all feedback was presented at the end of the scenario, similar to an after-action review. Participants in the Control condition did not receive any feedback. In addition to sending the feedback messages, the checklist system automatically generated a percentage-based performance score that was presented to all participants regardless of condition at the end of each mission.

Following each mission, participants completed a cognitive load questionnaire and the NASA-TLX to determine if there were any perceived differences in workload across the different conditions. Lastly, participants were thanked and debriefed.

RESULTS

Preliminary Analyses

Preliminary analyses were conducted to determine whether or not the data met the general assumptions for parametric statistics, if participants differed on any characteristics across conditions, and whether or not there were outliers with undue influence on the model. With regard to cross-condition differences, one-way ANOVAs and chi-squared tests revealed that participants did not differ on pre-test or demographic characteristics across conditions.

Feedback Timing and Mission Scores

Prior to analysis of the Mission Score data, spatial ability was identified as a possible covariate with mission score performance. An examination of the correlations between the spatial ability metrics and mission scores indicated

that the PFT may pose a good candidate for inclusion as a covariate. Although the PTSOT did exhibit a medium to large correlation with performance in the first mission (r = -.41, p < .001) and a small to medium correlation with performance in the third mission (r = -.20, p = .035) according to Cohen's (1992) effect size conventions, the PFT consistently exhibited statistically significant correlations with Mission 1 (r = .31, p = .001), Mission 2 (r = .28, p = .003), Mission 3 (r = .27, p = .004), and the performance mission (r = .29, p = .002). The PFT also met the assumption of homogeneity of regression slopes required for appropriately conducting ANCOVA. Lastly, due to a violation of the assumption of sphericity in repeated measures designs, the Greenhouse-Geisser correction for degrees of freedom for testing within-subjects effects was applied to this analysis.

In order to determine whether or not modes of feedback timing had an impact on the mission performance scores, a 4 x 4 repeated measures analysis of covariance (ANCOVA) was conducted with mission number as a within-subjects factor (3 training missions and 1 performance mission) and condition as a between-subjects factor (immediate, chunked, delayed, and control). Figure 1 presents the estimated marginal means across all four missions and for each feedack condition, as adjusted for scores on the PFT. The effect of condition was significant, F(3, 106) = 9.35, p < .001, partial η^2 = .21, such that the study conditions differed with regard to their mission performance. Post-hoc tests adjusted for multiple comparisons using the Bonferroni correction revealed that those in the Immediate, Chunked, and Delayed conditions scored significantly higher than the Control group. There were no significant differences between the various feedback timing conditions, however simple effects analyses showed that the Immediate condition did score marginally higher than both the Delayed (p = .053) and Control (p = .056) conditions during Mission 1, suggesting that receiving feedback had a positive impact after only performing in one mission. In addition, there was a significant within-subjects main effect of the mission number on mission scores, F(2.42, 256.68) = 10.06, p < .001, partial η2 = .09, demonstrating that mission scores generally increased across missions, while generally decreasing at the performance mission. All conditions saw decrements in scores on the performance mission due to the difficulty of the scenario (i.e., increased time pressure and reduced visibility). There also was a significant interaction between mission and condition, F (7.27, 256.68) = 3.18, p = .003, partial η2 = .08, indicating that the pattern of change in mission scores varied as a function of condition. That is, the performance in the feedback conditions tended to improve across missions 1-3, while the control condition did not show improvement. The main effect of the covariate, the PFT, was significant, F(1, 106) = 18.67, p < .001, partial η^2 = .15. In summary, receiving specific feedback on errors benefited performance relative to receiving only an outcome score.

Comparison of Feedback Timing Conditions by Cognitive Load Questionnaire and NASA RTLX

To examine the effect of feedback timing condition on participants' perceived cognitive load, two separate 4 x 4 repeated measures analyses of variance (ANOVA) were conducted with mission number as a within-subjects factor and condition as a between-subjects factor on the CLQ and RTLX scores. Four cases were excluded from the CLQ analysis for non-response and two cases were excluded from the RTLX analysis for a combination of non-response and improper completion of the survey.

Figure 2 presents the CLQ means and across all four missions and for each feedback condition. For the CLQ, the assumption of sphericity was violated and the Greenhouse-Geisser correction was applied in this analysis. There was a significant, moderate within-subjects main effect of the mission number on CLQ scores, F(2.55, 262.25) = 43.03, p < .001, partial η^2 = .30, demonstrating that cognitive load generally decreased across missions, while generally increasing at the performance mission. The effect of condition was also significant, F(3, 103) = 5.15, p = .002, partial η^2 = .13. Post-hoc tests adjusted for multiple comparisons using the Bonferroni correction revealed that the Delayed condition exhibited higher levels of Cognitive Load across all four missions. All other differences between conditions were not significant. To clarify the within-subjects pattern of change in CLQ across the conditions, there was a significant interaction between mission number and condition, F(7.64, 262.25) = 2.15, p = .035, partial η^2 = .06. The interaction suggests that the pattern of change for the three conditions varied across the delayed, chunked, and immediate conditions such that the delayed condition slightly increased or reached a plateau in Cognitive Load during the first three missions, whereas the chunked and immediate conditions exhibited decreases in Cognitive Load.

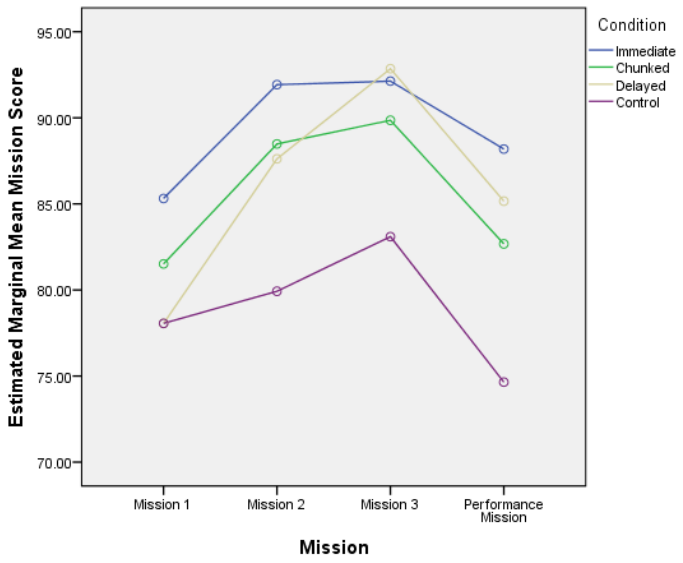

Figure 1: Estimated Marginal Means of the Mission Scores by Mission for all Study Conditions

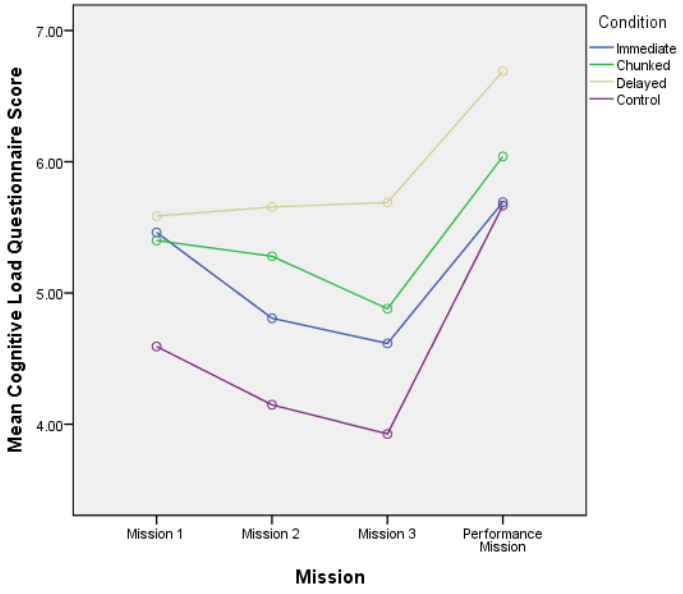

Figure 2: Means of the CLQ Scores by Mission for all Study Conditions

The RTLX estimated marginal means across all four missions and for each feedback condition are presented in Figure 3. The assumption of sphericity was also violated for the RTLX scores across missions and the Greenhouse-Geisser correction was applied in this analysis as well. There was a significant within-subjects main effect of the mission number on RTLX scores, $F(2.405, 250.13) = 12.52$, $p < .001$, partial $\eta^2 = .11$, demonstrating that the RTLX scores generally decreased across missions, while generally increasing at the performance mission. Furthermore, the effect of the confidence in video games as a covariate was significant, $F(1, 104) = 16.87$, $p < .001$, partial $\eta^2 = .14$. Confidence in video games was identified as a possible covariate with mental workload as Hart and Staveland (1988) identified the operator's perception of preconceptions and biases as well as task goals and structure as related

to mental workload. As the task primarily involves navigating through a video game environment, confidence in using video games sufficiently represents the perception of the task structure as well as preconceptions/biases regarding the task. After adjusting for confidence in video games the main effect of condition was significant, $F(3, 104) = 4.31$, $p = .007$, partial $\eta^2 = .11$. Post-hoc tests adjusted for multiple comparisons using the Bonferroni correction revealed that those in the Chunked and Delayed conditions tended to express that they were experiencing a higher mental workload than the Control group. All other differences between conditions were not significant.

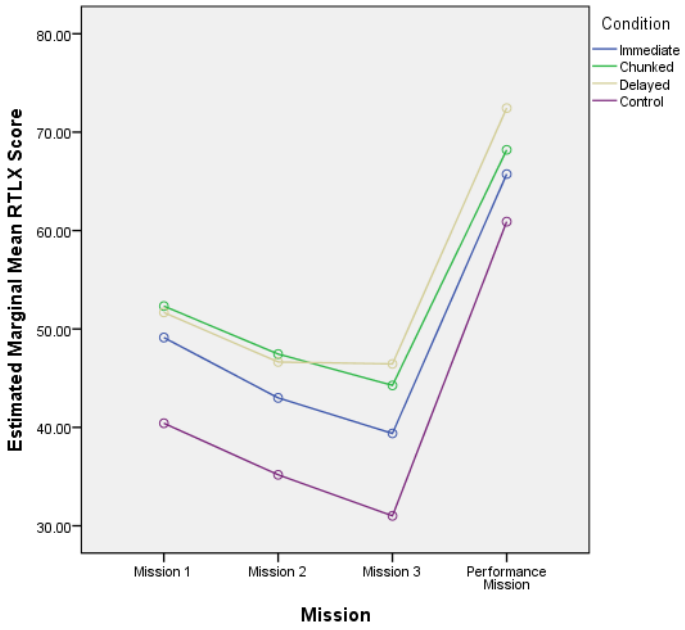

Figure 3: Estimated Marginal Means of the RTLX Scores by Mission for all Study Conditions

Knowledge Test Gains Across Conditions

Table 1 presents the means and standard deviations of knowledge test scores for each study condition. To examine the effect of feedback timing condition on the knowledge test scores, a 2 x 4 repeated measures analyses of variance (ANOVA) was conducted with test administration (i.e., Pre vs. Post) as a within-subjects factor and condition as a between-subjects factor on the knowledge test scores. Before examining these effects, it was first necessary to determine whether or not all participants started out with the similar emergency search procedure knowledge levels. As such, a one-way ANOVA was conducted to determine whether or not participants differed across condition on their Knowledge Pre-test scores. Participants across conditions scored similarly on the Knowledge Pre-test, $F(3, 106) = 0.91$, $p = .437$. There was a significant, large main effect for the within-subjects factor of test administration indicated that knowledge of emergency search procedures increased from pre-test to post-test, $F(1, 106) = 1272.70$, $p < .001$, partial $\eta^2 = .92$. However, the effect of condition was not significant suggesting that these knowledge gains were not a function of levels of feedback timing.

Table 1: Means and Standard Deviations of Knowledge Test Scores by Condition

	Pre-Test	Post-Test
Condition	*M* (*SD*)	*M* (*SD*)
Immediate	2.12 (1.18)	8.58 (0.90)
Chunked	2.46 (1.61)	8.27 (1.15)
Delayed	1.93 (1.03)	8.76 (0.95)
Control	2.38 (1.50)	7.90 (1.32)

DISCUSSION

The goal of the present experiment was to compare methods of feedback timing in a game-based training environment for novice trainees learning a procedural task. According to the temporal contiguity hypothesis, presenting immediate feedback would lead to better performance, while the distraction hypothesis predicts that presenting immediate feedback would impair performance. The data clearly demonstrate that the administration of feedback during the mission improves in-game performance relative to receiving only a feedback score, but there is little evidence to suggest that the timing of feedback impacts performance in the game-based simulation. During the first mission, those that received immediate feedback did perform marginally better than those who received no feedback on their performance up to that point (i.e., the delayed group received feedback at the end of the mission and the control group received only a performance score), demonstrating an early benefit for receiving immediate feedback and providing some support to the temporal contiguity hypothesis. However, participants who received immediate, chunked, or delayed feedback performed at statistically equivalent levels during Missions 2-4, suggesting that the benefit of immediate feedback timing disappears after Mission 1 when all the treatment groups had received feedback. Interestingly, participants who received delayed feedback reported the highest levels of perceived cognitive load and workload, despite performing similarly to the other feedback timing groups. This finding suggests that the timing of feedback can impact cognitive load; delayed feedback may lead to degraded performance in a more difficult task if it indeed induces greater cognitive load than other feedback timing conditions (see Sweller, van Merrienboer, & Paas, 1998). Furthermore, mission performance in a game-based simulation may be affected by trainees' prior spatial ability as demonstrated by the observed effect of including the PFT as a covariate. Likewise, trainees' confidence with video games may have an impact on their perceived mental workload.

One limitation of this experiment that may have reduced the impact of feedback timing was that the feedback messages were presented to trainees on a screen separate from the screen they used to perform the task. While the results indicate that the feedback was still beneficial (as the feedback groups performed better than the control), the advantage of receiving feedback may have been reduced due to the need for participants to split their attention across two screens. Research has shown that people learn better when the relevant information in a multimedia presentation is presented close together rather than far apart in a finding known as the spatial contiguity effect (and also called the split attention effect; Chandler & Sweller, 1992; Ginns, 2006; Johnson & Mayer, 2012). It could be the case that individuals who received immediate feedback could selectively pay attention to it, since it was presented on a separate screen. Follow-up experiments should pursue the effect of feedback timing when the feedback is presented on the same screen. Additionally, the particular task may not have been sensitive enough to show differences between the feedback groups, as performance tended to be fairly high even after the first mission, so follow up experiments are also worth pursuing.

Another area that shows promise for future research is adaptive feedback—that is, feedback that adapts to the trainee's performance (Durlach & Ray, 2011; Shute & Zapata-Rivera, 2008). Research has demonstrated that certain instructional strategies that benefit novices may actually impair performance for those with more experience—this finding is called the expertise reversal effect (Kalyuga, Ayres, Chandler, & Sweller, 2003;

Kalyuga, Chandler, Tuovinen, & Sweller, 2001). More research is needed to determine how providing feedback in game-based training environments affects individuals with more expertise with a given task—it could be the case that one method of timing would be more suitable for novices, and another more suitable for experts. With an adaptive system, as one gains expertise with a task, the feedback timing could be adjusted accordingly. More research is needed to determine how and when to implement feedback that adapts to the needs of the trainee.

In conclusion, the results of this experiment point to the importance of providing feedback to trainees on their performance in a game-based training exercise. Feedback provides trainees the opportunity to quickly correct mistakes, which could potentially lead to more effective and efficient learning. In addition, individual difference factors such as spatial ability and confidence in video games could have an impact on how well trainees can learn in game-based environments, and training developers should take these factors in account when designing games for training.

ACKNOWLEDGEMENTS

We would like to thank Dr. Glenn Martin and Jaime Flores (IST) for their assistance with programming the AFS. We also wish to thank Anthony Baker, Katherine Hood, and Angela Kramperth (UCF) for their assistance with data collection.

REFERENCES

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Chandler, P. & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal* of *Educational Psychology, 62*(2), 233-246.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Diaz, D. D. & Sims, V. K. (2003). Augmenting virtual environments: The influence of spatial ability on learning from integrated displays. *High Ability Studies, 14*, 191-212.

Durlach, P. J. & Ray, J. M. (2011). Designing adaptive instructional environments: Insights from empirical evidence (Technical Report 1297). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Science.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction, 16*(6), 511-525.

Hannafin, R. D., & Vermillion, J. R. (2008). Technology in the classroom. In T. L. Good (Ed.), *21st century education: A reference handbook* (vol. 2; pp. 209-218). Thousand Oaks, CA: Sage.

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.139-183). Amsterdam: North Holland.

Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion.* (Technical Report 2005-004). Orlando, FL: Naval Air Warfare Center, Training Systems Division.

Hegarty, M., Richardson, A. E., Montello, D.R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence, 30*, 425-447.

Hegarty, M. & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32*, 175-191.

Johnson, C. I. & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied, 18*(2), 178-191.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38, 23-31.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. Journal of Educational Psychology, 93, 579-588.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Kulik, J. A. & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*, 79-97.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63,* 505-512.

Maslovat, D., Brunke, K. M., Chua, R., & Franks, I.M. (2009). Feedback effects on learning a novel bimanual coordination pattern: Support for the guidance hypothesis. *Journal of Motor Behavior, 41*(1), 45-54.

Mayer, R. E. & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86*, 389-401.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*(1), 99-113.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of Research on Educational Communications and Technology* (pp.745-783). Mahwah, NJ: Lawrence Erlbaum.

Munro, A., Fehling, M. R., & Towne, D. M. (1985). Instruction intrusiveness in dynamic simulation training. *Journal of Computer Based Instruction, 12*(2), 50-53.

O'Neil, H. F., & Perez, R. S. (Eds.). (2008). Computer *games and team and individual learning*. Oxford, UK: Elsevier.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A cognitive load approach. *Journal of Educational Psychology, 84*, 429-434.

Richardson, A. E., Powers, M. E., & Bousquet, L. G. (2011). Video game experience predicts virtual, but not real navigation performance. *Computers in Human Behavior, 27*(1), 552-560.

Schmidt, R. A., Young, D. E.; Swinnen, S., & Shapiro, D. C. (1989). Summary of knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(2), 352-359.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Shute, V. J. & Zapata-Rivera, D. (2008). Adaptive technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology* (3rd Ed.) (pp. 277–294). New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10(3)*, 251-296.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing.* Burlington, MA: Elsevier.