Military Psychology

ISSN: 0899-5605 (Print) 1532-7876 (Online) Journal homepage: https://www.tandfonline.com/loi/hmlp20

The impact of surface projection on military tactics comprehension

Michael W. Boyce, Charles P. Rowan, Paul L. Shorter, Jason D. Moss, Charles R. Amburn, Christopher J. Garneau & Robert A. Sottilare

To cite this article: Michael W. Boyce, Charles P. Rowan, Paul L. Shorter, Jason D. Moss, Charles R. Amburn, Christopher J. Garneau & Robert A. Sottilare (2019) The impact of surface projection on military tactics comprehension, Military Psychology, 31:1, 45-59, DOI: 10.1080/08995605.2018.1529487

To link to this article: https://doi.org/10.1080/08995605.2018.1529487

MILITARY PSYCHOLOGY
2019, VOL. 31, NO. 1, 45–59
https://doi.org/10.1080/08995605.2018.1529487

Routledge
Taylor & Francis Group

The impact of surface projection on military tactics comprehension

Michael W. Boyce[a], Charles P. Rowan[b], Paul L. Shorter[a], Jason D. Moss[a], Charles R. Amburn[a], Christopher J. Garneau[a], and Robert A. Sottilare[a]

[a]Human Research and Engineering Directorate, US Army Research Laboratory, Orlando, Florida; [b]Department of Behavioral Sciences and Leadership, United States Military Academy at West Point, West Point, New York

ABSTRACT

This experiment assessed how displaying information onto different surfaces (flat vs. raised) influenced the performance, workload, and engagement of cadets answering questions on military tactics. Sixty-two cadets in a within-subjects design each answered 24 tactics-related questions across 2 conditions (12 on flat, 12 on raised) which were measured by accuracy and time on task. After each set of 12 questions, the cadets took postsurveys assessing engagement, measured by a modified User Engagement Scale and the System Usability Scale, and workload measured by the NASA-TLX. Findings indicated that raised terrain surface led to reduced workload and increased engagement and time on task as compared to the flat terrain surface. A practice effect drove performance metrics (time on task and accuracy), where the learner performed better on the second surface type displayed. This research contributes to expanding the literature base that supports alternative display methods to increase engagement and augment instruction of military tactics tasks.

ARTICLE HISTORY
Received 1 September 2017
Accepted 25 September 2018

KEYWORDS
Augmented reality sandtable (ARES); generalized intelligent framework for tutoring (GIFT); military tactics; terrain; United States military academy (USMA)

What is the public significance of this article? –– This study suggests that using raised terrain surfaces during military tactics instruction can reduce learner workload and increase engagement. Cadets demonstrated a decrease in workload and an increase in willingness to explore and interact with the raised terrain surface.

Current and future operational environments will place increased responsibility on Soldiers to make decisions with strategic, operational, and tactical implications while operating in complex environments. (US Department of the Army, 2011b, p. 12)

A military university classroom is a place where instruction has direct, real-world implications. With the increased reliance on technology, there is no shortage of technological solutions that promise great opportunities to help the future soldier. This research examines novice learners' understanding of displays of topographical information for military tactical instruction using different surface types. Comprehension of military tactics requires the ability to understand battlefield layout and constraints to assess scenarios. Assisting learning of cadets in tactical instruction requires an approach that considers the proper presentation of stimuli and a consideration of both traditional and innovative forms of delivery of instruction.

One of the most common tools in a military classroom, such as the United States Military Academy at West Point (USMA), is a sand table. A sand table is a tool that represents a battlespace and facilitates military planning (Amburn, Vey, Boyce, & Mize, 2015; US Department of the Army, 2011b). The practical benefit of using sand tables is that they allow cadets to experiment with different courses of action, rehearse operations, and generate discussions for learning (Brewster, 2002; Smith, 2010).

The Army Research Laboratory has developed the Augmented REality Sandtable (ARES) to support traditional instruction while enhancing visualization capabilities via technology. The ARES consists of a traditional sand table augmented with a Microsoft Kinect™, an LCD monitor, and a laptop computer. The ARES uses projection to display military scenarios and topographical features onto the sand, providing new avenues of battlespace visualization to support learning (Amburn et al., 2015). Previous research has shown that the additional information provided by the ARES can support tasks such as landmark identification and distance estimation (Schmidt-Daly, Riley, Hale, Yacht, & Hart, 2016). This research seeks to extend these findings to how terrain elevation information can support student performance in tactical situations.

CONTACT Michael W. Boyce michael.w.boyce11.civ@mail.mil United States Military Academy at West Point, Thayer Hall Room 258B, West Point, NY 10996.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmlp.

In preparation for this experiment, we conducted a pilot study with 19 Reserve Officers' Training Corps (ROTC) cadets. As a between-subjects design, this experiment looked at the interpretation of eight military tactics questions using a single map on either a raised or flat surface (Boyce et al., 2016). Data from the pilot study indicated significant differences in favor of the raised surface indicating higher engagement and trends toward significance in favor of the raised surface leading to lower workload scores but nonsignificant differences for performance. From these results, we expected that the present study would indicate significant differences in favor of the raised surface for engagement and workload (Boyce et al., 2016).

Background/related work

We used the three-stage model of human information processing and simple four-stage model of human information processing as the foundations for this research. The three-stage model of human information processing consists of perception, cognition, and action (Proctor & Van Zandt, 2018), and the simple four-stage model of human information processing consists of sensory processing, perception/working memory, decision-making, and response selection (Parasuraman, Sheridan, & Wickens, 2000). This research also had an interest in exploring engagement and understanding how information processes would influence engagement. Therefore, the research examines cadet understanding across a four-phase continuum. This continuum begins by providing additional information via visual cues (perception) and displaying how those cues support or detract from the cognitive demand for the learner (workload). It continues to show how the cues and demand impact choices on each display (performance assessment) and ends with a holistic understanding of a learner's experience (engagement; see Figure 1).

Perception via pictorial depth cues

Being able to assess topographic maps has been described as a complex and challenging task, especially because topographic maps require an interpretation of slope and elevation (Newcombe et al., 2015). When viewing a two-dimensional (2D) topographic map, the map itself is distorted and abstracted from the actual terrain, and this makes it difficult to interpret (Li, Willett, Sharlin, & Sousa, 2017). The human visual system uses visual cues for distance estimation, shape, and depth when examining objects in a three-dimensional (3D) space (Reichelt, Häussler, Fütterer, & Leister, 2010). Past human factors research has found that additional depth cues can be used to enhance visualization of terrain as well as support tasks such as map reading (Schmidt-Daly et al., 2016; Wickens, 2000).

This experiment targets the use of monocular or pictorial depth cues in the perception of topography. Pictorial depth cues are those that require only one eye to perceive depth (McIntire, Havig, & Geiselman, 2014) and include cues relative size, occlusion, shading, and perspective. Wickens (2000) noted that pictorial cues are those which are created on a 2D plane but use projection to create a 3D view. This use of cues creates an example of pictorial realism (Roscoe, 1968) where the structure of the display is closer to real-world displays (Wickens, 2000).

Through the combination of several pictorial cues, such as perspective and relative size, it is possible to create what Pomerantz and Pristach (1989) called an emergent feature. *Emergent features* are relations that can be more salient to human perception than any individual segment or cue by themselves (Martin & Pomerantz, 1978; Pomerantz & Pristach, 1989; Treisman & Paterson, 1984). Previous research has also demonstrated that providing hybrid cues and techniques such as animations, shadows, or droplines can improve terrain visualization in discriminating elevation between surfaces (Barfield & Rosenberg, 1995; Hendrix & Barfield, 1995; Van Beurden, 2013; Willett, Jenny, Isenberg, & Dragicevic, 2015). The ARES currently uses contour lines, shadows, and elevation numerical markers as well as color to help distinguish differences between surfaces. Because the raised surface will be able to provide emergent features, we expect that it will be easier for cadets to comprehend than the flat surface.

Workload and display design. Workload is one of the most studied areas of human factors, describing the characteristics of the task, operator, and the environment (Wickens, 2017; Young, Brookhuis, Wickens, & Hancock, 2015). *Workload* is the amount of an individual's memory resources being used to complete a cognitive task (Kalyuga & Singh, 2016). The mental workload experienced during

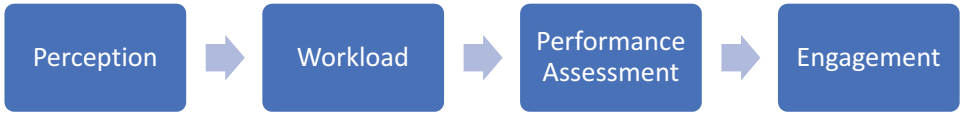

Figure 1. Four phase continuum guiding research.

learning, commonly referred to as *cognitive load*, can be one of three types: intrinsic, extrinsic, and germane. *Intrinsic* is load associated with the content itself, whereas *extrinsic* deals with the delivery of information, and *germane* seeks to develop mental models or schemas to support automatic processing (Paas, Renkl, & Sweller, 2003; Sweller, 1994, 2016; Sweller, van Merrienboer, & Paas, 1998).

The workload associated with a task can impact performance (Galán & Beal, 2012). Previous research has argued that the difficulty of the task, the context in which the task occurs, and the skill of the person performing the task are factors impacting workload (Young et al., 2015). One of the myths surrounding workload is that lower workload tasks are beneficial to the participant. However, having lower workload does not necessarily lead to better performance. For example, Wickens (2017) explained how listening may be more comfortable than taking notes (i.e., demonstrating lower levels of workload), but he pointed to the "generation effect" where retention via active processes outperforms similar passive processing (Slamecka & Graf, 1978; Wickens, 2017). The appropriate amount of workload when working with a display helps prevent cognitive underload or overload from occuring (Fallahi, Motamedzade, Heidarimoghadam, Soltanian, & Miyake, 2016; Hancock & Chignell, 1988; Young & Stanton, 2002). Underload or overload according to Young et al. (2015), is a mismatch between requirements and capabilities and the goal is to find the optimal zone to maximize performance, neither too high nor too low (Wickens, 2017; Yerkes & Dodson, 1908).

Wickens' multiple resource theory provides a helpful paradigm for modeling workload and determining human performance (1991). This theory describes a person's ability to process information as a function of various resource channels (e.g., visual, perception, and processing). Human limits on available resources for each channel as well as the interaction among resources limits an individual's performance on various tasks. Nonetheless, research has shown that multimodal interfaces (i.e., those using multiple resource channels) enable better performance for specific learning tasks by expanding students' working memory capacity (Mousavi, Low, & Sweller, 1995). Other studies have demonstrated performance advantages for multimodal interfaces displaying dynamic map data (Oviatt, DeAngeli, & Kuhn, 1997). Because the raised condition displays spatial information more clearly, helping to move through the stages of perception and processing of the multiple resource theory, we expected it to provide lower workload.

Matching displays and task type to support performance. With the introduction of novel technologies into a classroom setting, understanding how that technology impacts the perceptual and cognitive processes of learners can help instructors decide which technology is most applicable. Past research indicates that the appropriate display is dependent on the task at hand (Haskell & Wickens, 1993; Herbert & Chen, 2015; St. John, Harvey, Smallman, Oonk, & Cowen, 2000; Wickens, 2000). The research determining the best configuration of display type and the task has been inconsistent, which is further complicated because tasks have varying levels of difficulty within a single display. As an example, one of the more common findings is that 3D displays were better for general understanding and shape tasks and that 2D displays were better for accuracy tasks (St. John, Cowen, Smallman, & Oonk, 2001; Tory, Kirkpatrick, Atkins, & Moller, 2006). Related to this research, it has also shown that there can be challenges translating information from a traditional (2D) map to a more experimental display (Atit, Weisberg, Newcombe, & Shipley, 2016). Previous research has indicated that student learners prefer topographic maps that use 3D cues over more traditional 2D cues due to ease of interpretation (Rapp, Culpepper, Kirkby, & Morin, 2007). The ARES is similar to a 3D display in that it provides additional elevation (e.g., height) information as well as length and width.

There is existing research looking at the ability to understand maps using novel displays. In one study, Carbonell Carrera, Avarvarei, Chelariu, Draghia, and Avarvarei (2016) held a workshop that either taught just 2D maps or 2D maps plus a digital 3D and physical representation. They found significant differences in pre–post knowledge scores for the experimental groups and then found a significant difference between the 2D condition and the 3D digital and physical condition. They note that when trying to determine the steepest slope or locating terrain features that using 3D visualizations can help.

Wickens and Carswell noted in their work on the proximity compatibility principle (1995) that the actual features of the display are less important than how those display elements map to the mental model that the user is attempting to understand (i.e., the structure of the information). When considering using displays to ensure measurable benefits, Dixon, Fitzhugh, and Aleva (2009) recommended using a hybrid of 2D and 3D perspective views, which speaks to the need for this research. Regardless of the task, our research follows Christopher Wickens' concept of using 3D displays to support the integration of information from displays to support comprehension (Wickens, Merwin, & Lin, 1994).

User engagement across displays

According to qualitative data collected during a pilot study, using a raised surface to represent terrain information provides greater engagement (Boyce et al., 2016). Engagement is a multidimensional concept consisting of cognitive activity (mental effort), motivational orientation (approach vs. avoidance), and affect changes (Fairclough, Ewing, & Roberts, 2009). Further, research has characterized engagement as the quality of the user experience that emphasizes the positive aspects of the experience including the ability to attract and maintain the user's interest in the technology (Lalmas, O'Brien, & Yom-Tov, 2014).

Because the experiment compares a traditional 2D display of tactical information to a novel raised display, there is a possibility that a novelty effect will exist, leading to higher engagement for the novel display (McIntire et al., 2014). Previous research has demonstrated that novel problem-solving interfaces can serve to increase student engagement. Rowe, Shores, Mott, and Lester (2011) found that students who worked with a narrative-centered learning environment had higher levels of engagement (more motivated, greater focus, and more attention) which correlated to improved learning gains and problem-solving skills irrespective of prior knowledge. Likewise, in studying an alternate reality game, Liu and colleagues found that the GPS-based game increased student motivation, creativity, and exploration, more than its paper-based counterpart (2009).

Measuring engagement requires the proper use of metrics. O'Brien and colleagues have extensively examined the dimensions of engagement and advocate for an approach, which includes mixed-methods consisting of self-report, physiological, and performance-based factors (O'Brien & Cairns, 2015; O'Brien & Lebow, 2013). To further support research on user engagement, O'Brien and Toms (2010) developed the User Engagement Scale (UES), which is the primary measure of engagement in this study.

Research questions and hypotheses

The purpose of this experiment is to answer the question: How does the additional information provided by the raised terrain surfaces support the instruction of military tactics concepts? More specifically, do the raised surfaces increase performance (time on task and accuracy), reduce workload on learners, and increase their engagement and willingness to explore the technology?

We hypothesized that increasing the amount of relief would increase accuracy and reduce time on task (Hypothesis 1) due to the emergent cues presented in the raised display. Because the raised condition assists in providing height and depth cues of the terrain to participants, we expected that participants would have reduced global workload, mental demand, effort, and frustration according to the NASA TLX (Hypothesis 2). We also hypothesized that participants in the raised condition would show increased levels of user engagement as measured by the UES due to data from the pilot study (Boyce et al., 2016; Hypothesis 3). Finally, we hypothesized that there would be a correlation between the perceived usability subscale of the UES and the perceived system usability based on the System Usability Scale (SUS; Hypothesis 4). We based this hypothesis on findings from O'Brien and Lebow (2013), who used the UES and SUS in a similar study and found a Pearson's correlation coefficient of 0.86 between the UES perceived usability subscale and the SUS score.

Method

Participants

Sixty-two cadets (45 males, 17 females) from the USMA participated in this study. We recruited participants from introductory psychology courses via USMA Sona Systems, and they received extra credit in their classes for their participation. Fifty-nine participants were first-year cadets, and three were second-year cadets. Of the total number of participants, 29 experienced the flat condition first, and 33 experienced the raised condition first.

Upon completion of the study, a post hoc power analysis was run to determine whether the final sample of 62 individuals was adequate to obtain statistically significant effects. With a medium effect size of .25 and an alpha level of .05, power for a repeated-measures analysis of variance (ANOVA) with four variables and two groups indicated .998 power, which was well above the desired .80. Therefore, the sample was large enough to achieve significant results.

Equipment

ARES

The ARES (Amburn et al., 2015) is 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. The ARES is an example of central perspective projection (Willett et al., 2015). Central perspective involves projecting directly, without tilt, onto a surface below, and it provides depth cues information for the terrain. Because the ARES projects a 2D image onto a 3D plane, it is important to recognize differences in how the displays project information. For this experiment, we used the ARES projection technology combined with terrain boards rather than the actual sand table (see Figure 2). The reasoning behind the

Figure 2. Maps projected onto flat (left) and raised (right) surfaces.

use of the terrain boards was to eliminate variability in topography due to someone accidentally touching the sand since the terrain boards could not be modified based on accidental touch.

Generalized intelligent framework for tutoring

Generalized intelligent framework for tutoring (GIFT; Sottilare, Brawner, Goldberg, & Holden, 2012; Sottilare, Brawner, Sinatra, & Johnston, 2017) is an open source adaptive tutoring engine that can provide tailored learning experiences to each student based on their attributes (e.g., states and traits) and their preferences. For this research experiment, GIFT served as a content delivery and data acquisition tool. We chose GIFT because of its ability to structure lesson content and gather assessment data, and we presented questions on a monitor above the surfaces. Participants needed to answer by clicking the appropriate choice via mouse click (see Figure 3). GIFT managed the presentation of content with questions shown on the monitor and maps shown on the surfaces via projection. It also logged each multiple-choice answer for analysis, as well as the time taken to answer. GIFT linearly delivered content, providing the same content in the same way to each student with no adaptations.

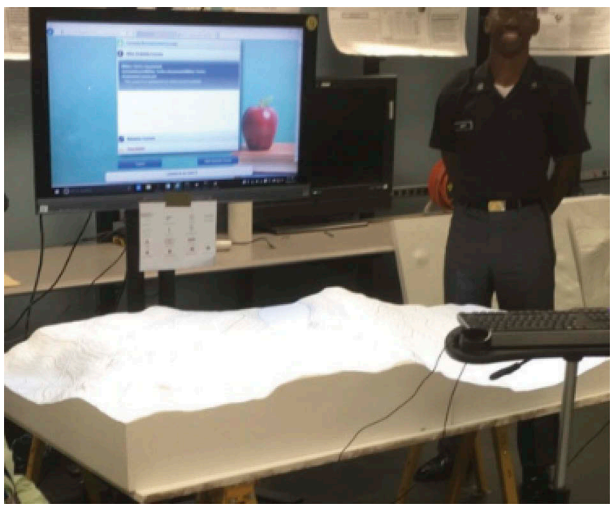

Figure 3. Experimental setup using GIFT and ARES.

Question and terrain development

As a part of cadet instruction at USMA, there is a series of military science courses. These courses focus on tactical decision making at various Army echelons. One of the authors, who is also a military faculty member at USMA, developed the 24 military tactics questions using content from an entry-level military science course. We designed the tactics questions with small-unit organizations as the focus of the scenarios. The authors derived the questions after referencing small-unit Army manuals. These manuals included the Supplemental Handbook 21–76: The Ranger Handbook and the Army Training Publication 3–21.8, The Infantry Rifle Platoon and Squad (US Department of the Army, 2007, 2011a). Below (Figure 4) is an example of a tactics question that we presented to participants. Each of the tactics questions presented four choices to the participant: three distractors and one correct answer in a multiple-choice format. We scored the questions according to percentage correct on each experimental condition (i.e., percentage correct of the 12 flat questions versus percentage correct of the 12 raised questions).

To explore the understanding of military tactics across multiple environments, we used four distinct terrains. We realized a need for a higher level of complexity because using one terrain in the pilot study yielded a ceiling effect concerning accuracy. With the increased variability of terrain, we expected that complexity would increase (Boyce et al., 2016). These terrains included the following: mountainous forest in New York (Map 1), low-level swamp terrain in Louisiana (Map 2), elevated desert terrain in California (Map 3), and mostly flat desert terrain in southern New Mexico (Map 4). The questions received several rounds of validation by Army subject matter experts in dismounted tactics and training and military science course instruction. We developed three questions for each of the four terrains for the two experimental conditions, yielding 24 questions.

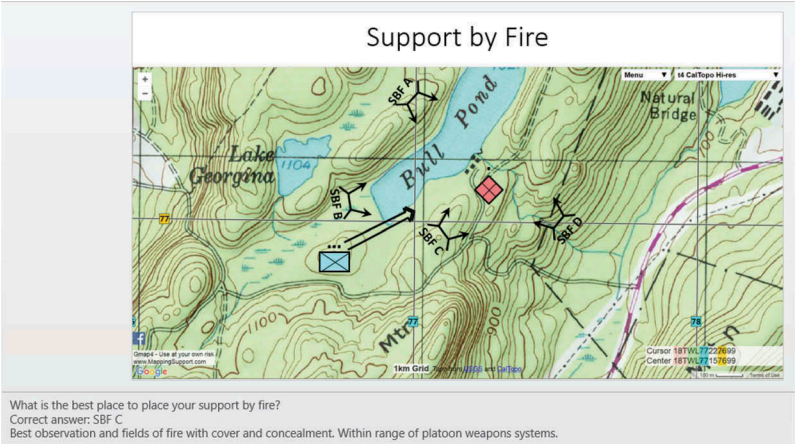

Figure 4. Sample tactics question.

Procedure

Upon arriving, participants were randomly assigned to one of the two ordering conditions: (a) the flat condition followed by the raised condition, or (b) the raised condition followed by the flat condition. After reading and signing a consent form approved by the Institutional Review Board of the Army Research Laboratory, participants received a brief explanation of the experimental task and how to respond to the test questions.

Participants then reviewed a series of training slides that went over tactical symbology to ensure they were not confused about what the displays were showing tactically. Following their review, we gave participants the opportunity to ask questions, with the stipulation that we might not be able to answer some questions until after the experiment.

Finally, we presented 24 military tactics questions while still showing the map and collected their responses via GIFT. We divided the 24 questions into two halves, with 12 questions presented per surface type. The surface types were manually changed in front of the participants. Subject matter experts validated the two question sets to ensure that they were of equal difficulty. After each question group, participants completed a postassessment consisting of the NASA-TLX, a modified UES, and the SUS. The experiment took approximately 60 min to complete.

Dependent variables

Time on task

We examined time on task by breaking out each condition into four variables based on map type (eight variables). We measured time on task by the amount of time, measured in seconds, that cadets took to answer each question. This was measured by when the map appeared to when a choice was selected, and the participant pressed the submit button. We captured time on task using system log files captured through GIFT.

Accuracy

We examined accuracy by dividing each condition into four variables based on map type (eight variables). We measured accuracy as the number of questions that the cadets answered correctly against the total number of questions presented in each experimental condition (12).

Workload (NASA TLX)

We measured workload using the NASA TLX, which is a multidimensional scale examining six types of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). The NASA-TLX raw scales range from 0 to 100, with 16 pairwise comparisons, which ask the participant to select which of two types of workload are more demanding. We examined the NASA-TLX both regarding its global scale and the individual rating scales for each condition (14 variables). Previous studies using the NASA-TLX have indicated reliability values in the .70 to .90 range, which demonstrates optimal reliability (Battiste & Bortolussi, 1988; Hoonakker et al., 2011; Xiao, Wang, Wang, & Lan, 2005).

UES

The UES has six associated dimensions: aesthetics (AE), endurability (EN), focused attention (FA), felt involvement (FI), novelty (NO), and perceived usability (PU; O'Brien & Toms, 2008, 2010). These dimensions do not amount to an aggregate score but stand on their own as measures. Therefore, there are

six measures across the two conditions (12 variables). The UES scale has demonstrated high reliability (Cronbach's α = .92) in previous research (Wiebe, Lamb, Hardy, & Sharek, 2014). The reliability is confirmed by O'Brien and Cairns (2015), who found reliability values that exceeded .90 on the FA, AE, and NO subscales. These high values could indicate overlap between subscales. They also found that the FI, EN, and PU subscales demonstrated optimal reliability, between .70 and .90 (O'Brien & Cairns, 2015). For this experiment, we decided to maintain all six of the subscales to use the UES to gather the most robust data and maintain its original factor structure.

The UES has seven response options per question ranging from strongly disagree to strongly agree in a Likert scale format. The AE subscale consists of five questions such as "This display is attractive." The EN subscale was also five items such as "I would recommend this display to my friends and family." The FA subscale consists of seven questions such as "During the experience, I let myself go." The FI subscale has three questions such as "This experience was fun," and the NO subscale has two questions such as "I continued to examine the display out of curiosity." Finally, the PU subscale consists of eight questions such as "I found the display confusing to use."

SUS

The SUS is a 10-item questionnaire with five response options per question ranging from strongly disagree to strongly agree in a Likert scale format (Brooke, 1996). The items include questions such as, "I thought the system was easy to use" and "I found the system unnecessarily complex." Because the target of this study was specifically on displays, the SUS was modified, replacing the word *system* with *display*.

We report the SUS as an aggregate score from its ten questions; thus, there is one score for each of the two conditions (2 variables). Data values emerging from the SUS will have a range from 0 to 100, with an SUS score of 70 indicating acceptable usability (Bangor, Kortum, & Miller, 2009). The SUS has been cited in over 3,500 papers and provides an easy to use metric with which to compare the UES (Bangor et al., 2009; Brooke, 2013). Analysis across 10 years of research indicated that the SUS demonstrated strong reliability, Cronbach's α = .91 (Bangor et al., 2009).

Research design

The research design is a mixed design with both between- and within-subjects factors. For time on task and accuracy, it was a 2 Order (Between) × 2 Condition (Within) × 4 Map (Within) design. For the NASA TLX, UES, and SUS, it was a 2 Order (Between) × 2 Condition (Within) design. The design consisted of two independent variables: condition, either flat or raised terrain board surface, and order, which describes which order of presentation the participant received. Each participant was shown both conditions but received only one order.

Results

Checking for missing data

Dong and Peng (2013) citing Schafer (1999) state that missing data of 5% or less should not have an impact on a study's results. After preparing the data of this study for analysis, and looking across cases and variables, there was less than 5% of missing information in the dataset, indicating no impact on the study.

To assess whether the pattern of missing values was missing completely at random (MCAR), we conducted Little's MCAR test (1988). The null hypothesis of Little's MCAR test is that the pattern of the data is MCAR and follows a χ^2 distribution. The results revealed that the pattern of missing values in the data is MCAR, $\chi^2(103) = 136.21$, $p = 1.00$, indicating that the data is missing completely at random, so the final sample size will not be affected by list-wise or pairwise deletion when running the analysis.

Testing of assumptions

Normality

Because of the overall sample size ($N = 62$), we examined Shapiro-Wilks and Kolmogorov-Smirnov tests and found the tests to be nonnormally distributed. The histograms, Q-Q plots, and box plots indicated that the variables have adequately normal distributions. Because ANOVAs are robust to deviations of normality for analysis, we decided to proceed with the analysis (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010).

High-level results for variables of interest

Table 1 shows the means and standard deviations for the variables of interest. At a high level, performance variables were driven by a practice effect, producing nonsignificant results for accuracy and time on task. Statistically significant results were obtained for workload and engagement.

Table 1. Summary of means and standard deviations.

	Flat		Raised	
Variable	*M*	*SD*	*M*	*SD*
Total time on task (s)	257.08	95.40	292.71	94.09
Accuracy	0.58	0.15	0.54	0.19
NASA-TLX Global	55.03	15.70	49.20	14.64
UES-Aesthetics	3.73	0.84	4.50	0.45
UES-Endurability	3.62	0.53	4.02	0.42
UES-Focused attention	2.76	0.75	3.05	0.77
UES-Self-involvement	3.56	0.74	4.16	0.56
UES-Novelty	3.32	1.00	4.28	0.50
UES-Perceived usability	2.37	0.63	1.96	0.45

Note. UES = User Engagement Scale.

H1: When level of relief is increased, there will be greater levels or performance, such that there are higher levels of accuracy and faster time on task.

Preliminary analysis to support H1

Because this was a within-subjects design, we suspected that order may be impacting the results. We ran a mixed ANOVA for time on task (broken down according to map) to determine impact of order and condition. The results indicated a significant interaction between order and condition, Wilks' Lambda = .267, $F(4,52) = 35.70$, $p < .001$, $\eta_p^2 = .733$. Looking at the univariate tests, the interaction effect of order and condition for all four map types was significant, $p < .001$. For Map 1, $F(1,55) = 108.94$, $p < .001$, $\eta_p^2 = .665$; Map 2, $F(1,55) = 20.87$, $p < .001$, $\eta_p^2 = .275$; Map 3, $F(1,55) = 32.46$, $p < .001$, $\eta_p^2 = .371$; Map 4, $F(1,55) = 26.52$, $p < .001$, $\eta_p^2 = .325$, indicating a practice effect. The practice effect shows that a participant's speed to answer a question was not based on the map itself, but rather if the participant had seen the map before.

Based on this finding, we ran a mixed ANOVA for accuracy (broken down according to map) to determine impact of order and condition. The results again indicated a significant interaction between order and condition, Wilks' Lambda = .538, $F(4,57) = 12.258$, $p < .001$, $\eta_p^2 = .462$. Looking at the univariate tests, the interaction effect of order and condition for three of the four map types was significant, $p < .05$. For Map 1, $F(1,60) = 8.23$, $p = .006$, $\eta_p^2 = .121$; Map 3, $F(1,60) = 23.07$, $p < .001$, $\eta_p^2 = .278$; and Map 4, $F(1,60) = 26.96$, $p < .001$, $\eta_p^2 = .310$. There was also a significant main effect for condition for Map 4, $F(1,60) = 4.07$, $p = .048$, $\eta_p^2 = .063$. This is also indicative of a practice effect. Map 2 was not statistically significant, $F(1,60) = 3.62$, $p = .062$, $\eta_p^2 = .06$, however, at a significance level of .06 with a larger sample, this marginally significant result may become significant.

Follow-up pairwise comparisons using the Bonferroni Correction indicated a significant difference for participants on Map 1 ($p = .03$) when the flat condition is presented first, with participants answering more questions correctly on the raised condition ($M = .62$, $SD = .29$), as opposed to the flat condition ($M = .44$, $SD = .24$). When the raised condition is presented first, participants answered more questions correctly on the flat condition ($M = .55$, $SD = .34$), as compared to the raised condition ($M = .41$, $SD = .31$). Regardless of which condition we showed first, participants answered more accurately on the second condition we showed. For Maps 3 and 4, there are significant differences on both when the flat is displayed first (Map 3, $p = .023$; Map 4, $p = .033$) and when the raised is displayed first (both maps, $p < .001$). The second map shown had higher accuracy scores than the first map. Map 2 showed the same effect; however, results were nonsignificant ($p > .05$).

Main analysis for H1

We ran a mixed ANOVA for time on task (broken down according to map type) to determine the impact of order and condition for both accuracy and time on task when answering tactics questions. Because high accuracy and low time on task indicate positive performance, we produced z-scores for both accuracy and time on task to provide easier comparison of results. We inverse transformed time on task variables before standardization to indicate that higher values represented less time to complete a task. The results indicated a significant interaction between order, time on task, and accuracy: Wilks' Lambda = .786, $F(1,57) = 3.54$, $p = .012$, $\eta_p^2 = .214$.

In univariate tests, the interaction effect of order, time on task, and accuracy for Map 1 was significant: $F(1,57) = 6.17$, $p = .004$, $\eta_p^2 = .141$. However, the other three maps were not significant, $p > .05$. Follow-up pairwise comparisons using the Bonferroni Correction for Map 1 indicated that irrespective of whether they received the flat condition or the raised condition first, the second condition presented always exhibited reduced time on task (see Figure 5) and increased levels of accuracy (see Figure 6). The practice effect supports the findings of the preliminary analysis. All comparisons were significant, $p < .001$.

H2: Upon completion of a condition, individuals in the raised condition will exhibit lower levels of global workload, as well as lower levels of the subscales of mental demand, effort, and frustration as compared to the flat condition, according to the NASA-TLX. We did not expect differences between the flat and raised conditions for the physical demand, temporal demand, and performance subscales.

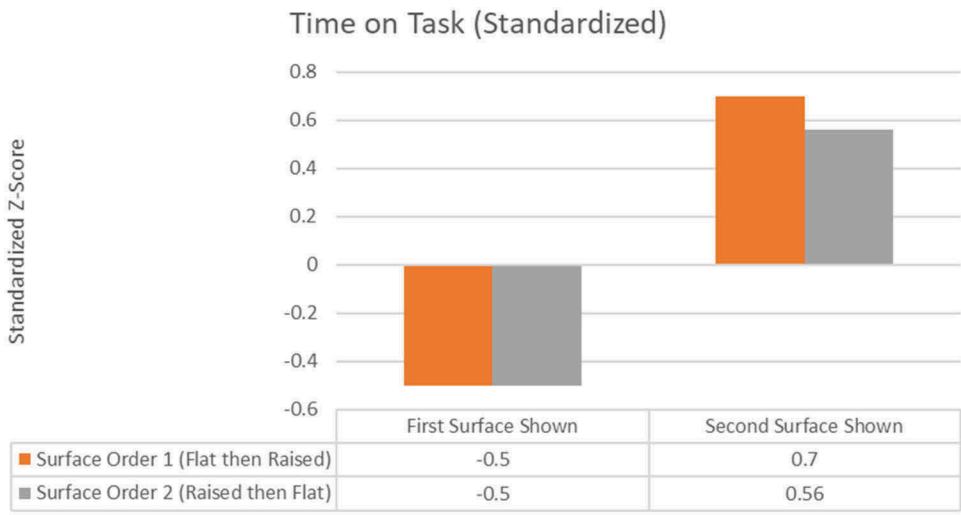

Figure 5. Time on task according to z-score (whichever surface is presented first has poorer performance than the surface presented second).

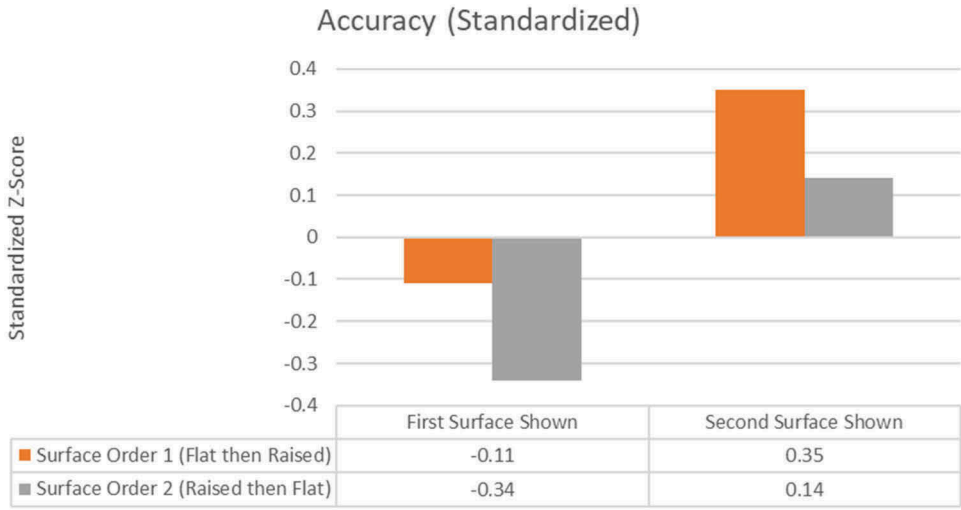

Figure 6. Accuracy according to z-score (whichever surface is presented first has poorer performance than the surface presented second).

We ran a within-subjects ANOVA for workload (NASA-TLX) to determine impact of surface. The results indicated a significant main effect for condition: Wilks' Lambda = .673, $F(7,44) = 3.051$, $p = .01$, $\eta_p^2 = .327$. Individuals in the raised condition ($M = 49.20$, $SD = 14.64$) reported lower levels of global workload than individuals in the flat ($M = 55.03$, $SD = 15.70$) condition. It was also significant for the following subscales (see Table 2), but the other three subscales were not significant, $p > .05$.

Table 2. Significance values for NASA-TLX.

Variable	*F(1,50)*	*p*	η^2_p
Mental demand	9.53	.003	.160
Performance	8.17	.006	.140
Frustration	8.00	.007	.138

Individuals in the flat condition reported higher mental workload ($M = 62.88$, $SD = 18.51$) than individuals in the raised condition ($M = 55.65$, $SD = 17.67$). Individuals in the flat condition also reported higher levels of performance, where higher numbers indicate poorer subjective performance and lower numbers indicate positive subjective performance, ($M = 53.22$, $SD = 21.79$) and higher levels of frustration ($M = 40.43$, $SD = 25.05$) than individuals in the raised condition reported for performance ($M = 44.10$, $SD = 19.04$) and frustration ($M = 31.94$, $SD = 21.32$). See Figure 7 below.

H3: Participants in the raised condition will exhibit higher levels of engagement as measured by the UES than participants in the flat condition.

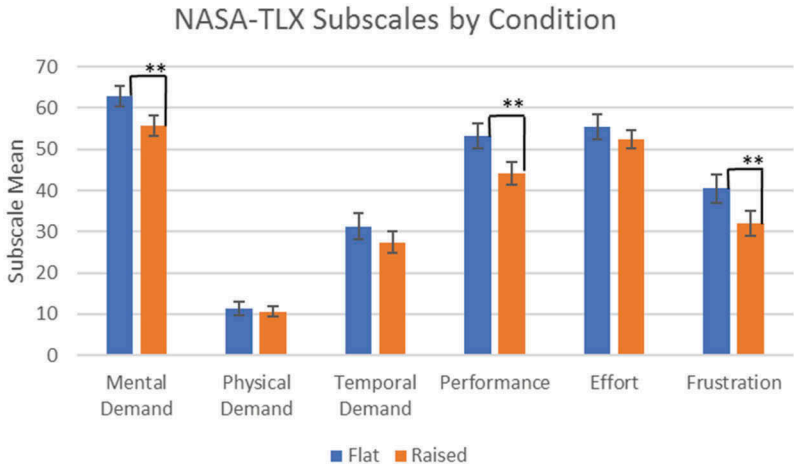

Figure 7. NASA-TLX according to subscale (Error bars indicate SEM).

We ran a within-subjects ANOVA for user engagement (UES) to determine impact of the terrain condition. The results indicated a significant main effect for the raised terrain condition: Wilks' Lambda = .428, $F(6,52) = 11.60$, $p < .001$, $\eta_p^2 = .572$.

Examining the univariate tests, the main effect for terrain condition was significant for all six dimensions of the UES (see Table 3 and Figure 8):

It should be noted that for the PU subscale, the flat condition outperformed the raised condition (see Figure 8). In the study, the raised condition took more time to set up due to the presence of terrain boards. This may have contributed to a lower PU score for the raised condition, especially because the PU scale contains items such as "This experience was demanding." To investigate this, we conducted ancillary analysis looking at the SUS. This analysis showed the opposite of the PU scale, with the perception of usability for the raised condition rated higher ($M = 79.27$, $SD = 9.71$) than the flat condition ($M = 69.11$, $SD = 14.86$).

> H4: There will be a correlation between the perceived usability subscale of the UES and the perceived system usability based on the SUS.

We ran Pearson correlations to assess the correlation between the perceived usability dimension of the UES and the overall score of the SUS per condition. When looking at the flat condition, the two variables are strongly negatively correlated: $r(58) = -.739$, $p < .001$. The pattern of negative correlation also occurs in the raised condition; the two variables are also strongly negatively correlated: $r(58) = -.55$, $p < .001$. This finding is consistent with the ancillary analysis discussed above in Hypothesis 3 such that the UES and the SUS were showing opposite results.

Table 3. Significance values for user engagement scale subscales.

Subscale	$F(1,57)$	p	η^2_p
Aesthetics	17.14	< .001	.451
Endurability	28.83	< .001	.336
Focused attention	19.15	< .001	.251
Felt involvement	51.26	< .001	.473
Novelty	46.26	< .001	.448
Perceived usability	27.13	< .001	.322

Discussion

The purpose of this experiment was to examine the effect of displays onto different topographical surfaces to support tactics assessment as measured by workload, engagement, and performance. We designed the raised surface to mimic a military sand table, whereas the flat surface represents a paper map. We expected that the raised surface would provide additional information through the emergent visual cues of the surface, which would lead to an increase in performance (Pomerantz & Pristach, 1989). We also hypothesized that the raised surface would be easier to distinguish and understand, which would reduce workload and improve user engagement in working with the display.

In Hypothesis 1, we expected that as the amount of relief increased, there would be a decrease in time on task and an increase in accuracy. We could not support this hypothesis since a practice effect drove the relationship such that increased exposure to the assessment maps led to improved performance for the second condition shown, regardless of surface type. In hindsight, the two experimental groups were similar to one another, and with the exception of Map 1, which had mountainous terrain, differences between surfaces were not substantial enough to demonstrate significant differences. In

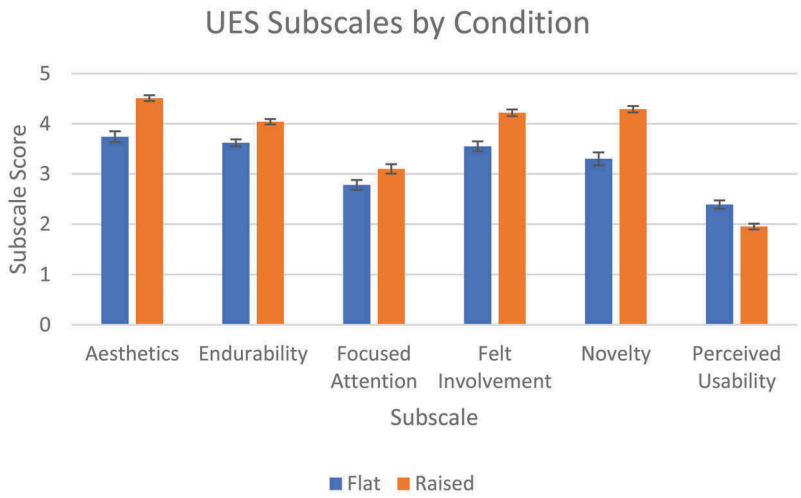

Figure 8. UES subscales by condition (Error bars indicate SEM).

addition, the increased performance on the flat display could be due to the familiarity of cadets with 2D maps.

Because we expected the raised display to help integrate information on elevation in conjunction with each assessment question, the raised display was hypothesized to have lower workload (Hypothesis 2; Wickens et al., 1994). The flat condition, where individuals would have to use the visual information presented in the map and mentally translate from 2D to 3D to answer questions, was expected to have higher workload (Rapp et al., 2007). This hypothesis was supported with individuals in the flat condition demonstrating higher workload. This result is consistent with previous research looking at mental demand on military-related tasks where augmented displays supported reduced mental workload (Davis, 2006). For the subscale of performance on the NASA-TLX, the higher performance results on the flat condition indicate that the participants felt that they did poorer since lower scores indicate good perceived performance on this subscale. These results, along with the trends in workload discovered in the pilot study of this research, indicate the potential for a raised topographical display to reduce workload while not negatively impacting performance, compared to their flat counterparts. Longitudinal experiments controlling for level of information and variations in surface type are needed to examine workload in greater depth.

In Hypothesis 3, we expected that the raised condition would lead to increased user engagement. Because of the finding supporting the lower workload of the raised condition, it is not surprising that there was also a finding supporting increased user engagement with the raised display, supporting Hypothesis 3. It is easier to explain the result of increased engagement when looking at the questions of the UES in more detail. Several of the questions discuss characteristics that would be favorable to the raised display, such as being aesthetically pleasing, being fun, inciting the user's curiosity, and being drawn into the training experience.

For the PU scale, which was the only scale in which the flat condition outperformed the raised, participant bias was identified by the researchers as a possible factor. Possible sources of bias included differences in setup time for each terrain. The raised condition had to be set up by two individuals, which required physically changing terrains every three questions, whereas the flat condition did not require additional set up once started. The participants saw the difference in set up time, and this may have contributed to the difference in usability scores.

The majority of the scales, however, indicated favorably for the raised condition. This result may also be explained due to the novelty of the raised surface, as opposed to a flat surface, which is boring, traditional, and less aesthetically pleasing (McIntire et al., 2014). Research has shown that user preference does not equate to user performance (Andre & Wickens, 1995). Therefore, using more in-depth metrics like the UES can provide insight when examining perceptions between different surfaces.

Hypothesis 4 investigated the relationship between the SUS and the PU scale of the UES. We chose to investigate this relationship because the SUS is often looked at as the gold standard for usability and being able to understand how the two measures relate to one another can assist in building sets of metrics when assessing user experience and different display surfaces. We expected that, based on previous research looking at the correlation between these two measures, the

results in one would correlate with the other (O'Brien & Cairns, 2015). The negative correlation reveals the impact of the UES question wording on the overall score. It is possible that because several of the questions were negatively worded (i.e., "I felt frustrated using this display"), participants not agreeing with these statements could drive lower scores. The implication of this finding encourages caution when examining subjective rating scales, as potentially misunderstanding questions could lead to inflated or deflated scores. This finding further stresses the importance of using multiple subjective measures as the underlying psychological constructs might be slightly different, yielding a more complete picture of user experience.

Limitations of the study

Potential changes for future studies include improved management of sample size. This study had unequal sample sizes for each of the experimental conditions. We generated the ordering for the experimental conditions via a random number generator that assumed that all participant data would be analyzed. However, a system error, errors in data collection, and situations beyond experimental control (e.g., a fire drill) caused the sample sizes to be uneven.

A between-subjects design would also have improved the study. We decided to use a within-subjects design because of the number of participants that were going to be available and the goal of looking for differences between the groups. To do a between-subjects design, roughly double the number of participants would have been required, holding effect size levels and analyses constant. This design had a substantial impact to the research due to the presence of the practice effect between the two conditions.

The milling process for creating the terrain boards had limited amounts of relief, which was another limitation of the study. Although technically possible to scale relief to any level, this was not practically possible because it would render the terrain unable to match the contour lines of the tactical maps. Furthermore, the greater the relief, the higher the cost, which meant that it was costlier to render terrains with high topographic variability (i.e., with extremely high peaks and low valleys). However, the findings indicate variation in performance according to the amount of relief, leading to possibilities for future research.

The number of tactics questions available for each terrain condition was another limiting factor of this study. We had limited time in which to execute the experiment (one semester) and wanted to conduct a thorough validation of the questions. These factors resulted in a limited number of questions. In future iterations of this study, a more substantial number of questions would allow better comparison of specific tactical conceptual knowledge.

Potential avenues for expansion

This research represents one small component of a much larger research effort looking at the interaction of displays and systems to support battlespace visualization. The goal of the ARES program is to investigate a common operating picture at the point of need for soldiers using innovative user interfaces. Projecting imagery onto different surfaces is a starting point that can help to understand how to project content using mobile devices or large-scale virtual environments (e.g., holodecks where an entire room is projection technology). It can also assist in the design of content for heads-up displays such as the Microsoft HoloLens where the insights on visual cues and workload/engagement can extend to augmented reality. At the time of this writing, there is currently a new research study about to begin looking at the effects of the sand table versus a virtual environment versus augmented reality, which is a follow up from Schmidt-Daly and colleagues work (2016). The ARES is also being looked at to support land navigation training via mobile application design, where a cadet first does the classroom training using the ARES and then receives additional instructions in the wild via a cellular connected land navigation course (Goldberg, Davis, Riley, & Boyce, 2017).

Conclusion

This experiment sought to verify and extend the findings of a previous study investigating the use of the ARES to support military tactics decisions. The findings indicate that the amount of relief provided in a topographic map can alter performance, increase engagement, and reduce workload. The applicability of the ARES to support military tactics instruction was confirmed, and the groundwork for further research investigating more complex terrains is warranted. From a human factors psychology perspective, this research contributes to an expanding base in the literature that supports methods to increase engagement to increase performance during the instruction of military tactics tasks.

Funding

This research was supported in part by grants from the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-12-2-0019.

References

Amburn, C. R., Vey, N. L., Boyce, M. W., & Mize, J. R. (2015). *The augmented reality sandtable (ARES)* (No. ARL-SR-0340). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Andre, A. D., & Wickens, C. D. (1995). When users want what's not best for them. *Ergonomics in Design: the Quarterly of Human Factors Applications*, *3*(4), 10–14. doi:10.1177/106480469500300403

Atit, K., Weisberg, S. M., Newcombe, N. S., & Shipley, T. F. (2016). Learning to interpret topographic maps: Understanding layered spatial information. *Cognitive Research: Principles and Implications*, *1*(1), 2. doi:10.1186/s41235-016-0002-y

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Barfield, W., & Rosenberg, C. (1995). Judgments of azimuth and elevation as a function of monoscopic and binocular depth cues using a perspective display. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, *37*(1), 173–181. doi:10.1518/001872095779049453

Battiste, V., & Bortolussi, M. (1988). Transport pilot workload: A comparison of two subjective techniques. *Proceedings of the Human Factors Society Annual Meeting*, *32*(2), 150–154. doi:10.1177/154193128803200232

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)* (No. ARL-TR-7792). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Brewster, F. W. (2002). Using tactical decision exercises to study tactics. *Military Review*, *82*(6), 3.

Brooke, J. (1996). SUS – A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (Vol. 189, pp. 4–7). London, England: Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Carbonell Carrera, C., Avarvarei, B. V., Chelariu, E. L., Draghia, L., & Avarvarei, S. C. (2016). Map-reading skill development with 3D technologies. *Journal of Geography*, *116*, 1–9.

Davis, B. M. (2006). Effects of tactical navigation display modality on navigation performance, situation awareness, and mental workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(17), 2089–2093. doi:10.1177/154193120605001780

Dixon, S., Fitzhugh, E., & Aleva, D. (2009). *Human factors guidelines for applications of 3D perspectives: A literature review*. Paper presented at the display technologies and applications for defense, security, and avionics III, Orlando, FL.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. doi:10.1186/2193-1801-2-222

Fairclough, S. H., Ewing, K. C., & Roberts, J. (2009). *Measuring task engagement as an input to physiological computing*. Paper presented at the 3rd international conference on affective computing and intelligent interaction (ACII 2009), Amsterdam, Netherlands.

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., & Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied Ergonomics*, *52*, 95–103. doi:10.1016/j.apergo.2015.07.009

Galán, F. C., & Beal, C. R. (2012). *EEG estimates of engagement and workload predict math problem solving outcomes*. Paper presented at the international conference on user modeling, adaptation, and personalization, Montreal, Canada.

Goldberg, B., Davis, F., Riley, J. M., & Boyce, M. W. (2017). *Adaptive training across simulations in support of a crawl-walk-run model of interaction*. Paper presented at the international conference on augmented cognition, Vancouver, BC, Canada.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*(4), 647–658. doi:10.1109/21.17382

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Haskell, I. D., & Wickens, C. D. (1993). Two- and three-dimensional displays for aviation: A theoretical and empirical comparison. *The International Journal of Aviation Psychology*, *3*(2), 87–109. doi:10.1207/s15327108ijap0302_1

Hendrix, C., & Barfield, W. (1995, March 11–15). *Presence in virtual environments as a function of visual and auditory cues*. Paper presented at the virtual reality annual international symposium, 1995, Research Triangle Park, NC.

Herbert, G., & Chen, X. (2015). A comparison of usefulness of 2D and 3D representations of urban planning. *Cartography and Geographic Information Science*, *42*(1), 22–32. doi:10.1080/15230406.2014.987694

Hoonakker, P., Carayon, P., Gurses, A. P., Brown, R., Khunlertkit, A., McGuire, K., & Walker, J. M. (2011). Measuring workload of ICU nurses with a questionnaire survey: The NASA task load index (TLX). *IIE Transactions on Healthcare Systems Engineering*, *1*(2), 131–143. doi:10.1080/19488300.2011.609524

Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831–852. doi:10.1007/s10648-015-9352-0

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014). Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *6*(4), 1–132. doi:10.2200/S00605ED1V01Y201410ICR038

Li, N., Willett, W., Sharlin, E., & Sousa, M. C. (2017). *Visibility perception and dynamic viewsheds for topographic maps and models*. Paper presented at the 5th symposium on spatial user interaction, Brighton, England.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. doi:10.1080/01621459.1988.10478722

Liu, T. Y., Tan, T. H., & Chu, Y. L. (2009). Outdoor natural science learning with an RFID-supported immersive ubiquitous learning environment. *Journal of Educational Technology & Society*, *12*(4), 161–175.

Martin, R. C., & Pomerantz, J. R. (1978). Visual discrimination of texture. *Perception & Psychophysics*, *24*(5), 420–428. doi:10.3758/bf03199739

McIntire, J. P., Havig, P. R., & Geiselman, E. E. (2014). Stereoscopic 3D displays and human performance: A comprehensive review. *Displays*, *35*(1), 18–26. doi:10.1016/j.displa.2013.10.004

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, *87*(2), 319–334. doi:10.1037/0022-0663.87.2.319

Newcombe, N. S., Weisberg, S. M., Atit, K., Jacovina, M. E., Ormand, C. J., & Shipley, T. F. (2015). The lay of the land: Sensing and representing topography. *Baltic International Yearbook of Cognition, Logic and Communication*, *10*(1), 6. doi:10.4148/1944-3676.1099

O'Brien, H. L., & Cairns, P. (2015). An empirical evaluation of the User Engagement Scale (UES) in online news environments. *Information Processing & Management*, *51*(4), 413–427. doi:10.1016/j.ipm.2015.03.003

O'Brien, H. L., & Lebow, M. (2013). Mixed-methods approach to measuring user experience in online news interactions. *Journal of the American Society for Information Science and Technology*, *64*(8), 1543–1556. doi:10.1002/asi.22871

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. doi:10.1002/asi.20801

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69. doi:10.1002/asi.21229

Oviatt, S., DeAngeli, A., & Kuhn, K. (1997). *Integration and synchronization of input modes during multimodal human-computer interaction*. Paper presented at the referring phenomena in a multimedia context and their computational treatment, Madrid, Spain.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. doi:10.1207/S15326985EP3801_1

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. doi:10.1109/3468.844354

Pomerantz, J. R., & Pristach, E. A. (1989). Emergent features, attention, and perceptual glue in visual form perception. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 635.

Proctor, R. W., & Van Zandt, T. (2018). *Human factors in simple and complex systems*. Boca Raton, FL: CRC press.

Rapp, D. N., Culpepper, S. A., Kirkby, K., & Morin, P. (2007). Fostering students' comprehension of topographic maps. *Journal of Geoscience Education*, *55*(1), 5–16. doi:10.5408/1089-9995-55.1.5

Reichelt, S., Häussler, R., Fütterer, G., & Leister, N. (2010). *Depth cues in human visual perception and their realization in 3D displays*. Paper presented at the SPIE defense, security, and sensing, Orlando, FL.

Roscoe, S. N. (1968). Airborne displays for flight and navigation. *Human Factors*, *10*(4), 321–332. doi:10.1177/001872086801000402

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, *21*(1–2), 115–133.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. doi:10.1177/096228029900800102

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology*, *6*, 147–151. doi:10.1027/1614-2241/a000016

Schmidt-Daly, T. N., Riley, J. M., Hale, K. S., Yacht, D. P., & Hart, J. (2016). *Augmented reality sandtables (ARESs) impact on learning* (No. ARL-CR-0803). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. doi:10.1037/0278-7393.4.6.592

Smith, R. (2010). The long history of gaming in military training. *Simulation & Gaming*, *41*(1), 6–19. doi:10.1177/1046878109334330

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). *The generalized intelligent framework for tutoring (GIFT)*. Orlando, FL: U.S. Army Research Laboratory.

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). *An updated concept for a generalized intelligent framework for tutoring (GIFT)*. Orlando, FL: U.S. Army Research Laboratory.

St. John, M., Cowen, M. B., Smallman, H. S., & Oonk, H. M. (2001). The use of 2D and 3D displays for shape-understanding versus relative-position tasks. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, *43*(1), 79–98. doi:10.1518/001872001775992534

St. John, M., Harvey, S., Smallman, H., Oonk, H., & Cowen, M. (2000). *Navigating two-dimensional and perspective views of terrain*. San Diego, CA: Space and Naval Warfare Systems Center.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. doi:10.1016/0959-4752(94)90003-5

Sweller, J. (2016). Working memory, long-term memory, and instructional design. *Journal of Applied Research in Memory and Cognition*, *5*(4), 360–367. doi:10.1016/j.jarmac.2015.12.002

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. doi:10.1023/a:1022193728205

Tory, M., Kirkpatrick, A. E., Atkins, M. S., & Moller, T. (2006). Visualization task performance with 2D, 3D, and combination displays. *IEEE Transactions on Visualization and Computer Graphics*, *12*(1), 2–13. doi:10.1109/TVCG.2006.17

Treisman, A., & Paterson, R. (1984). Emergent features, attention, and object perception. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 12–31. doi:10.1037/0096-1523.10.1.12

US Department of the Army. (2007). *The infantry rifle and platoon and squad FM 3-21.8*. Washington, D.C: Army Training and Doctrine Command (TRADOC):

US Department of the Army. (2011a). *Ranger handbook SH 21-76*. Retrieved from http://www.benning.army.mil/infantry/rtb/4thrtb/content/PDF/Handbook.pdf

US Department of the Army. (2011b). *The U.S. army learning concept for 2015 (TRADOC Pam 525-8-2)*. Retrieved from http://sill-www.army.mil/DOTD/divisions/pdd/docs/Army%20Learning%20Model%202015.pdf

Van Beurden, M. (2013). *Interaction in depth* (Thesis). Interaction group, Eindhoven University of Technology, Eindhoven, The Netherlands.

Wickens, C. D. (1991). Processing resources and attention. In D. L. Damos (Ed.), *Multiple task performance* (pp. 3–34). London, England: Taylor & Francis.

Wickens, C. D. (2000). The when and how of using 2-D and 3-D displays for operational tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(21), 3–403-403–406. doi:10.1177/154193120004402107

Wickens, C. D. (2017). Mental workload: Assessment, prediction and consequences. In L. Longo & M. C. Leva (Eds.), *Human mental workload: Models and applications: First international symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28- 30,2017, revised selected papers* (pp. 18–29). Cham, Switzerland: Springer International Publishing.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, *37*(3), 473–494. doi:10.1518/001872095779049408

Wickens, C. D., Merwin, D. H., & Lin, E. L. (1994). Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh. *Human Factors*, *36*(1), 44–61. doi:10.1177/001872089403600103

Wiebe, E. N., Lamb, A., Hardy, M., & Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the user engagement scale. *Computers in Human Behavior*, *32*, 123–132. doi:10.1016/j.chb.2013.12.001

Willett, W., Jenny, B., Isenberg, T., & Dragicevic, P. (2015). *Lightweight relief shearing for enhanced terrain perception on interactive maps*. Paper presented at the 33rd ACM conference on human factors in computing systems (CHI 2015), Seoul, South Korea.

Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index (English translation). *Chinese Journal of Industrial Hygiene and Occupational Diseases*, *23*(3), 178–181.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482. doi:10.1002/cne.920180503

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. doi:10.1080/00140139.2014.956151

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194. doi:10.1080/14639220210123789