TRUST IN SYNTHETIC TRAINING ENVIRONMENTS: APPLICATIONS FOR MILITARY SOLDIERS

Conference Paper · July 2019

DOI: 10.33965/ihci2019_201906C050

7 authors, including:

Michael Boyce
Yale University
59 PUBLICATIONS **627** CITATIONS
SEE PROFILE

Ericka Rovira
United States Military Academy
55 PUBLICATIONS **1,007** CITATIONS
SEE PROFILE

ISBN: 978-989-8533-91-3 © 2019

TRUST IN SYNTHETIC TRAINING ENVIRONMENTS: APPLICATIONS FOR MILITARY SOLDIERS

Michael Boyce[1], Ericka Rovira[2], Joshua Rea[2], Payton Rengel[2], John Emezie[2], Christian Ackerman[2] and Camilla Knott[3]

[1]*Simulation and Training Technology Center – West Point Field Element, West Point, NY, USA*
[2]*U.S. Military Academy, West Point, NY, USA*
[3]*TiER1 Performance Solutions, Washington, DC, USA*

ABSTRACT

Familiarity with simulations and game-based environments, including the use of technologies like Augmented Reality (AR) and Virtual Reality in which these games are played, may make young Soldiers more willing to use training approaches that leverage the use of gaming technologies. The goal is to introduce sophisticated training technologies and to use technological developments to tailor training to Soldier needs. AR/VR technologies have been identified as a low-cost solution to enhance training. However, little is known about the impact of AR/VR technology system reliability as well as how best to use VR for training. Decades of human automation interaction research suggests that technology reliability impacts trust in the system which then impacts SA and task performance. An understanding of how trust in VR interplays with performance outcomes is critical for enhancing Soldier performance with VR technology. Concurrently, individual differences will impact responses to new technologies and should therefore be accounted for in the introduction of new forms of technology and training. Thus there is a need to understand (1) the impact of trust in VR-based training on training objectives, e.g., situational awareness (SA) or performance, given challenges with reliability of the technology, and (2) how to optimize the use of VR given individual differences in trust in technology. Thirty-six participants volunteered and participated in a single factor three level within subjects design tactical mission simulation. Results indicate less trust with Microsoft HaloLens technology as compared to a tablet or ARES table.

KEYWORDS

Augmented Reality, Virtual Reality, Training, Military, Usability, Trust, Individual Differences

1. INTRODUCTION

1.1 Synthetic Training Environment

All soldiers are expected to perform to a designated level of performance to achieve an accepted standard. This is often represented by creating training environments that mimic military operation conditions, also known as "training as you fight" (ADRP 7.0). This extends to the military classroom where performing exercises aimed at specific learning objectives and realistic training scenarios can enhance learning experiences for situated learning. The Synthetic Training Environment (STE) has been identified as one of the key priorities in the U.S. Army Modernization strategy. The STE refers to having a collective training capability to conduct complex, diverse and multi-domain battles at the point of need. The STE will leverage technologies like artificial intelligence (AI), machine learning (ML), and augmented reality to reduce the cost of executing collective training events. Developed products from the STE will have a strong emphasis on augmented reality technology, such as the Microsoft HoloLens ®. Hence it is important to establish how the technical capabilities and weaknesses of these technologies impact usability, soldier technology acceptance, and trust.

1.2 Trust and Automation

There is a vast literature on how humans interact with automation. Traditional automation is static, in contrast synthetic training environments are designed to learn and adapt based on the knowledge of the user as well as display vast amounts of information that would support the operator in decision-making.

The literature has demonstrated that the reliability of automation impacts human trust and reliance in automation (Parasuraman & Riley, 1997; PSW, 2000; Wickens & Xu, 2002; Parasuraman, Sheridan, & Wickens, 2008). And further suggests that with highly reliable, but imperfect automation, there is a differential cost associated with automation that supports operator decision making versus perceptual processes (Rovira, McGarry, & Parasuraman, 2007). It is unknown if this finding would apply to a synthetic training environment or more specifically to STE in the context of a military task.

Existing literature has conclusively shown the explanatory power of trust (Lee & See, 2004; Hoff & Bashir, 2015) in human automation interaction. According to Lee & See (2004) model of appropriateness of trust and automation, the key elements to develop trust consist of information simulation, trust evolution, intention formation and reliance action. To provide appropriate calibration, mixed reality devices will need to understand the soldiers trust in the system and its own limitations in terms of the current environment.

1.3 Human Partner Characteristics

The human factors literature suggests that the human element is a critical factor when designing technology. However, there has yet to be a thorough exploration of the factors that impact successful AR/VR in training. Given that the goal of STE is to act flexibly and adaptively i.e. more human like, human partner characteristics are likely to be helpful in understanding trust and performance with STE.

A small yet growing body of literature has explored cognitive individual differences with automation (perceived attentional control: Chen & Terrence, 2009; working memory: de Visser et al., 2010; Rovira et al., 2016; genetics: Parasuraman et al., 2012). STE are designed in some instances are similar to the highest types of automation, hence understanding the role of trust in performance with STE is imperative.

1.4 Research Questions

This research focuses on AR/VR technology, individual differences, and multimodal interaction with different combinations of technology to assess trust, usability, workload and performance while conducting a simulated attack. The primary research questions guiding this line of research include:

1. Does trust in VR-based synthetic training systems impact performance on training objectives, e.g., situational awareness (SA) or performance?
2. Do individual differences predict trust in AR/VR equipment?

2. METHOD

2.1 Participants

Thirty-seven cadets from the U.S. Military Academy volunteered and received extra credit in introductory psychology courses for their participation.

2.2 Equipment

2.2.1 ARES

The Ares Sandtable is an advanced battlespace visualization framework; a traditional sand table, filled with play sand, augmented with a commercial, off the shelf (COTS) projector, LCD monitor, laptop, and Microsoft Kinect and Xbox Controllers. Ares was developed internally at ARL, and has been used in previous experiments (Boyce, Reyes, Cruz, Amburn, Goldberg, Moss, & Sottilare, 2016; Boyce, Rowan, Amburn, Shorter, Moss, Goldberg, & Sottilare, 2018). The sandtable can be used in conjunction with the HoloLens peripherals to provide multi-modal visualization of the terrain, urban structures, and assess above/below the terrain.

Figure 1.

2.2.2 Microsoft HoloLens

The Microsoft HoloLens is an AR HMD that creates images through a projection system with holographic high-definition in full color with low latency in real time. The headset can capture photos, record video and allow users to navigate with air tap gestures With the Microsoft HoloLens in ARES users can see urban structures, artillery visualization, and other elements above the terrain.

2.3 Materials, Tests, Tasks, and Stimuli

Individual differences in attitudes (trust) towards technology will be assessed with Automation Induced Complacency Potential – Revised Scale (AICP-R). The AICP-R assesses propensity to trust technology. It is a ten item scale using a five point likert rating. The scale was administered at the start of the experiment. The following scales are used in between each condition. The System Usability Scale (SUS) uses a Likert scale format consisting of 10 questions that range with five responses from "strongly agree" to "strongly disagree". The SUS can be utilized as a tool to cover system usability, support and training. Analysis across ten years of research indicated that the SUS demonstrated strong reliability for measuring usability of a system, Cronbach's $\alpha = .91$ (Bangor et al., 2009). The technique captures examples of extreme expressions on a spectrum. For example, the individual might be asked to respond to statements such as "I thought the system was easy to navigate" or "I can't imagine myself using something like this". The Lee and Moray Trust Scale is a multi-item scale that examines the operators trust in automated systems specifically process control of an experimental task. A subjective 10 point rating scale measures participants' perception of the trustworthiness and reliability of the technology.

2.4 Experimental Design

A single factor three level within subjects design was employed. The first condition is using the ARES tablet tactical planner by itself, the second is to use the tablet tactical planner in conjunction with the ARES table, and the third is to use the tactical planner app in conjunction with the Microsoft HoloLens.

2.5 Procedure

Participants first completed the AICP-R scale. Participants then observed a map of a company attack and answered two types of questions. One type of question was what they observe in front of them at the given time. For example these questions consisted of asking "What phase line did the company just cross?" or "Where is the enemy location?" The second set of questions test what the participant anticipated would happen next according to the display in front of them. For example these questions consisted of asking "What is the enemies' most likely course of action?" or "Where should you place the FO?"After each condition, the participants were given a series of surveys: the SUS, the NASA-TLX, the SEQ, TAM, and trust questions modeled after Lee & Moray (1994). Performance time was measured via stopwatch. Accuracy was measured by percentage correct of the scenario- based questions.

3. RESULTS

All dependent variables were checked for outliers and one extreme outlier was removed. This left a total of 36 data points. Testing for normality was accomplished via histograms and assessing the shapiro-wilk test since

the sample was less than 50. Although the Shapiro-Wilk showed violations of normality for two of the dependent variables, (Time on Task Sandtable, and Accuracy Tablet) inspection of both the histogram plots and QQ plots, along with the robustness of the ANOVA to withstand violations of normality it was decided to continue with the analysis.

A repeated-measures analysis of variance (RM ANOVA) was conducted to assess the effect of technology on Time on Task, Accuracy, Workload according to the NASA TLX, System Usability according to the System Usability Scale, and Automation Aid Trust using the four question Lee & Moray Questionnaire. Upon running an RM ANOVA, there is a need to check Mauchly's test for Sphericity. Only one of the dependent variables (SUS) violated sphericity (p = .023), but upon checking the corrections, all the corrections demonstrated significance (p < .001).

Analysis were completed by looking at within-subject Helmert Contrasts. The reason for this is that due to our population (West Point Cadets) familiarity with both flat maps as they are shown in the classroom, the fact that every cadet has a tablet, and our primary research interest of understanding how the additional technology (hololens / sandtable) would impact our outcome variables. Below is a table of those contrasts that were significant at the .05 threshold.

Table 1. Contrasts

Tests of Within-Subjects Contrasts				
Source			F	Sig.
condition	Time_On_Task	Tablet vs. Later	25.328	0.000*
		Sandtable vs. HoloLens	4.739	0.036*
	SUS	Tablet vs. Later	14.826	0.000*
		Sandtable vs. HoloLens	16.126	0.0008
	Lee_Moray_Q2	Tablet vs. Later	0.003	0.954
		Sandtable vs. HoloLens	4.831	0.035*
	Lee_Moray_Q4	Tablet vs. Later	0.986	0.328
		Sandtable vs. HoloLens	7.029	0.012*

3.1 Time on Task

Planned contrasts indicated a significant difference in the predicted direction for both Tablet versus the experimental conditions F(1,35) = 25.328, p < .001. Further, there is also a significant difference between the two experimental conditions in the predicted direction Sandtable vs. Hololens, F(1,35) = 4.739, p < .04. Taking the means and standard deviations of the three conditions and comparing them:

Tablet (M = 78.44, SD = 16.34) and Sandtable (M = 89.00, SD = 22.91), Cohen's d=.53 (medium)
Tablet (M = 78.44, SD = 16.34) and Hololens (M = 96.84, SD = 21.85), Cohen's d = .95 (large)
Sandtable (M = 89.00, SD = 22.91) and Hololens (M = 96.84, SD = 21.85), Cohen's d = .35 (small)

3.2 System Usability Scale

Planned contrasts indicated a significant difference in the predicted direction for both Tablet versus the experimental conditions F(1,35) = 14.826, p < .001. Further, there is also a significant difference between the two experimental conditions in the predicted direction Sandtable vs. Hololens, F(1,35) = 16.126, p < .001. Taking the means and standard deviations of the three conditions and comparing them:

Tablet (M = 75.90, SD = 11.68) and Sandtable (M = 74.16, SD = 12.56), Cohen's d=.14 (below small)
Tablet (M = 75.90, SD = 11.68) and Hololens (M = 61.95, SD = 15.63), Cohen's d = 1.01 (large)
Sandtable (M = 74.16, SD = 12.56) and Hololens (M = 61.95, SD = 15.63), Cohen's d = .86 (large)

3.3 Lee & Moray Question 2: To what extent did you rely on the automation aid in this scenario?

Planned contrasts indicated a significant difference between the two experimental conditions in the predicted direction Sandtable vs. Hololens, F(1,35) = 4.831, p < .04. Taking the means and standard deviations of the conditions and comparing them:

Sandtable (M = 7.97, SD =1.40) and Hololens (M = 7.39, SD = 1.38), Cohen's d = .41 (small)

Follow on analysis looking at the correlation between the alleviating workload factor of the AICP and reliance on the automated aid indicate a significant moderate positive correlation r(34) = .50, p < .01. Therefore

participants who answered that they relied more on the automation also rated more highly that they use automation to relieve workload.

3.4 Lee & Moray Question 4: To what extent do you think the automation aid improved your performance in this scenario compared to performance without the automation?

Planned contrasts indicated a significant difference between the two experimental conditions in the predicted direction Sandtable vs. Hololens, $F(1,35) = 7.029$, $p < .02$. Taking the means and standard deviations of the conditions and comparing them:

Sandtable ($M = 7.47$, $SD = 1.46$) and Hololens ($M = 6.53$, $SD = 2.12$), Cohen's $d = .52$ (medium)

4. CONCLUSION

Our data suggests that as compared to other technologies the HoloLens results in less trust and worse performance as measured by time on task. This is a first in a series of studies scheduled to investigate trust in AR/VR technology.

REFERENCES

Bangor, A., Kortum, P., Miller, J. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies, 4:* 114-123

Boyce, M.W., Reyes, R. J., Cruz, D. E., Amburn, C. R., Goldberg, B., Moss, J. D., & Sottilare, R. A. 2016. Effect of Topography on Learning Military Tactics - Integration of Generalized Intelligent Framework for Tutoring (GIFT) and Augmented REality Sandtable (ARES). Technical Report AD1017876.

Boyce, M. W., Rowan, C. P., Moss, J. D., Amburn, C. R., Shorter, P. L., Garneau, C. J., & Sottilare, R. A. (2018). The Impact of Surface Projection on Military Tactics Comprehension. *Military Psychology, 31:*45-59.

Chen, J.Y.C., and P.I. Terrence. 2009. "Effect of Imperfect Automation and Individual Differences Concurrent Performance of Military Robotics Tasks in a Simulated Multitasking Environment." *Ergonomics 52:* 907–920.

de Visser, E., T. Shaw, A. Mohamed-Ameen, and R. Parasuraman. 2010. "Modeling Human-Automation Team Performance in Networked Systems: Individual Differences in Working Memory Count." In Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting, 1087–1091. Santa Monica, CA: Human Factors and Ergonomics Society.

Hoff, K.A., and M. Bashir. 2015. "Trust in Automation Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors* 57 (3): 407–434.

Lee, J.D., and N. Moray. 1994. "Trust, Self-Confidence, and Operators' Adaptation to Automation." *International Journal of Human-Computer Studies* 40 (1): 153–184.

Lee, J., & See, J. (2004). Trust in automation and technology: Designing for appropriate reliance. *Human Factors, 46,* 50–80

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, & Shirase, L. Automaiton-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers Psychology, 10:*225

Parasuraman, R., E. de Visser, M.-K. Lin, and P.M. Greenwood. 2012. "Dopamine Beta Hydroxylase Genotype Identifies Individuals Less Susceptible to Bias in Computer-Assisted Decision Making." *PLoS ONE 7* (6): e39675.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230–253.

Parasuraman, R., T.B. Sheridan, and C.D. Wickens. 2000. "A Model of Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics – Part A* 30: 286–297.

Parasuraman, R., T.B. Sheridan, and C.D. Wickens. 2000. "A Model of Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics – Part A 30*: 286–297.Savoy: University of Illinois, Aviation Research Lab.

Rovira, E., K. McGarry, and R. Parasuraman. 2007. "Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task." *Human Factors 49:* 76–87.

Rovira, E., Pak, R., McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science, 18*: 573-591.

Wickens, C.D., and X. Xu. 2002. Automation Trust, Reliability and Attention (Tech. Rep. AHFD02-14/MAAD-02-2).