

Evaluating a virtual human storyteller for improved decision support

K.L. Blackmore ^a , S.P. Smith ^a , K.V. Nesbitt ^a , L. North ^a , S. Wark ^b and M. Nowina-Krowicki ^b

^a School of Electrical Engineering and Computing, The University of Newcastle, New South Wales, ^b Decision Support Systems, Joint Warfare & Operations, Joint & Operations Analysis Division, Defence Science & Technology Group, South Australia

Email: karen.blackmore@newcastle.edu.au

Abstract: Defence and security organisations face increasing challenges to maximise the real-time integration of data and information for improved decision making. This is particularly true in applications of human-machine teaming. These teams create complex interaction environments for human operators that integrate autonomous systems, multi-modal displays, and information streams at differing temporal scales. Examples of these complex environments can be seen in simulation pilots operating in large scale synthetic training scenarios, and operators of multi-unmanned aerial vehicle (multi-UAV) systems. In both these cases, an individual must perform a primary task (directing simulated or real entity behaviour), whilst simultaneously following a pre-defined scenario plan. On top of this, emergent information, with varying levels of uncertainty, must be acted on or ignored as appropriate. Additional contextual information may also be required such that these actions can occur with an understanding of broader contextual factors. The efficiency with which this primary, planned, emergent, and contextual information is presented underpins decision support and superiority in increasingly complex military contexts. In this research, we evaluate the use of storytelling via a virtual human to deliver complex geo-spatial conflict briefing information.

The role of virtual humans in information delivery is an area of active research. At the intersection of education, psychology, and computing, the development of effective virtual humans is a difficult task. Many factors contribute to their effectiveness, including cognitive loading associated with different forms of multimedia, as well as the design and features of the virtual humans themselves. Further complicating this is the inherent learning abilities and inclinations of individuals, and their underlying preferences for different learning and information delivery approaches.



Figure 1. DST Virtual Human Storytelling system

In this research we conducted an initial evaluation of the delivery of narrative content using the Defence Science and Technology (DST) Virtual Human Storytelling system (Figure 1). An experimental methodology was adopted, using a between-subjects design. A total of 87 participants viewed a 3-minute multimedia video on a hypothetical conflict scenario in a fictitious location that included details of actors in the conflict, actions occurring between these actors, and the geographic location. Three experimental conditions were used: a control group with the scenario delivered using text and images; a virtual human condition, with the scenario delivered using an animated avatar, synthesised speech, and images; and an audio condition, with the scenario delivered using synthesised speech and images. Preliminary analysis of the results suggests differences in performance on the information recall task, with participants in the virtual human group recording a greater range of, and higher maximum, scores. While simple between group comparisons of mean scores were not statistically significant, the higher maximum recall scores for participants in the virtual human group are promising. Directions for future work are also outlined, including incorporation of a measure of real-time cognitive load to isolate learning content for optimum delivery of information at a more granular level.

Keywords: Decision support, virtual humans, simulation, cognitive load, storytelling

1. INTRODUCTION

Simultaneous control of multiple unmanned aerial vehicles (multi-UAVs) during operations creates a complex environment that places high information processing demands on operators. Within these environments, operators must simultaneously execute an operational plan, developed and enacted under pre-existing contextual information, whilst directing tasks and responding to real-time multi-UAV events. Within this operating environment, all information can be considered to filter through an Information Communication Layer (ICL). In this context, the ICL consists of the technology, modality, and temporality of information and data that operators use to develop the situation awareness that forms the basis of their decisions and resultant actions. Figure 2 provides a high-level conceptual overview of the role of the ICL in multi-UAV tasks, including specific human and task/system factors that may also impact on the overall operating environment.

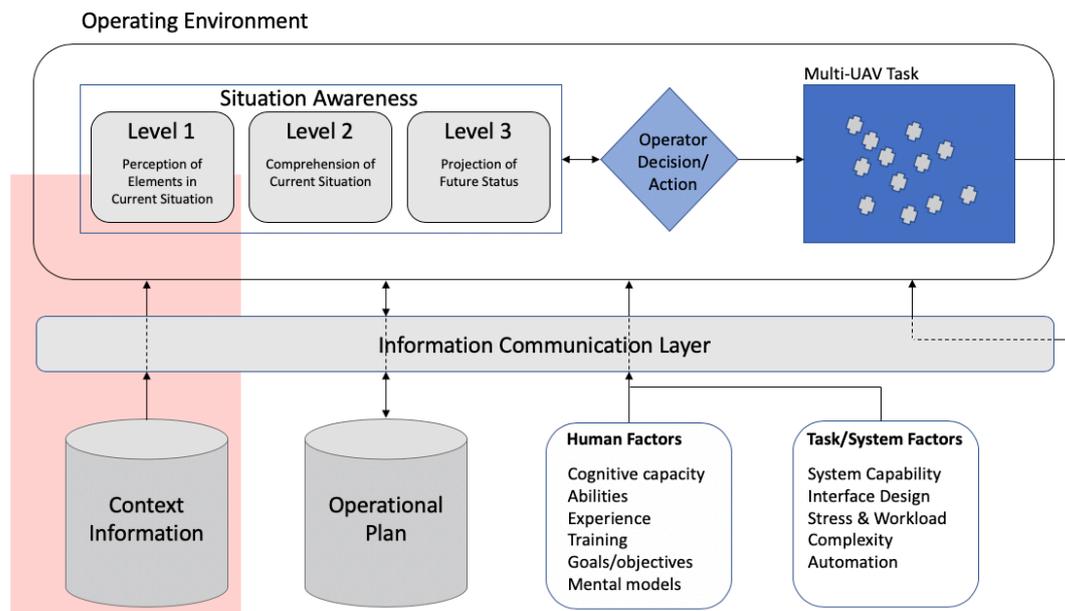


Figure 2. Conceptual model of the Information Communication Layer (ICL) in multi-UAV tasks (showing focus of current work (Adapted from Endsley and Garland 2000)). Dotted lines indicate implied information transfer, and pink shading shows focus area for the work presented here.

There has been considerable interest in minimising the cognitive load associated with complex tasks such as multi-UAV operation. For example, research has focussed on minimising the cognitive workload associated with the transition of control between human operators and unmanned vehicles (Zhang *et al.* 2018). Similarly, Mercado *et al.* (2018) considered the effect of agent transparency on operator workload in human-agent teaming for multi-robot management. In this research, we focus on exploring the contextual information provided to operators that underpins understanding of the pre-existing operational environment (pink shading in Figure 2). This contextual information, synonymous with a briefing, may be delivered in large ‘chunks’ or distributed into smaller information modules. In the latter case, cognitive switching between multiple sources of information, for example the operational plan and the multi-UAV task, may pose a problem, as is evident in other high-workload situations such as driving. This switching between information sources may result in an effect known as perceptual tunnelling, which may in turn result in delayed reactions or indeed a lack of response to situation changes in the operating environment (Jakus *et al.* 2015). However, there is evidence to suggest that information delivered using different perceptual modalities may reduce workload (Wickens 2002).

Virtual human storytellers mimic human information delivery and thus incorporate multiple perceptual modalities. While existing research has considered the influence of display type and narrative medium in the context of sensemaking (Hibbard *et al.* 2016), the potential benefits of virtual human storytellers for delivering contextual information in the form of a narrative requires investigation. In this research, we evaluated narrative delivery using a virtual human storyteller developed by Defence Science and Technology (DST) in comparison to audio only and text only modalities. We specifically considered the ability for users to recall information delivered using these different modalities as a first step toward more effective decision making. We begin by providing background on the use of virtual humans in learning contexts, followed by details on the methods, experimental design, results and discussion.

2. BACKGROUND

There is evidence to suggest that people find it easier to understand information presented through stories rather than textual forms such as bulleted lists (Gershon and Page 2001). This also supports Mayer's modality principle of multimedia design; that deeper learning occurs when words are presented as spoken narration rather than text (Mayer 2002). Thus, storytelling may be a mechanism to establish the context, and appropriate mental models, needed to achieve situational awareness and understanding for operators, decision makers, and analysts. Narration associated with multimedia content may help explain a visualisation or geospatial display and point out the significance and key aspects of what is being presented. Further, narrative delivered by virtual human storytellers can potentially value-add to this presentation through affective behaviours that convey uncertainty, importance, and urgency and more fully immerse the user in the narrative.

With the continued integration of virtual humans into our everyday world in film and television, home and device automation, and gaming (Craig *et al.* 2015), exploration of the role that these virtual humans may play in other fields of interest such as online learning, training, and communications is of keen interest (Wang and Antonenko 2017). Virtual humans (or in a learning context, pedagogical agents), are computer generated representations of "natural" humans, including facial expressions, gestures, and spoken intonations. Advances in display technology, animation tools, sensors, and machine learning make virtual humans viable alternatives for simulation training and information delivery. However, the effectiveness of these forms of training and communication, when compared to traditional approaches, is of critical importance. This is particularly true in the complex human-machine teaming environments that exhibit high cognitive workload demands.

Cognitive Load Theory (CLT) (Chandler and Sweller 1991) and Cognitive Theory of Multimedia Learning (CTML) (Mayer 2002) are used to provide guidelines for effective multimedia tools for learning. In brief, these theories suggest that the use of multimedia materials in conjunction with audio visual mediums allows for integration of new information with prior knowledge to construct new knowledge, with effective materials reducing the overall cognitive load. However, redundant information, or representation of two types of information through the one information method such as two forms of visual cues can result in split-attention and an increase in cognitive load (Craig *et al.* 2015; Homer *et al.* 2008).

The absence of a physical instructor *in situ* has been identified as a common problem in learning contexts (Wang and Antonenko 2017). Sustained attention to the activity is also crucial in effective learning and memory tasks (Broadbent 2013). However, evidence on the importance of a physical instructor presence is mixed. Wang and Antonenko (2017) found no difference in knowledge transfer regardless of the presence of an instructor but noted that the presence of an instructor had a positive influence on the participants perceived learning as well as lower degrees of mental loading. Homer *et al.* (2008) found that participant response to the presence of an instructor depended on their preference for visual or verbal information. van Wermeskerken and colleagues (2018) considered the presence and absence of a presenter, repeated across two experiments. They reported that participants paid significant attention to the face of the instructor and paid less attention to the task area. However, they found no difference in the learning performance of participants between the two presented scenarios.

Learning behaviour with the presence of virtual humans rather than a "natural" human has been the focus of many studies. A review conducted in 2011 concluded that as with studies featuring natural humans, mixed results on information retention were reported with the majority of studies yielding no improvement in learning (Heidig and Clarebout 2011). Lane (2016) extends this conclusion, suggesting that well designed and deployed pedagogical agents may have a small improvement on learning. Sadzak *et al.* (2007) tested information recall in a virtual heritage storytelling scenario using real (video recorded human) and virtual humans. Their findings indicated that information perception was better when delivered by a real human character. However, the avatar used was a low visual fidelity humanoid rather than a realistic virtual human character which may impact results. The focus of these studies is on information recall and retention; they do not consider cognitive load associated with these communication mediums. In terms of information delivery for decision support for multi-UAV operators, cognitive workload presents as key measure of information communication effectiveness in addition to information recall ability.

3. METHOD

The results of this study were collected between May and June 2019. This study was approved by the University of Newcastle Human Research Ethics Committee (H-2015-0163), with written informed consent obtained from all participants. In total, 98 participants were recruited via direct approach to University of Newcastle computer science, IT, software engineering, and psychology students as part of their coursework. Of these, 87 participants successfully completed all tasks included in the experiment.

3.1. Experiment Conditions

The DST Virtual Human Storyteller system was evaluated using a randomized control, pre-test, post-test design. Participants were randomly assigned to either one of three (3) groups: 0 = the control (static pages with picture, caption, text); 1 = virtual human condition (animated avatar with picture and text); 2 = audio condition (voice over with picture and caption) (see Figure 3). All media were MP4 format, at the same resolution, with a duration of 2min 45sec. The videos were played using the native Windows 10 Films&TV app (Version 10.19031.11411.0) running under Windows 10 Home (Version 1803) on an Alienware 15 RT laptop, with an i7 CPU 2.6GHz with 32 GB of RAM (64 operating system) and a GeForce GTX1070 graphics card. In all analyses, the between-groups factor was the presented scenario, and the procedure was conducted in a consistent manner, with the only change between studies the variation in the presented scenario.

 <p style="text-align: center;">Dragonia and Saxonia in dispute over Solace Bay (DZ) region</p> <p>There are long-standing tensions between the countries of Saxonia to the north, and Dragonia to the south, over sovereignty of the Solace Bay region with a large ethnic Saxonian population. The neighbouring countries of Avalon and Ryania have remained neutral.</p>	<p>Scenario 0 – Control Condition</p> <p>The control condition presents the participants with a video that includes text overlay on a news report style series of images. All information to score the total possible 25 marks on the recall task is provided within the text overlay, and it is supplemented by the images; no audio is present.</p>
 <p style="text-align: center;">Dragonia and Saxonia in dispute over Solace Bay (DZ) region</p>	<p>Scenario 1 – Virtual Human Condition</p> <p>In addition to the images shown in the control condition, a virtual human bust is overlaid on the news report, akin to a newsreader in a news report. Text presented in the control condition is instead provided using audio. The virtual human mimics human facial expressions and mouth movements when speaking. The speech is generated based on the control condition text using text-to-speech conversion software. Minimal text is provided via a news report style headline banner at the bottom of the screen.</p>
 <p style="text-align: center;">Dragonia and Saxonia in dispute over Solace Bay (DZ) region</p>	<p>Scenario 2 – Audio Condition</p> <p>The audio condition presents the audio from the virtual human scenario instead of the text from the control condition, and without the presence of the virtual human newsreader. Minimal text is provided via a news report style headline banner at the bottom of the screen.</p>

Figure 3. Stimuli for the experiment with control, virtual human, and audio conditions.

3.2. Measurement of Baseline Neuropsychological Status and Task Load

Baseline status was measured using the Inquisit List Learning Task (Borchert, 2018). This task presents a basic computer-adapted list learning task inspired by the Repeatable Battery for Neuropsychological Status (RBANS) (Randolph et al. 1998). The intent of this task was to determine both the delayed recall and delayed recognition behaviour of the participant. As the name suggests, delayed recall involves the recollection of a previously memorised word. Delayed recognition measures whether a word has been memorised and consolidated (Borchert, 2018). The test was implemented using Inquisit 5 Lab (www.millisecond.com) under Windows 10 Home (Version 1803) on an Alienware 15 RT laptop, with an i7 CPU 2.6GHz with 32 GB of

RAM (64 operating system) and a GeForce GTX1070 graphics card. The test was implemented as a pre-test task to collect participant data on immediate memory capacity

The task load associated with the allocated scenario was assessed using the NASA-TLX method (Hart and Staveland, 1988). This subjective self-assessment method reports the perceived work load across six subscales, namely the mental, physical and temporal demands associated with the task, as well as the degree of frustration experienced, their perception of performance, and the effort required to complete the task. In this study, the pen and paper implementation of this task was used.

4. RESULTS

The summary profile of each demographic group is provided in Figure 4.

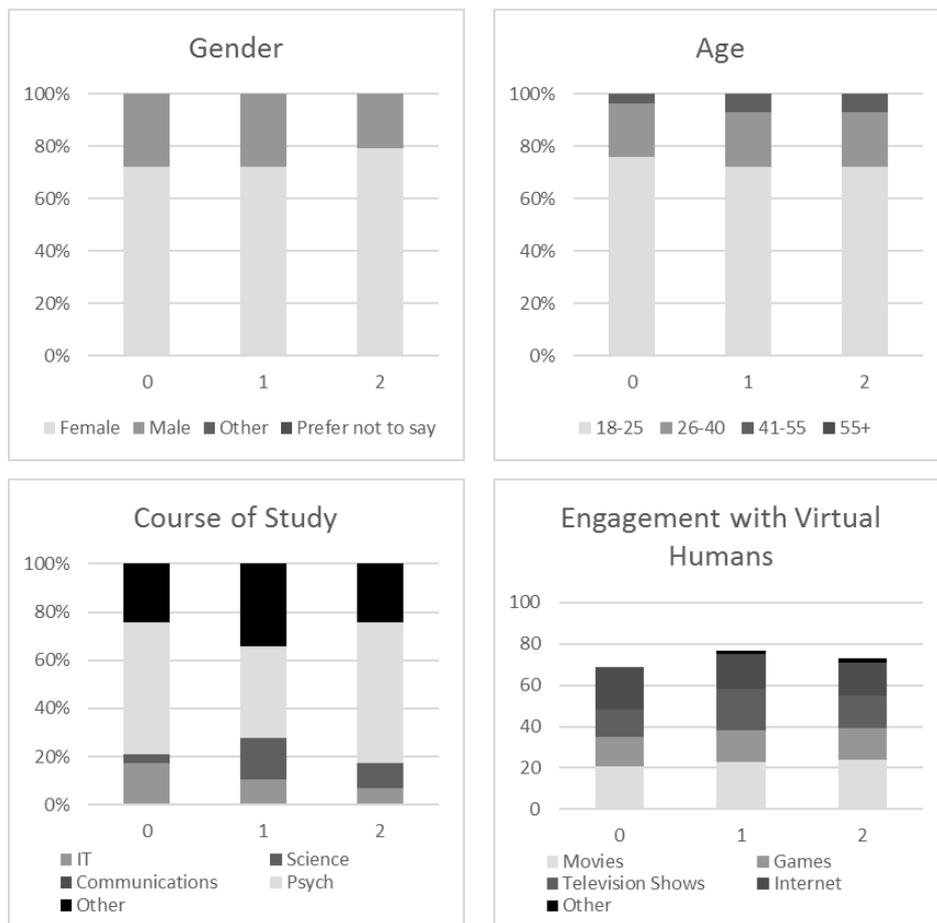


Figure 4. Summary profile for participants, grouped by experiment condition, where 0 = Control, 1= Virtual human, and 2 = Audio.

Categorical chi-square testing could not demonstrate any differences in age ($\chi^2=1.05$, p-value = 0.593 for both alternate scenarios) or gender ($\chi^2=0$ and 0.690, p-value = 1 and 0.406, respectively for scenarios 1 and 2) when compared to the control scenario (0) as reference.

4.1. Performance on Baseline Neurological Task

No statistically significant difference was evident from the baseline neurological status results, indicating similar information recall capacity for participants across groups. Summary results, as well as one-way ANOVA testing, are shown in Table 1.

4.2. Measure of Recall Activity and NASA-TLX

No statistically significant difference in the mean scores was observed in participant performance in the response to the three different scenarios. One-way ANOVA testing could not demonstrate significant differences between the performance of any group in the testing ($F(2, 84)=0.138$, $p = 0.871$). However, the range of results does provide some interesting insights. The results for the virtual human group (Group 1)

indicate a much greater range of scores, and it is also noted that the highest information recall scores (excluding outliers) were obtained by participants in this group. These results do suggest that some, but not all, participants may find it easier to recall information delivered via the virtual human storyteller.

Table 1. Baseline neurological status summary statistics

Group	Mean			SD			F Value	P-Value
	0	1	2	0	1	2		
Immediate Recall	7.65	7.89	7.74	1.10	1.02	1.31	0.306	0.737
Delayed Recall	7.89	8.10	8.27	1.85	1.50	1.36	0.407	0.667
Recognition	19.79	19.86	19.83	0.50	0.44	0.46	0.194	0.824

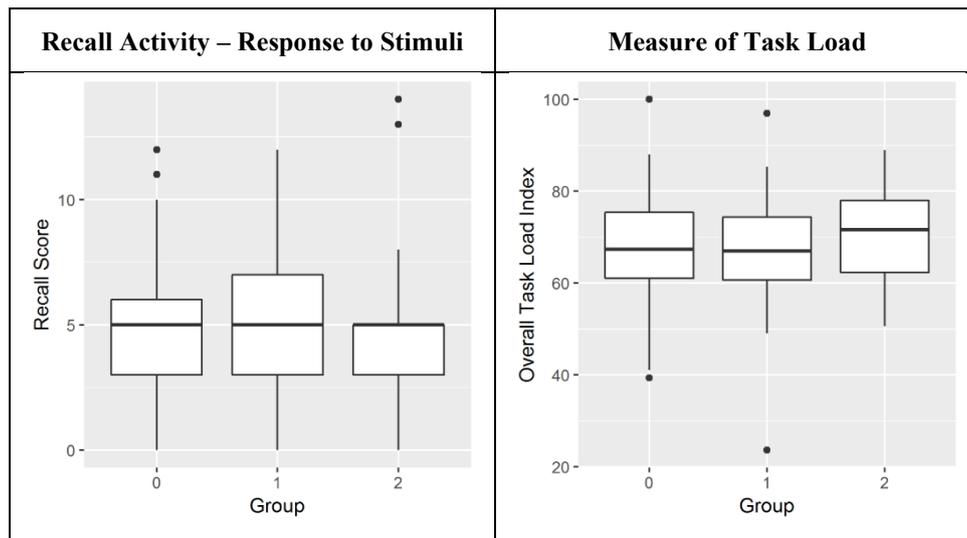


Figure 5. Boxplots of total score on the recall task and NASA-TLX Overall Task Load Index based on the differing stimuli. Group 0 is the Control scenario, Group 1 is the Virtual human scenario, and Group 2 is the Audio scenario.

No statistically significant difference was observed between participants overall perceptions of task loading using the NASA-TLX Overall Task Load Index (Figure 5; one-way ANOVA ($F(2, 84) = 0.340, p = 0.713$)). No differentiation was identified in the three scenarios on any of the individual subscales (Mental demand, $F(2, 84) = 0.600, p = 0.551$, Physical demand; $F(2, 84) = 1.387, p = 0.255$, Temporal demand; $F(2, 84) = 1.192, p = 0.309$, Effort; $F(2, 84) = 0.653, p = 0.523$, Performance; $F(2, 84) = 1.652, p = 0.198$, Frustration; $F(2, 84) = 0.656, p = 0.522$). Some studies have found that using the raw (unweighted) Task Load Index has yielded more representative results (Hart, 2006). The same one-way ANOVA analysis was repeated using this value, yielding similar results ($F(2, 84) = 0.391, p = 0.678$). It is interesting to note however that there does again appear to be some variability in the results across groups, with audio condition (Group 2 – Audio scenario) participants recording higher overall cognitive load. This is in keeping with existing literature (Wickens, 2002).

5. DISCUSSION AND CONCLUSION

While our preliminary results suggest a potential benefit in the delivery of complex information using a virtual human storytelling system, overall statistical significance was not demonstrated. The results of this study demonstrate that for participants with comparable performance in baseline testing, there is no statistical difference in response to the three alternate stimuli. Further, the perceived task difficulty was statistically equivalent in the three scenarios. However, the results also suggest that the recall task was very difficult. The average performance in the recall task across all groups was 5.03 out of a total possible 25 marks, with the best performing participant yielding a score of 14 out of 25. Whilst the best performing participant indicated a marginally lower than average overall task load index, no general correlation could be observed between the overall task load index and the recall score.

A number of limitations have been identified with this study; the overall task difficulty for the participant pool in this study proved to be a key issue. A clear way to address this limitation in future studies is to provide a more task-oriented activity for participants that better reflects the multi-UAV decision task that is of primary

interest in this research. Additionally, the narrative presented did not differ across the three information delivery scenarios and thus there is the potential that task dependence influenced these results. Future research will consider virtual human storytelling as an information delivery medium across differing tasks (eg. navigation, visual search, and construction) to account for this potential limitation. Lastly, post-hoc evaluation of cognitive task load does not provide fine grain insights into variability of workload across a task, and therefore the impact of the virtual human storyteller on cognitive task load may be better assessed using real-time cognitive load measures. In conclusion, the proposed Information Communication Layer (ICL) model for improving decision making in multi-UAV models provides a more general lens through which to structure research aimed at improving situational awareness for real-time decision-making tasks. Further analysis of virtual human storytelling in these high information processing environments using real-time measures of cognitive load is warranted.

ACKNOWLEDGMENTS

This project is funded by Defence Science and Technology (DST) under a Defence Science Partnering Deed (ID8837) with the University of Newcastle i3 Lab.

REFERENCES

- Borchert, K. (2018, August 24). User Manual for Inquisit's List Learning Task. Retrieved from <https://www.millisecond.com/download/library/v5/listlearningtask/listlearningtask.manual>
- Broadbent, D. E. (1958). *Perception and communication*. Oxford: Pergamon Press.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4), 293-332.
- Craig, S. D., Twyford, J., Irigoyen, N., & Zipp, S. A. (2015). A test of spatial contiguity for virtual human's gestures in multimedia learning environments. *Journal of Educational Computing Research*, 53(1), 3-14.
- Endsley, M. R., & Garland, D. J. (2000). Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1, 24.
- Gershon, N., & Page, W. (2001). What storytelling can do for information visualization. *Association for Computing Machinery. Communications of the ACM*, 44(8), 31-31.
- Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland.
- Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning?. *Educational Research Review*, 6(1), 27-54.
- Hibbard, S. J., Whitney, S. J., Carter, L., Fidock, J. J., Temby, P., & Thiele, L. (2016). Making Virtual Sense: Display Type and Narrative Medium Influence Sensemaking in Virtual Environments. In *Intersections in Simulation and Gaming* (pp. 222-236). Springer, Cham.
- Homer, B. D., Plass, J. L., & Blake, L. (2008). The effects of video on cognitive load and social presence in multimedia-learning. *Computers in Human Behavior*, 24(3), 786-797.
- Jakus, G., Dicke, C., & Sodnik, J. (2015). A user study of auditory, head-up and multi-modal displays in vehicles. *Applied Ergonomics*, 46, 184-192.
- Lane, H. C. (2016). Pedagogical agents and affect: Molding positive learning interactions. In *Emotions, Technology, Design, and Learning* (pp. 47-62). Academic Press.
- Mayer, R. E. (2005). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 41, 31-48.
- Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for Multi-UxV management. *Human factors*, 58(3), 401-415.
- Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology*, 20(3), 310-319.
- Sadzak, A., S. Rizvic, Colin Dalton, and A. Chalmers. (2007). Information perception in virtual heritage storytelling using animated and real avatars. Paper presented at 23rd Spring Conference on Computer Graphics, SCCG 2007, April 26-28, 2007. Budmerice, Slovakia. ACM.
- van Wermeskerken, M., Ravensbergen, S., & van Gog, T. (2018). Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior*, 89, 430-438.
- Wang, J., & Antonenko, P. D. (2017). Instructor presence in instructional video: Effects on visual attention, recall, and perceived learning. *Computers in Human Behavior*, 71, 79-89.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159-177.
- Zhang, W., Shirley, J., Deng, Y., Kim, N. Y., & Kaber, D. (2018). Effects of dynamic automation on situation awareness and workload in UAV control decision tasks. In *International Conference on Applied Human Factors and Ergonomics* (pp. 193-203). Springer, Cham.