Dylan D. Schmorrow
Cali M. Fidopiastis (Eds.)

LNAI 10916

Augmented Cognition

Users and Contexts

12th International Conference, AC 2018
Held as Part of HCI International 2018
Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II

2 Part II

Springer

Lecture Notes in Artificial Intelligence 10916

Subseries of Lecture Notes in Computer Science

More information about this series at http://www.springer.com/series/1244

Dylan D. Schmorrow · Cali M. Fidopiastis (Eds.)

Augmented Cognition

Users and Contexts

12th International Conference, AC 2018
Held as Part of HCI International 2018
Las Vegas, NV, USA, July 15–20, 2018
Proceedings, Part II

Springer

Editors
Dylan D. Schmorrow
Office of Naval Research
Orlando, FL
USA

Cali M. Fidopiastis
Design Interactive, Inc.
Orlando, FL
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-91466-4 ISBN 978-3-319-91467-1 (eBook)
https://doi.org/10.1007/978-3-319-91467-1

Library of Congress Control Number: 2018944376

LNCS Sublibrary: SL7 – Artificial Intelligence

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

The 20th International Conference on Human-Computer Interaction, HCI International 2018, was held in Las Vegas, NV, USA, during July 15–20, 2018. The event incorporated the 14 conferences/thematic areas listed on the following page.

A total of 4,373 individuals from academia, research institutes, industry, and governmental agencies from 76 countries submitted contributions, and 1,170 papers and 195 posters have been included in the proceedings. These contributions address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The contributions thoroughly cover the entire field of human-computer interaction, addressing major advances in knowledge and effective use of computers in a variety of application areas. The volumes constituting the full set of the conference proceedings are listed in the following pages.

I would like to thank the program board chairs and the members of the program boards of all thematic areas and affiliated conferences for their contribution to the highest scientific quality and the overall success of the HCI International 2018 conference.

This conference would not have been possible without the continuous and unwavering support and advice of the founder, Conference General Chair Emeritus and Conference Scientific Advisor Prof. Gavriel Salvendy. For his outstanding efforts, I would like to express my appreciation to the communications chair and editor of *HCI International News*, Dr. Abbas Moallem.

July 2018 Constantine Stephanidis

HCI International 2018 Thematic Areas and Affiliated Conferences

Thematic areas:

- Human-Computer Interaction (HCI 2018)
- Human Interface and the Management of Information (HIMI 2018)

Affiliated conferences:

- 15th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2018)
- 12th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2018)
- 10th International Conference on Virtual, Augmented, and Mixed Reality (VAMR 2018)
- 10th International Conference on Cross-Cultural Design (CCD 2018)
- 10th International Conference on Social Computing and Social Media (SCSM 2018)
- 12th International Conference on Augmented Cognition (AC 2018)
- 9th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management (DHM 2018)
- 7th International Conference on Design, User Experience, and Usability (DUXU 2018)
- 6th International Conference on Distributed, Ambient, and Pervasive Interactions (DAPI 2018)
- 5th International Conference on HCI in Business, Government, and Organizations (HCIBGO)
- 5th International Conference on Learning and Collaboration Technologies (LCT 2018)
- 4th International Conference on Human Aspects of IT for the Aged Population (ITAP 2018)

Conference Proceedings Volumes Full List

1. LNCS 10901, Human-Computer Interaction: Theories, Methods, and Human Issues (Part I), edited by Masaaki Kurosu
2. LNCS 10902, Human-Computer Interaction: Interaction in Context (Part II), edited by Masaaki Kurosu
3. LNCS 10903, Human-Computer Interaction: Interaction Technologies (Part III), edited by Masaaki Kurosu
4. LNCS 10904, Human Interface and the Management of Information: Interaction, Visualization, and Analytics (Part I), edited by Sakae Yamamoto and Hirohiko Mori
5. LNCS 10905, Human Interface and the Management of Information: Information in Applications and Services (Part II), edited by Sakae Yamamoto and Hirohiko Mori
6. LNAI 10906, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris
7. LNCS 10907, Universal Access in Human-Computer Interaction: Methods, Technologies, and Users (Part I), edited by Margherita Antona and Constantine Stephanidis
8. LNCS 10908, Universal Access in Human-Computer Interaction: Virtual, Augmented, and Intelligent Environments (Part II), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 10909, Virtual, Augmented and Mixed Reality: Interaction, Navigation, Visualization, Embodiment, and Simulation (Part I), edited by Jessie Y. C. Chen and Gino Fragomeni
10. LNCS 10910, Virtual, Augmented and Mixed Reality: Applications in Health, Cultural Heritage, and Industry (Part II), edited by Jessie Y. C. Chen and Gino Fragomeni
11. LNCS 10911, Cross-Cultural Design: Methods, Tools, and Users (Part I), edited by Pei-Luen Patrick Rau
12. LNCS 10912, Cross-Cultural Design: Applications in Cultural Heritage, Creativity, and Social Development (Part II), edited by Pei-Luen Patrick Rau
13. LNCS 10913, Social Computing and Social Media: User Experience and Behavior (Part I), edited by Gabriele Meiselwitz
14. LNCS 10914, Social Computing and Social Media: Technologies and Analytics (Part II), edited by Gabriele Meiselwitz
15. LNAI 10915, Augmented Cognition: Intelligent Technologies (Part I), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
16. LNAI 10916, Augmented Cognition: Users and Contexts (Part II), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
17. LNCS 10917, Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management, edited by Vincent G. Duffy
18. LNCS 10918, Design, User Experience, and Usability: Theory and Practice (Part I), edited by Aaron Marcus and Wentao Wang

19. LNCS 10919, Design, User Experience, and Usability: Designing Interactions (Part II), edited by Aaron Marcus and Wentao Wang
20. LNCS 10920, Design, User Experience, and Usability: Users, Contexts, and Case Studies (Part III), edited by Aaron Marcus and Wentao Wang
21. LNCS 10921, Distributed, Ambient, and Pervasive Interactions: Understanding Humans (Part I), edited by Norbert Streitz and Shin'ichi Konomi
22. LNCS 10922, Distributed, Ambient, and Pervasive Interactions: Technologies and Contexts (Part II), edited by Norbert Streitz and Shin'ichi Konomi
23. LNCS 10923, HCI in Business, Government, and Organizations, edited by Fiona Fui-Hoon Nah and Bo Sophia Xiao
24. LNCS 10924, Learning and Collaboration Technologies: Design, Development and Technological Innovation (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
25. LNCS 10925, Learning and Collaboration Technologies: Learning and Teaching (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
26. LNCS 10926, Human Aspects of IT for the Aged Population: Acceptance, Communication, and Participation (Part I), edited by Jia Zhou and Gavriel Salvendy
27. LNCS 10927, Human Aspects of IT for the Aged Population: Applications in Health, Assistance, and Entertainment (Part II), edited by Jia Zhou and Gavriel Salvendy
28. CCIS 850, HCI International 2018 Posters Extended Abstracts (Part I), edited by Constantine Stephanidis
29. CCIS 851, HCI International 2018 Posters Extended Abstracts (Part II), edited by Constantine Stephanidis
30. CCIS 852, HCI International 2018 Posters Extended Abstracts (Part III), edited by Constantine Stephanidis

http://2018.hci.international/proceedings

12th International Conference on Augmented Cognition

Program Board Chair(s): **Dylan D. Schmorrow and Cali M. Fidopiastis, USA**

- Micah Clark, USA
- Martha Crosby, USA
- Dan Dolgin, USA
- Sven Fuchs, Germany
- Rodolphe Gentili, USA
- Scott Grigsby, USA
- Monte Hancock, USA
- Frank Hannigan, USA
- Robert Hubal, USA
- Øyvind Jøsok, Norway
- Ion Juvina, USA
- Benjamin Knott, USA
- Benjamin J. Knox, Norway
- Julie Marble, USA
- Chang S. Nam, USA
- Banu Onaral, USA
- Robinson Pino, USA
- Mannes Poel, The Netherlands
- Lauren Reinerman-Jones, USA
- Stefan Sütterlin, Norway
- Robert Sottilare, USA
- Ayoung Suh, Hong Kong, SAR China
- Christian Wagner, Hong Kong, SAR China
- Melissa Walwanis, USA
- Quan Wang, USA
- Martin Westhoven, Germany

The full list with the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences is available online at:

http://www.hci.international/board-members-2018.php

HCI International 2019

The 21st International Conference on Human-Computer Interaction, HCI International 2019, will be held jointly with the affiliated conferences in Orlando, FL, USA, at Walt Disney World Swan and Dolphin Resort, July 26–31, 2019. It will cover a broad spectrum of themes related to Human-Computer Interaction, including theoretical issues, methods, tools, processes, and case studies in HCI design, as well as novel interaction techniques, interfaces, and applications. The proceedings will be published by Springer. More information will be available on the conference website: http://2019.hci.international/.

General Chair
Prof. Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
E-mail: general_chair@hcii2019.org

http://2019.hci.international/

Contents – Part II

Cognitive Modeling, Perception, Emotion and Interaction

Augmented Learning and Training

Shared Cognition, Team Performance and Decision-Making

Contents – Part I

Context Aware Adaptation Strategies in Augmented Cognition

Brain Sensors and Measures for Operational Environments

Artificial Intelligence and Machine Learning in Augmented Cognition

Augmented Cognition in Virtual and Mixed Reality

Cognitive Modeling, Perception, Emotion and Interaction

Multi-modal Interruptions on Primary Task Performance

Pooja P. Bovard[1(✉)], Kelly A. Sprehn[1], Meredith G. Cunha[1], Jaemin Chun[2], SeungJun Kim[3], Jana L. Schwartz[1], Sara K. Garver[1], and Anind K. Dey[4]

[1] Draper, Cambridge, MA, USA
ppatnaik@draper.com
[2] UX Innovation Lab, Samsung Research, Seoul, Korea
[3] Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju, Korea
[4] Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. In this paper we have investigated a range of multi-modal displays (visual, auditory, haptic) to understand the effects of interruptions across various modalities on response times. Understanding these effects is particularly relevant in complex tasks that require perceptual attention, where pertinent information needs to be delivered to a user, e.g., driving. Multi-modal signal presentation, based on the Multiple Resource Theory framework, is a potential solution. To explore this solution, we conducted a study in which participants perceived and responded to a secondary task while conducting a visual, auditory, and haptic vigilance task during a driving scenario. We analyzed response times, errors, misses, and subjective responses and our results indicated that haptic interruptions of a primarily haptic task can be responded to the fastest, and visual interruptions are not the preferred modality in a driving scenario. With the results of this study, we can define logic for a context-based framework to better determine how to deliver incoming information in a driving scenario.

Keywords: Augmented reality · Interruptibility · Multi-modal signaling

1 Introduction

Each time a digital device, such as a smart phone or GPS, proactively provides information, it is competing for the user's attention and possibly interrupting ongoing tasks. Interruptions occur when the user is forced to shift attention away from the primary task. However, interruptions can be detrimental to accomplishing a primary task. Interruptions could increase the time required to accomplish the primary task, cause more errors, and elicit increased feelings of stress and anxiety (Adamcyzk and Bailey 2004). In addition, several characteristics of interruptions have been shown to be disruptive, including how closely the interrupting and primary tasks are related (Cutrell et al. 2001) and how much one has control over the interruption engagement (McFarlane 2002). The results of this work can help inform a context-aware framework that can more appropriately provide information to users proactively, particularly

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 3–14, 2018.
https://doi.org/10.1007/978-3-319-91467-1_1

focusing on the modalities of the task that the user is engaged in, and the modalities of the interruption.

2 Background

2.1 Context-Aware Computing

Context-aware computing might be a way to mitigate the effects of interruptions that decrease task performance. According to Wickens, a multiple resource framework can be used to assess a situation, internal and external to the operator, to evaluate the potential for interference in multi-task scenarios and even multi-modal scenarios (Wickens 2002). Such a framework can be embedded into a context-aware system to help decide how to present the information to the operator to allow for efficient use of resources since some information presented during an interruption may not be relevant to an operator's current set of primary tasks. Instead, signals are filtered, categorized, prioritized, and subsequently acted upon. Context-aware computing might be a way to mitigate the effects of interruptions that decrease task performance.

Through this context-aware computing system, the most relevant information can be presented when appropriate, through the modality (visual, auditory, haptic) that least interferes with the primary task and across the most useful interface or presentation (Abowd et al. 1999). A driving situation is a common and relevant situation in which we can see the effects of these challenges. In order to design a system to meet the level of complexity required of a context-aware computing system, various factors must be understood such as (1) how much information (relevant and irrelevant, e.g., GPS directions to the destination vs. a text about making plans next week) can be processed while driving; (2) the speed at which new information can be processed; and (3) through what modalities or channels should information be presented in order to mitigate overload while simultaneously allowing for greater information handling. This will be further explored in the Discussion to help define the parameters of a context-aware framework based on the results of this experiment.

In terms of context-aware systems, the relevant context is equally as important as understanding the level of situational awareness the operator has developed both in that instance and over time through repeated exposure. Employing a context-aware computing system and framework to a driving scenario in particular means that interruptions can be correctly prioritized and handled by systems and operators, leading potentially to fewer accidents and better understanding of upcoming hazards (Alghamdi et al. 2012).

2.2 Multiple Resource Theory

Multiple resource theory (MRT), as defined by Wickens (1984, 2002) states that there are different pools of resources available that can be leveraged at the same time depending on the nature of the task. Issues occur when multiple tasks pull from the same pool of resources; performance can drop, time-to-completion for the tasks can extend, and less information may be processed. Further, as tasks become more difficult,

performance will start to vary depending on the types of resources required to process and prioritize between the different tasks (Wickens 1984). As a result, if one modality is being utilized heavily, then presenting a signal across a different modality may result in better task performance.

The principles behind MRT suggest that input from haptic displays will not interfere with inputs from auditory or visual displays. The haptic modality has neither been incorporated nor studied as extensively as the visual and auditory modalities. The MRT model has been investigated mainly in the visual and auditory modalities and it is unclear whether the same principles apply to the haptic modality. However, Scerra and Brill (2016) did look at a primary counting task in the tactile modality and presented participants with a secondary task in the visual, auditory, and tactile modalities. They found evidence supporting the inclusion of the modality in MRT. In addition, Grane and Bengtsson (2011) found that a haptic interface reduced the visual load needed to enable effective multitasking and agreed that the Wickens' MRT model held true. Therefore, it is likely that lower response times will occur when the interrupting modality is different from the modality of the primary task. For example, an interrupting haptic modality will have a lower response time when the primary task leverages the visual modality than when it leverages the haptic modality.

2.3 Interruptions While Operating a Vehicle

Interruptions while driving can lead to increased errors and impair the ability to safely operate a vehicle (Moray 1988). The reason is that driving is a complex activity that requires high attentional resources and interruptions can exceed the cognitive capacity of the decision maker.

As a way to reduce workload and increase the safety of the driver, there has been a recent surge of interest in automated driving. Automated driving is not necessarily the answer to decreasing errors and reaction times. Highly automated systems can initiate actions on their own but must notify drivers of those actions. Drivers must continue to monitor the environment and determine if and when certain situations call for driver takeover. This role still places emphasis on the driver processing mainly visual information. A driver's situation awareness (SA) is a critical piece in interruption management, since attentional resources may need to be devoted elsewhere. Poorly designed warnings have the potential to disturb driving and distract driving (Fagerlonn 2010; Wiese and Lee 2004). For example, advanced driver assistance systems (ADAS) send audible sounds when parameters such as a driver's speed exceeds a given threshold. Findings show that the abrupt onset of beeping startles drivers, causing them to take their foot off the accelerator and momentarily deviate from the correct trajectory within a lane (Biondi et al. 2014). As a result, there is a need to understand the best modality and time to present a notification so that a driver is not overloaded with too much information.

2.4 Hypotheses

To increase our understanding, we designed a study that specifically explores modality and response time. By understanding the effects of interruptions on response times, we

can define logic for a supporting framework to better allocate incoming information and decide when and how to interrupt a user.

H1: Our first hypothesis is based on Wickens' MRT where we believe that reaction time will be reduced for the haptic modality when the interrupting modality is visual or auditory.
H2: Although the primary tasks were spread across modalities, we hypothesize that visual interruptions would perform worse than the auditory and haptic modalities since the driving scenario was more visually taxing.

3 Methods

3.1 Design

Our experiment was a 3 × 3 within-subjects design. The within-subjects factors were the Primary Task (PT) with three modes (Visual, Auditory, Haptic) and Secondary Task (ST) with three modes (Visual, Auditory, Haptic). A power analysis indicated that a sample of 30 participants was sufficient to detect large effects on outcome measures with a probability of at least 0.80.

3.2 Participants

We recruited thirty-one participants (17 male, 14 female) from Craigslist. The average age of the participants was 27.8 years with a range from 19 to 42 years. The participants had an average of 67 months (5 years, 7 months) of driving experience, which ranged from 2 to 168 months.

3.3 Protocol

We chose a driving scenario for our experimental design. Driving is a dual-task situation since drivers have to concentrate on high levels of cognitive functions (route to destination) while paying attention to immediate concerns (avoiding pedestrians). In addition, driving is a real-time task, and will provide a realistic understanding of the time it takes to react to interrupting stimuli.

We designed the experiment to present stimuli that a driver would typically encounter while driving. During the experiment, participants watched a first-person video of someone driving, while they were presented with background stimuli that required low level attention: visually, participants were presented with a driving scene; in the background, participants heard music playing softly and street noise; and the haptic stimuli consisted of gentle, constant vibrations beneath the seat pan of a participant's chair.

The stimuli for the PTs were designed to engage the participants. The primary visual task (PT-V) was a vigilance task where participants had to detect numeric information from road signs. The primary auditory task (PT-A) presented numeric information through spoken GPS guidance (e.g., "Turn left at Main street and drive

point five miles."). The primary haptic task (PT-H) consisted of receiving 3 or 5 bursts of vibro-tactile stimuli over 3 s on the left wrist and providing feedback about what participants felt.

The STs were issued during the presentation of the PTs. The secondary task visual (ST-V) stimuli were green numbers that turned red and then back to green after one second in the lower left corner of the computer screen. The secondary auditory task (ST-A) was a 1-second beep sound. The secondary haptic task (ST-H) presented a 1-second vibro-tactile stimulus on the right wrist. For every PT and ST presentation, participants were instructed to press pre-assigned keys to indicate what they detected.

PT-A was different from the other PTs because the numeric information was presented after the ST was presented. Although this was originally a concern, there were no significant differences in response times (RTs) between PT-A and PT-V/H (Fig. 1).

Fig. 1. In this example screenshot, a green number is in the periphery (called out by the arrow here) while participants drove through the scenario (Color figure online).

3.4 Materials

For both training and main experiment videos, participants "drove" through similar city settings. Ambient visual and auditory signals emitted from the video clip, and haptic vibrations were provided using a vibration actuator attached beneath the seat pan to simulate the vibration of a vehicle. Headphones were used for the audio task and covered the entire ear but did not occlude noise unrelated to the experiment. Vibration actuators were attached to both wrists.

Participants were instructed to press pre-assigned keys on a computer keyboard for each of the PTs and STs for each modality (6 keys total). The single experimental video presented all combinations of visual, haptic and audio for the primary and secondary

tasks. After the experiment, participants were asked to rate the difficulty on a 5-point Likert scale (1: very easy, 3: normal, 5: very hard) for each of the STs for a given PTs (STs × PTs) and provide reasons.

3.5 Procedures

Participants signed a consent form when they first arrived for the study. They received experimental instructions for the driving simulator and experimental task. Once participants understood the instructions, they underwent a training session where one-third of the PTs were interrupted by STs with one stimuli combination for each modality (visual, auditory, and haptic) was presented (9 PT × ST combinations). Upon completion of the training, participants were given the opportunity to ask questions before starting the main experiment. In the main experiment, participants were presented with 12 primary tasks in each modality for a total of 36 total PTs. All participants were given a questionnaire after completing the main experiment.

3.6 Measures

We collected data on response time to the stimuli (i.e., elapsed time between presentation and when the participant presses the correct key), errors, misses, and subjective rating during the study. Errors measured whether the participant incorrectly matched the assigned keys to stimuli. Misses were recorded when the subject did not provide an answer to an ST. Although the participant was not given a set amount of time to answer, stimuli were presented quickly and the participant's attention could have been directed towards pressing a key for the next stimulus presentation. A qualitative subjective rating was collected during the questionnaire at the end, which rated the difficulty of ST when PT was presented.

4 Results

Three Repeated Measures Anovas (RMANOVAs) were performed. A 3 (PT-V, PT-H, PT-A) × 3 (ST-V, ST-H, ST-A) RMANOVA assessed differences in the response times (RTs) of the secondary task for each primary task that was interrupted and primary task response times without interruptions. Significance for tests involving a repeating factor used Huynh-Feldt corrections for degrees of freedom with an alpha level of .05 and a p value less than .05. The PTs that were interrupted were significant, $F(1.71, 13.66) = 12.2$, partial $\eta^2 = .604$. The PTs that were not interrupted were significant as well, $F(1.50, 43.58) = 52.13$, partial $\eta^2 = .20$. The group means are plotted in Fig. 2 and means and standard deviations are presented in Table 1. Response times for the primary tasks are affected by the secondary task interruption. However, results of the primary task without interruption suggest that the primary tasks have varying response times depending on modality.

The PT × ST interaction was significant for secondary task $F(3.83, 114.86) = 8.18$, partial $\eta^2 = .214$. Response times for the secondary task were

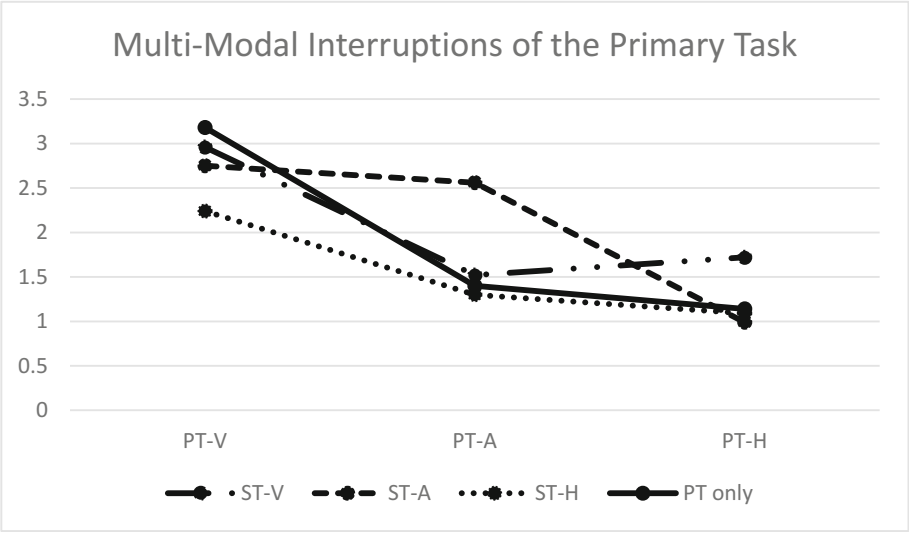

Fig. 2. Response time means for PT × ST combinations and PT only

Table 1. Stimuli ordered from fastest average response time to slowest average response time excluding the misses along with subjective ratings from the questionnaire.

Primary task	Response time means (SD) for primary task only	Secondary task	Response time means (SD) for secondary task only	Subjective ratings
PT-H	.99 (.444)	ST-A	1.44 (.563)	1.40
PT-H	1.09 (.676)	ST-H	1.15 (.409)	2.26
PT-H	1.14 (.560)	–	–	–
PT-A	1.30 (.947)	ST-H	1.67 (.719)	1.74
PT-A	1.40 (.643)	–	–	–
PT-A	1.52 (1.18)	ST-V	1.31 (.536)	3.57
PT-H	1.72 (.894)	ST-V	1.57 (.638)	2.98
PT-V	2.24 (1.35)	ST-H	1.65 (.845)	1.40
PT-A	2.56 (5.81)	ST-A	1.59 (.597)	1.86
PT-V	2.75 (1.01)	ST-A	1.19 (.375)	1.52
PT-V	2.96 (2.19)	ST-V	1.53 (.428)	3.79
PT-V	3.18 (1.42)	–	–	–

significantly different depending on which primary task was interrupted, indicating that secondary task response time is dependent upon the primary task modality.

The results show that primary tasks are affected by the secondary tasks. Subjective ratings indicate that participants found it difficult to detect ST-V compared to the other modalities (1 = very easy, 5 = very hard).

The number of misses and errors for PT × ST combinations and PT only are presented in Table 2. In general, participants missed more of the visual tasks than the other two modalities.

Table 2. Number of total misses and response errors across participants (N = 31).

Secondary task	Primary task								
	PT-V			PT-A			PT-H		
	Miss		Errors	Miss		Errors	Miss		Errors
	Pri	Sec		Pri	Sec		Pri	Sec	
ST-V	4	13	–	9	19	–	5	10	–
ST-A	8	–	–	8	–	1	1	–	1
ST-H	2	–	1	2	–	–	2	–	–

5 Discussion

Context-aware computing systems present an opportunity for leveraging multiple complex factors during complex tasks like driving. Driving is an example of a large class of use cases that involves cognitive functioning, situational awareness, and immediate concerns. Understanding situational awareness and interruptibility across and within modalities is a step forward to understanding how to leverage task and user information to improve user performance.

5.1 Hypotheses and Results

We did not find support for our first hypothesis that response times would be lower for the haptic modality when the interrupting modality was different from the primary task. However, visual and auditory modes did align with the MRT principles. Both were faster for the incongruent stimuli, rather than the congruent stimuli. We did not find that MRT applies to haptic signals in this experiment, since response time was the fastest for the PT-H ST-A combination but then PT-H ST-H was the second fastest. This was surprising since Scerra and Brill (2012) found that participants performed significantly worse in tactile-tactile dual task conditions.

The ST-H combinations were the fastest for PT-A and PT-V. One explanation for these results could come from Van Erp and Van Veen (2004). They found that drivers may benefit from haptic information, however haptic vibrations primarily provided on/off information since more complex information is difficult to convey through haptic signals. The same could be true in our study since participants did not have to remember how many buzzes they felt, merely that haptic signals were present. The finding is important because presenting interrupting haptic signals should be further investigated with respect to Wickens' MRT. We cannot incorporate multi-modal displays containing haptics without understanding the impact of haptics on visual and auditory primary tasks.

The ST-V combinations were the slowest with participants reporting that they had more difficulty detecting ST-V compared to all other PTs and STs except for PT-A. This supports the hypothesis that visual interruptions would produce slower response times than auditory and haptic modalities since driving is primarily a visual modality. Saccadic suppression could help explain that the participant's main focus was on the ongoing PT-V task, temporarily rendering them blind to other changes in the visual field (Peterson and Dugas 1972; Bridgeman et al. 1975; Burr and Ross 1982). In addition, when looking at the primary task only, the visual task had the slowest response time compared to the other two tasks. As a result, attentional resources may not have been available to notice a one second change occurring on the edge of the screen while focusing on driving.

5.2 Interruptions and Design Implications

Research has shown that distracted drivers experience "inattention blindness" where their field of view narrows (Maples et al. 2008), and they tend to look at, but not necessarily register the information in their driving environment (Strayer 2007), resulting in missing visual cues that are important for safe driving (Jacobson and Goston 2010). Inattention blindness could help explain why the ST-V was largely missed across modalities and was responded to the slowest. However, there is the possibility that since the focus was on looking for a numerical road sign, goal-directed attention and attentional priority could have been directed to a certain area on the screen, far enough away from the secondary visual stimulus to create a delay in processing (Egeth and Yantis 1997).

This finding supports research that says that even reading a text while driving is detrimental for driving (Drews et al. 2009; Hoffman et al. 2005). Based on these findings and results from the present study, we believe that autonomous cars should not warn drivers of complex decisions in a visual format. Highly automated driving allows the driver to take over at any time but especially in emergency situations; therefore drivers still need to pay attention to their environment. The findings from our study can be applied to determining how, when, and which modality information should be presented, depending on the situation, importance of the information, driver state, etc.

5.3 Multiple Resource Theory

A recent study found that presenting redundant, multi-modal signals to drivers had a positive influence on response time, with little added frustration or other negative effects (Biondi et al. 2017). In fact, they found that multi-modal presentation (auditory and haptic) at the same time resulted in faster brake and response times for drivers than in using auditory or haptic warnings individually. The more a context-aware system is able to adapt to different requirements based on driver expertise and experience, in addition to the physical and time constraints within different environments, the better it will be able to support future drivers. For example, results from Table 1 indicate that participants found the haptic modality easy to detect, however it is more difficult to use it to convey rich information than in the visual and auditory modalities.

5.4 A Context-Aware Framework

In the context of the growing challenge of information overload, we propose a theoretical framework which describes how context-aware computing technology can be strategically combined with multi-modal displays in order to provide users with the information they need, when they need it, and in a way in which they can utilize it to make decisions. In the case of a driver, the context framework would work to ensure that information from auxiliary digital devices, such as a smartphone or GPS device, is presented at the appropriate time. Results from our study indicate that there is an interaction between primary task modality and secondary (interruptive) task modality with respect to reaction time. As a result, a supporting context-aware framework should account for the modes in which a user is currently engaged (PT) and the proposed modes through which the system is considering interrupting the user (ST) when determining the timing of the interruption or the presentation of additional information.

The details of an interruptibility algorithm are a topic that merits further investigation; however, the current results do lead to some working hypotheses. Because participants in many cases failed to acknowledge ST-Vs across all three primary task modalities, interrupting a user with a visual cue should be a low probability event. Algorithmically, the effectiveness of a visual interruption should have a very low weight compared to interruptions in other modalities. As a result, the calculated cost to the user of an interruption in the visual modality should be high. This means that if the interruption is urgent, the suitability of an auditory or haptic modality should be considered. Additionally, if the information is only well-suited to the visual modality, the cost of delaying the information presentation until a time when it is not interrupting an existing PT may be lower than when compared to the cost of interrupting that task. The parameters of these cost and delay variables are a topic of future study.

Figure 3 is an example of the response time pairings informing the first-order rules for a context-aware model. This model will start to inform a framework which will come together from individual framework pieces. These guidelines serve as groundwork for developing the framework to better characterize and quantify the costs and benefits of signal combinations.

5.5 Limitations

This study was able to explore interruptions across and between modalities. While major factors contributing to the effect of the interruption on the response time were controlled for to the best of our ability, a few limitations provide opportunities for further exploration. A driving simulation has several challenges including limited physical, perceptual, and behavioral fidelity (Evans 2004), which limits high levels of experimental control. Exploring response time in a higher-fidelity driving simulation, or in a real-world driving task may alter the effect of the signals since we may find slightly different results when the scenario is more realistic.

Another limitation may have been the placement of the green number on the bottom left of the screen. Perhaps a number that flashed in the middle of the screen would have been more salient than numbers in the lower left corner. This could increase the performance of ST-V but distract from PT-V.

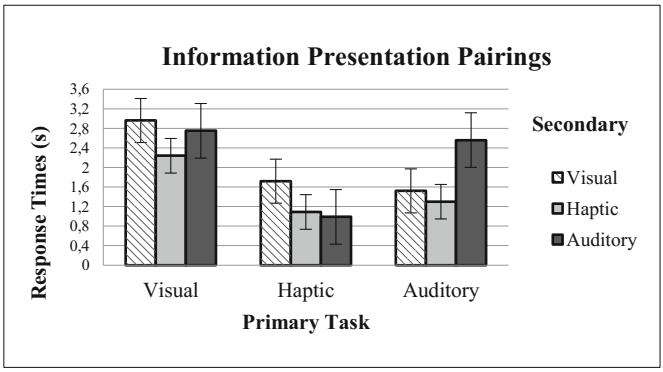

Fig. 3. Average response times for given interruptions (secondary tasks) across primary tasks. Lower response times indicate better pairings for information presentation.

Finally, this study was limited to a single interruption, categorized as a secondary task. We did not explore the effects of the importance of information, which may shift a secondary task to a primary task. The experiment could also be extended to multiple interrupting signals to assess their combined impact on response time.

5.6 Summary

The present study investigated the effects of visual, auditory, and haptic interruptions during a driving scenario. Haptic interruptions need to be further studied with regards to Multiple Resource Theory. Participants responded to the ST-V interruptions the slowest for PT-H and PT-V which suggests that interruptions during driving should not be presented visually since driving is mostly a visual task. These results inform components of a larger context-aware computing system for the purpose of distributing oncoming signals across modalities during the performance of complex tasks, such as driving.

References

Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: Gellersen, H.W. (ed.) Handheld and Ubiquitous Computing. LNCS, pp. 304–307. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48157-5_29

Adamczyk, P.D., Bailey, B.P.: If not now, when? The effects of interruption at different moments within task execution. In: Proceedings of CHI 2004 Human Factors in Computing Systems, New York, pp. 271–278. Association for Computing Machinery Press (2004)

Alghamdi, W., Shakshuki, E., Sheltami, T.R.: Context-aware driver assistance system. Procedia Comput. Sci. **10**, 785–794 (2012)

Biondi, F., Rossi, R., Gastaldi, M., Mulatti, C.: Beeping ADAS: reflexive effect on drivers' behavior. Transp. Res. Part F Traffic Psychol. Behav. **25**, 27–33 (2014)

Biondi, F., Strayer, D.L., Rossi, R., Gastaldi, M., Mulatti, C.: Advanced driver assistance systems: using multimodal redundant warnings to enhance road safety. Appl. Ergon. **58**, 238–244 (2017)

Bridgeman, G., Hendry, D., Stark, L.: Failure to detect displacement of visual world during saccadic eye movements. Vis. Res. **15**, 719–722 (1975)

Burr, D.C., Ross, J.: Contrast sensitivity at high velocities. Vis. Res. **22**(4), 479–484 (1982)

Cooper, R., Franks, B.: Interruptibility as a constraint on hybrid systems. Minds Mach. **3**(1), 73–96 (1993)

Cutrell, E., Czerwinski, M., Horvitz, E.: Notification, disruption, and memory: effects of messaging interruptions on memory and performance. In: Interact 2001: IFIP Conference on Human-Computer Interaction, pp. 263–269. IFIP, Tokyo (2001)

Drews, F.A., Yazdani, H., Godfrey, C.N., Cooper, J.M., Strayer, D.L.: Text messaging during simulated driving. Hum. Factors: J. Hum. Factors Ergon. Soc. **51**, 762–770 (2009)

Egeth, H.E., Yantis, S.: Visual attention: control, representation, and time course. Ann. Rev. Psychol. **48**(1), 269–297 (1997)

Evans, L.: Traffic Safety. Science Serving Society, Bloomfield Hills (2004)

Fagerlonn, J.: Distracting effects of auditory warnings on experienced drivers. In: 16th International Conference on Auditory Display (ICAD-2010), pp. 127–132 (2010)

Grane, C., Bengtsson, P.: Haptic addition to a visual menu selection interface controlled by an in-vehicle rotary device. Adv. Hum.-Comput. Interact. **2012**, 1–12 (2011)

Hoffman, J., Lee, J., McGehee, D., Macias, M., Gellatly, A.: Visual sampling of in-vehicle text messages: effects of number of lines, page presentation, and message control. Transp. Res. Rec. J Transp. Res. Board (1937), 22–30 (2005)

Jacobson, P.D., Gostin, L.O.: Reduced distracted driving: regulation and education to avert traffic injuries and fatalities. J. Am. Med. Assoc. **303**, 1419–1420 (2010)

Maples, W.C., DeRosier, W., Hoenes, R., Bendure, R., Moore, S.: The effects of cell phone use on peripheral vision. Optom.-J. Am. Optom. Assoc. **79**(1), 36–42 (2008)

McFarlane, D.C.: Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. Appl. Cogn. Psychol. **18**, 533–547 (2002)

Moray, N.: Mental workload since 1979. In: Oborne, D.J. (ed.) International Review of Ergonomics, vol. 2, pp. 123–150. Taylor & Francis, London (1988)

Petersen, H.E., Dugas, D.J.: The relative importance of contrast and motion in visual detection. J. Hum. Factors Ergon. **14**(3), 207–216 (1972)

Scerra, V.E., Brill, J.C.: Effect of task modality on dual-task performance, response time, and ratings of operator workload. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56, no 1, pp. 1456–1460. Sage Publications, Los Angeles, September 2012

Strayer, D.L.: Presentation at Cell Phones and Driver Distraction. Traffic Safety Coalition, Washington, DC (2007)

Van Erp, J.B., Van Veen, H.A.: Vibrotactile in-vehicle navigation system. Transp. Res. Part F Traffic Psychol. Behav. **7**(4), 247–256 (2004)

Wickens, C.D.: Processing resources in attention. In: Parasuraman, R., Davies, D.R. (eds.) Varieties of Attention, pp. 63–98. Academic Press Inc., London (1984)

Wickens, C.D.: Multiple resources and performance prediction. Theor. Issues Ergon. Sci. **3**(2), 159–177 (2002)

Wiese, E.E., Lee, J.D.: Auditory alerts for in-vehicle information systems: the effects of temporal conflict and sound parameters on driver attitudes and performance. Ergonomics **47**(9), 965–986 (2004)

Can University Students Use Basic Breathing Activities to Regulate Physiological Responses Caused by Computer Use? A Pilot Study

Hubert K. Brumback(✉)

Department of Educational Psychology and Hawai'i Interdisciplinary Neurobehavioral and Technology Laboratory (HINT Lab), University of Hawai'i at Mānoa, 1776 University Avenue, Honolulu, HI 96822, USA
brumback@hawaii.edu

Abstract. Advancements integrating computers and mobile devices with physiological sensors may increase individual opportunities for stress awareness. Conversely, computers and mobile devices, remain sources of stress, particularly in student populations. Use of these devices can have a negative effect on breathing patterns, which can, in turn, cause or exacerbate stress. This study explores whether breath counting and abdominal breathing activities may be useful to help manage stress induced by computer use. Nine male senior computer science students completed a repeated-measures experimental sequence which included Stroop color-word tasks followed by periods of sitting quietly or periods of conducting one of three breathing activities. The study highlights the possibility that college students may be able to use five-minute breathing activities to effect physiological responses. The study advocates the continued examination of these breathing activity effects on skin conductance (SC), heart rate (HR) and breath ratio. It also suggests that there may be similarity in overall effect on respiration rate (RR) for the three breathing activities used in this study. Future studies with larger sample sizes and demographic diversity may provide better fidelity on suggested relationships. This study expands upon a previously published work in progress [1].

Keywords: Breath counting · Abdominal breathing · Stress
Stroop color-word task · Students · Physiological response · Meditation
Kruskal-Wallis H test

1 Introduction

Portable and wearable sensor technology is becoming ubiquitous and enables one to potentially gain greater visibility of one's own physiological measures [2–10]. Integration of physiological sensors with computers, mobile devices and other platforms will continue to expand the possibilities for augmented cognition related to stress monitoring, especially through the use of sensors to monitor skin conductance (SC), heart rate (HR) and respiratory rate (RR) [5, 7, 10–18]. As this integration continues, individual users will have enhanced capacity, to monitor, record, and become more

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 15–31, 2018.
https://doi.org/10.1007/978-3-319-91467-1_2

aware of their own physiological signals and use this information to make decisions about their health, especially stress management.

2 Theoretical Background

2.1 Students, Stress and Technology

College and university students generally experience high levels of stress [19–23]. Some of this stress may be caused, or intensified by using computers and electronic devices, which in some cases can cause superficial or irregular breathing [24–26]. Computers and mobile devices could eventually be used to help remedy this phenomenon, particularly if used to build awareness of breathing and variations in physiological signals, and to enable individuals to make behavioral changes, which is consistent with approaches used in breath-related biofeedback for stress management [27–34].

2.2 Breathing and Stress Management

As noted previously [1], science has thoroughly documented the complex physiological relationship between breathing and stress [35–38], which is consistent with long-standing knowledge applied in a variety of meditative and other traditions [39–42]. Since breathing is both reflexive and voluntary, it has the potential to both cause and help manage stress. Breathing activities that have been used with children and students include breath counting [43, 44] and abdominal breathing [19, 32, 45–47].

One aspect of using breathing activities for stress management involves enabling individuals do develop an awareness of their own breathing. Breath awareness is used extensively in a variety of meditative practices, including meditative practices applied for stress management [43, 48–52].

2.3 Physiological Measures

From an augmented cognition perspective, the following physiological measures may be particularly useful for monitoring stress levels because sensors that measure these signals are becoming more integrated with computers and mobile devices.

Skin Conductance (SC). One measure of electrodermal activity (EDA), SC is the change in electrical conduciveness of the surface of the skin caused by eccrine sweat glands secretions. It is measured in microsiemens (μS) and increases in SC are generally related to increased physiological arousal [38, 53, 54].

Heart Rate (HR). HR is derived from electrocardiogram (ECG) measures by converting the heart period (inter-beat interval - IBI) into beats per minute to quantify cardiac activity. (It can also be determined by counting heart beats.) HR measures are converted to average beats per minute (BPM) and increases in HR can indicate physiological arousal [38, 53, 54].

Respiratory Rate (RR). RR is the quantification of breathing by counting respiratory cycles. One cycle is comprised of one inhalation and one exhalation [38, 54]. It is measured in average breaths per minute (BrPM[1]).

Breath Ratio. Breath ratio is a calculated value derived by dividing the standard deviation of the voltage obtained from an abdominal belt sensor by the standard deviation of the voltage obtained from a thoracic belt sensor. Values less than one indicate a thoracic breathing style and values greater than one indicate an abdominal breathing style [55].

2.4 Research Questions

Based on the aforementioned information, the following research questions have emerged as the basis of this study:

RQ1: Can college students use a five-minute breathing activity to regulate physiological responses caused by computer use?

RQ2: Are there significant differences in effect on physiological responses between quiet sitting and the breathing activities of breath counting, abdominal breathing and combined breath counting and abdominal breathing?

2.5 Hypotheses

The research questions are parsed into hypotheses related to specific breathing activities and physiological measures, which will guide data collection and analysis. Since the computer-based task is designed to elicit physiological response, any change in physiological response would be demonstrated by comparing the activity periods immediately following each computer-based task.

H1: Participants can use a five-minute breath counting activity to regulate physiological responses caused by computer use.
- H1.1: The breath counting activity will cause a change in participant SC when compared to the quiet sitting activity.
- H1.2: The breath counting activity will cause a change in participant HR when compared to the quiet sitting activity.
- H1.3: The breath counting activity will cause a change in participant RR when compared to the quiet sitting activity.
- H1.4: The breath counting activity will cause a change in participant breath ratio when compared to the quiet sitting activity.

H2: Participants can use a five-minute abdominal breathing activity to regulate physiological responses caused by computer use.
- H2.1: The abdominal breathing activity will cause a change in participant SC when compared to the quiet sitting activity.

[1] The acronym BPM appears in medical and scientific literature to refer to both breaths per minute for RR and beats per minute for HR. This study uses the acronym BrPM to clearly distinguish breaths per minute from beats per minute (BPM).

H2.2: The abdominal breathing activity will cause a change in participant HR when compared to the quiet sitting activity.
H2.3: The abdominal breathing activity will cause a change in participant RR when compared to the quiet sitting activity.
H2.4: The abdominal breathing activity will cause a change in participant breath ratio when compared to the quiet sitting activity.

H3: Participants can use a five-minute combined breathing activity (breath counting and abdominal breathing) to regulate physiological responses caused by a computer use.
H3.1: The combined breathing activity will cause a change in participant SC when compared to the quiet sitting activity.
H3.2: The combined breathing activity will cause a change in participant HR when compared to the quiet sitting activity.
H3.3: The combined breathing activity will cause a change in participant RR when compared to the quiet sitting activity.
H3.4: The combined breathing activity will cause a change in participant breath ratio when compared to the quiet sitting activity.

Evidence for or against these hypotheses involves a statistically significant change in the mean physiological signal measure between the periods of quiet sitting and the breathing activity periods. If there is a change, it may indicate that the participant is able to use the breathing activity to regulate physiological response, which supports the hypotheses. If there is no change it may indicate that the participant is not able to use the breathing activity to regulate physiological response, or that the breathing activity had no effect - both of which do not support the hypotheses.

2.6 Stroop Effect

The Stroop Effect [56] describes interference that occurs when two inconsistent stimuli are simultaneously presented to an individual and the individual's response to the designated stimulus is slower (in milliseconds) than when two stimuli are consistent.

Stroop color-word tasks present color words in font colors that match or do not match the color word. Participants are instructed to respond by identifying the font color, not the color-word. Stroop color-word tasks have been employed extensively to study stress and cognitive load [33, 34, 57–64]. As previously noted [1], Stroop tasks are effective laboratory stressors because practice effects only emerge after prolonged exposure [65] and because individual response to interference cannot be controlled [64].

3 Method

The purpose of this study is to investigate the efficacy of university student use of breath counting, abdominal breathing and combined breathing treatments to regulate physiological responses caused by a computer-based task. This study uses a repeated measures design to account for the range of participant physiological variation. Physiological responses were measured and recorded throughout each session across the experimental activity sequence.

3.1 Participants

Nine male university seniors were recruited from a summer computer science course. Participants age ranged from 20–24 years with the average age of 22 years. Participants received extra credit as compensation for their participation and all participants completed the experimental activity sequence. Each participant indicated he was in good health and had normal color vision.

3.2 Activity Sequence

Figure 1 shows the experimental activity sequence. The researcher first recorded baseline physiological measures for each participant during a two and one-half minute period of no activity (Fig. 1, item 1). Next, participants practiced the Stroop color-word task [66] for two and one-half minutes (Fig. 1, item 2). All participants then completed a series of three – two and one-half computer-based Stroop color-word tasks, each immediately followed by five minutes of quiet sitting (Fig. 1, items 3–8b).

After rehearsing a breathing activity that corresponded with their treatment group for two and one-half minutes (Fig. 1, item 9), participants subsequently completed another series of three – two and one-half minute Stroop tasks, each followed by five-minute periods of the breathing activity (Fig. 1, items 10–15b).

Stroop Color-Word Task. This study uses a modified version of an existing computer-based Stroop color-word task [66] for the specific purpose of eliciting physiological responses. Participants were instructed to complete the task as quickly and accurately as they were able. Each participant completed a total of seven iterations (including a practice period – Fig. 1, items 2, 3, 5, 7, 10, 12 and 14).

In the task, the program displays color words in a font color that matches or does not match the color word. Only four colors are used: red, green, yellow and blue. Items are randomly displayed to participant, half having matching color word and font color and half not. Participants are instructed to respond by pressing the key on the computer keyboard that corresponds with the first letter of the font color (not the color word). The message "Correct. Press SPACE to continue" appears after correct responses. The message "INCORRECT. Don't rush" is displayed for incorrect responses.

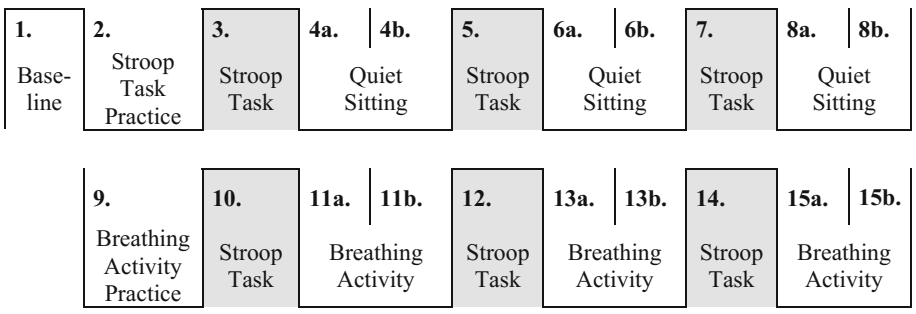

Fig. 1. Experimental activity sequence with each, two and one-half minute period delineated. Letters a and b after numbers designate the first and second halves of five-minute activities.

3.3 Treatments

The researcher used restricted random assignment, forcing equal sample sizes to place participants in one of three breathing activity conditions: (a) breath counting, (b) abdominal breathing and (c) combined - both breath counting and abdominal breathing. Each treatment group had three participants (Table 1).

Table 1. Participants by group

	Group		
	1. Breath Counting (n =3)	2. Abdominal Breathing (n =3)	3. Combined (n = 3)
Participants (N = 9)	2, 6, 7	1, 5, 8	3, 4, 9

Quiet Sitting. Participants completed a total of three, five-minute quiet sitting tasks, (Fig. 1, items 4a–b, 6a–b, and 8a–b). For quiet sitting, participants were instructed to maintain a comfortable sitting position with their backs straight and refrain from talking.

Breath Counting. The researcher provided participants in the breath counting treatment group a breath-counting worksheet with the following instructions and verbally reviewed the instructions with each participant:

1. Sit up comfortably with your back straight and both feet on the floor.
2. Breathe comfortably and focus your attention on your breathing.
3. When the time starts, mark the worksheet with the next breath event that occurs: inhale start, inhale end, exhale start, exhale end.
4. Continue to mark the worksheet with each subsequent breath event as it occurs: inhale start, inhale end, exhale start, exhale end.
5. Continue to breathe comfortably.

Abdominal Breathing. The researcher provided participants in the abdominal breathing treatment group a worksheet with the following instructions and verbally reviewed the instructions with each participant:

1. Sit up comfortably with your back straight and both feet on the floor.
2. Place your dominant hand on the center of your chest and your other hand on the center of your abdomen.
3. Inhale slowly through your nose and permit your abdomen to expand as your diaphragm descends and your lungs fill; while at the same time keeping your chest and upper body as still as you can.
4. Exhale slowly through your nose and permit your abdomen to reduce and move toward your spine as your diaphragm rises and your lungs empty; while at the same time keeping your chest and upper body as still as you can.
5. Keep your eyes open and observe your abdomen as it moves.

6. Focus your attention on the sensations of breathing and when you notice that your mind has wandered, bring your attention back to your breathing sensations.
7. Continue to breathe slowly, deeply and comfortably.

Combined. The researcher provided and verbally reviewed both instruction sets and worksheets to the participants. The participants completed the breath counting and abdominal breathing activities simultaneously with the following minor modification: participants used their dominant hand to mark the breath counting worksheet and placed the other hand on their abdomen.

Variables. The independent variable for this study is the treatment group: (a) breath counting, (b) abdominal breathing and (c) combined. The dependent variables are (a) SC, (b) HR, (c) RR and (d) breath ratio.

SC was measured by placing two disposable sensors side-by-side on the medial side of the right foot along the plantar surface of the longitudinal arch [53]. HR was measured by placing one pre-gelled disposable sensor on each of the participants forearms inferior to the antecubital space [53]. RR was measured using two belt sensors: one placed around the middle of the participant's abdomen and the other around the upper portion of the participant's thorax [55].

3.4 Procedure

Lab Description. The lab space is air conditioned and equipped with fluorescent lights and an acoustic tile drop ceiling. It contains two sections, the experimental workstation and the proctor workstation. The experimental workstation is comparable to a standard, 64-in. by 64-in. office cubicle, equipped with a small desk, computer workstation and static chair. The computer is equipped with a 20-in. monitor, standard keyboard and mouse. It is separated from the lab equipment and the proctor workstation by a 64-in. by 72-in. fabric covered partition. The proctor workstation has two displays and two sets of controls. One monitor, keyboard and mouse that mirrors and controls the participant's experimental workstation and another monitor, keyboard and mouse to view and record the physiological signals.

Instrumentation. The researcher used BIOPAC MP150 system to collect physiological data. Signals were recorded at 1000 Hz across four channels: channel one – EDA, channel two – electrocardiogram (ECG), channel three – abdominal breathing signal, channel four – thoracic breathing signal. All signals were recorded and prepared with BIOPAC AcqKnowledge software [67].

To determine HR, the researcher used the software to mark and count QRS peaks in the ECG channel. To calculate RR, the researcher used the software to mark the signals from the abdominal and thoracic RR channels at the beginning of each inspiration (inhale) and the beginning of each expiration (exhale). The researcher then used the software to count the markers and divided the count by two to derive one breath cycle (inhale, exhale) and then averaged the RR values from the abdominal and thoracic channels to establish composite RR values.

Experimental Procedure. This experiment was approved by the university institutional review board (IRB). The experimental session lasted approximately 90 min for each participant. After providing informed consent, participants were seated in the experimental space at the computer workstation. The proctor then applied the sensors and instructed the participant to complete a computer-based demographic survey. Next, the participant started the activity sequence (Fig. 1). After the final breathing activity in the activity sequence, the participant completed a post-activity questionnaire, after which the proctor removed the sensors and debriefed the participant.

4 Results

Due to the small sample size (N = 9) and subsequent small group sizes (n = 3) the researcher applied the Kruskal-Wallis H test [68] (non-parametric ANOVA) in SPSS [69] to compare means of the physiological signals between treatment groups and between selected activities in the experimental protocol.

The researcher compared the physiological data from the following selected activities: (a) baseline, (b) Stroop task before breathing activity exposure, (c) first half of the quiet sitting activity, (d) second half of the quiet sitting activity, (e) Stroop task after breathing activity exposure, (f) first half of the breathing activity and (g) second half of the breathing activity.

4.1 Between Groups

The Kruskal-Wallis H test suggests statistically significant difference between the three groups for SC, HR, and breath ratio. For SC, $\chi^2(2) = 16.19$, $p = 0.000$ with a mean rank SC score of 4.00 for Group 1, 17.29 for Group 2 and 11.71 for Group 3. For HR, $\chi^2(2) = 13.781$, $p = 0.001$ with a mean HR score of 6.43 for Group 1, 18.00 for Group 2 and 8.57 for Group 3. For breath ratio, $\chi^2(2) = 9.824$, $p = 0.010$ with a mean breath ratio score of 5.43 for Group 1, 12.29 for Group 2 and 15.29 for Group 3. The test did not show a statistically significant difference between the three groups for respiration rate (RR). Table 2 displays these results.

Table 2. Analysis of relationships between groups and physiological measures

Measure	$\chi^2(2)$	p	Mean Rank		
			Group 1 (Counting)	Group 2 (Abdominal)	Group 3 (Combined)
Skin Conductance	16.186	0.000**	4.00	17.29	11.71
Heart Rate	13.781	0.001**	6.43	18.00	8.57
Respiration Rate	2.545	0.280	11.00	8.36	13.64
Breath Ratio	9.284	0.010*	5.43	12.29	15.29

Note: N = 21 (three groups by seven activities), * $p < 0.05$ ** $p \leq 0.001$

Figure 2 displays box plots for the three physiological measures with statistically significant differences by group. Table 3 contains the corresponding box plot values for each measure by group. Breath ratio values correspond with a breathing style category. Values greater than one indicate a predominately abdominal breathing style. Values less than one indicate a predominately thoracic breathing style.

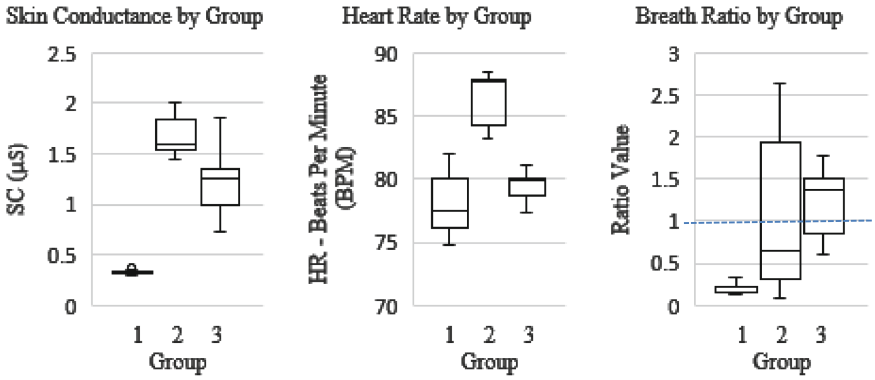

Fig. 2. Plots of statistically significant differences in physiological measures between treatment groups: Group 1 – Breath Counting, Group 2 – Abdominal Breathing, Group 3 – Combined.

Table 3. Box plot values for each measure by group

Measure	Group	Smallest Value	First Quartile	Median	Third Quartile	Largest Value	Outliers
SC (µS)	1	0.30	0.31	0.31	0.33	0.35	0.37
	2	1.45	1.55	1.59	1.84	2.00	
	3	0.74	0.99	1.25	1.34	1.86	
HR (BPM)	1	74.80	76.11	77.42	80.02	82.00	
	2	83.24	84.24	87.64	87.93	88.53	
	3	77.33	78.69	79.82	80.07	81.07	
Breath Ratio (Value)	1	0.13	0.15	0.15	0.22	0.34	
	2	0.08	0.30	0.65	1.93	2.64	
	3	0.61	0.85	1.36	1.51	1.77	

4.2 Across Activities

Graphing the mean values SC, HR, RR and breath ratio by group across the activities showed variations, but did not show any visibly consistent patterns for SC, HR and breath ratio. A consistent pattern did seem to emerge in the graph of mean RR for the three groups by activity period (Fig. 3).

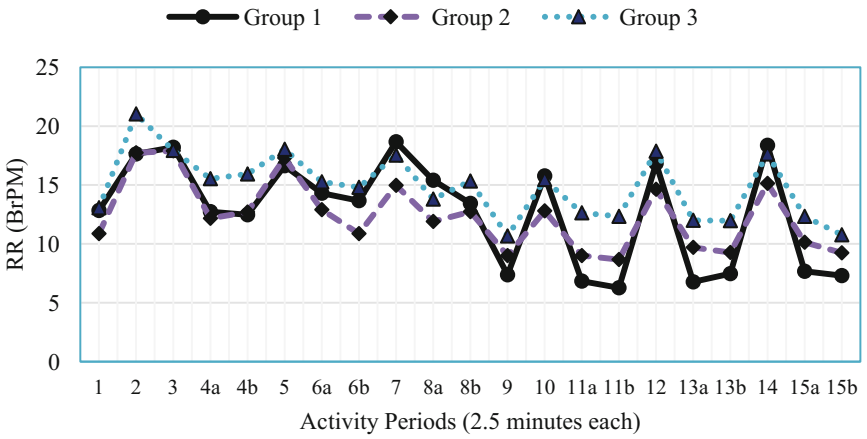

Fig. 3. Mean respiratory rate (RR) by activity

The Kruskal-Wallis H test suggests a statistically significant difference (Fig. 4) between the selected seven activities and RR, $\chi^2(6) = 16.357$, $p = 0.012$ with a mean rank RR score of (a) 8.67 for baseline, (b) 19.33 for the Stroop activities before exposure to the breathing activity (Stroop 1), (c) 12.17 for the first half of the quiet sitting activities (QSa), (d) 11.67 for the second half of the quiet sitting activities (QSb), (e) 17.00 for the Stroop activities after exposure to the breathing activity (Stroop 2), (f) 4.83 for the first half of the breathing activities (BAa) and (g) 3.33 for the second half of the breathing activities (BAb). Figure 4 shows the box plot of RR values for the seven selected activities and Table 4 contains the corresponding RR box plot values by activity. The Kruskal-Wallis H test did not show any statistically significant difference between the seven selected activities and SC, HR, and breath ratio.

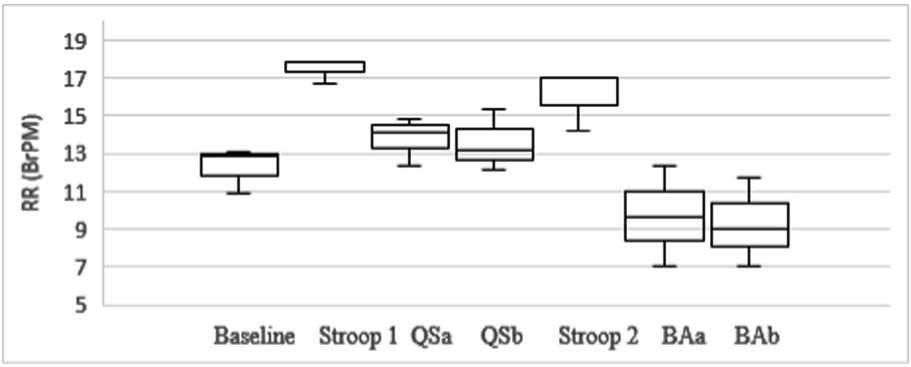

Fig. 4. Plots of RR measures across selected activities: baseline measure, Stroop 1 – pre-treatment Stroop tasks, QSa – first half quiet sitting periods, QSb – second half quiet sitting periods, Stroop 2 – post treatment Stroop tasks, BAa -first half breathing activities, BAb - second half breathing activities

Table 4. Box plot values for RR measure by activity

Activity	Smallest Value	First Quartile	Median	Third Quartile	Largest Value
1. Baseline	10.87	11.85	12.83	12.95	13.07
2. Stroop 1 (pre-treatment)	16.72	17.28	17.83	17.83	17.83
3. Quiet Sitting a (QSa)	12.32	13.23	14.14	14.51	14.87
4. Quiet Sitting b (QSb)	12.10	12.64	13.19	14.27	15.36
5. Stroop 2 (post-treatment)	14.19	15.59	16.99	16.99	16.99
6. Breathing Activity a (BAa)	7.09	8.35	9.61	10.97	12.32
7. Breathing Activity b (BAb)	7.01	8.03	9.06	10.37	11.69

5 Discussion

Small sample size precludes any definitive conclusions regarding the results of this study. The results may, however indicate potential relationships that are worth examining in future studies with more participants and with a greater range of participant demographic diversity.

The statistically significant difference between the mean physiological measures of SC, HR and breath ratio, (Fig. 2, Table 3) may lend evidence supporting the hypotheses specific to those particular measures when comparing the three treatment groups. It is also possible that the differences between the mean values was due to physical and physiological differences between individual participants, so further study with more participants has the potential to reveal evidence related to any individual differences.

5.1 Skin Conductance (SC) Between Groups

The statistically significant difference in mean SC between groups may be evidence which supports the hypotheses that the breathing activities for the treatment groups cause a change in SC (H1.1, H2.1 and H3.1). Participants in the breath counting group (group 1) showed a substantially lower ranked mean than the other two treatment groups (abdominal breathing - group 2, and combined - group 3). This could be attributed to individual differences or it is possible that the treatments were responsible for these results.

The abdominal breathing group showed the highest ranked mean score when compared to the other two groups, which could indicate that the abdominal breathing activity caused an increase in SC. Since abdominal breathing was also included in the combined breathing activity group, this could explain why the combined treatment group had a ranked mean score between the breath counting and abdominal breathing treatment groups.

5.2 Heart Rate (HR) Between Groups

The statistically significant difference in mean HR between groups may be evidence which supports the hypotheses that the breathing activities for the treatment groups cause a change in HR (H1.2, H2.2 and H3.2). The abdominal breathing group displayed the highest ranked mean score when compared with the other two groups. This could be due to individual differences or it may indicate that the abdominal breathing activity can cause a general increase in heart rate.

5.3 Breath Ratio Between Groups

The statistically significant difference in mean breath ratio values between groups may be evidence which supports the hypotheses that the breathing activities for the treatment groups cause a change in breath ratio (H1.4, H2.4 and H3.4). Because breath ratio values greater than one indicate a predominately abdominal breathing style and values less than one indicate a predominately thoracic breathing style, the fact that the abdominal breathing group displays the greatest range of ratio scores when compared with the other two groups is consistent with the experimental manipulation.

5.4 Respiratory Rate (RR) Between Groups

The absence of a statistically significant difference between the mean RR may mean that the breathing activities had no effect on RR when comparing treatment groups or that the treatments all had similar effects on RR. The possibility of similar effects for all three treatments may be supported by the similar pattern in mean values by activity (Fig. 3) and the statistically significant difference for RR across activities (Fig. 4).

Since the Stroop, quiet sitting and breathing activities all involve minimal levels of physical activity, it is possible that RR may not be a salient indicator of stress for individual participants for tasks performed while seated in front of a computer. It is also possible that individual differences and small sample size masked any statistically significant RR variation.

5.5 Respiratory Rate (RR) Across Selected Activities

The presence of a statistically significant difference in RR across the selected activities (Fig. 4, Table 4) may show a generally positive and similar effect of the breathing activities in reducing RR when comparing the baseline, Stroop, and quiet sitting composite activity periods. Both Stroop composite activity periods showed the two highest ranked values when compared to other activities, which is consistent with the used of the computer-based Stroop color-word task as a laboratory stressor. The two segments of the breathing activity period composites had the two lowest ranked values and were below baseline measures, which may show that the breath activities have a greater effect on RR when compared to the quiet sitting task. The composite baseline and the two segments of the sitting quietly composite activity periods have similar rankings falling between the Stroop activities and the breathing activities. This may

show a near equivalence between baseline measure and quiet sitting, which can be expected due to activity similarity.

6 Potential Implications

Towards gathering evidence related to the first research question, it does appear that some college students may be able to use breathing activities to mitigate changes in RR due to computer use. At this time, there is no evidence in this study to indicate that college students are able to use the breathing activities to mitigate changes in SC, HR and breath ratio measures due to computer use but these potential relationships might still be considered in future studies.

Towards gathering evidence related to the second research question, it appears that there could be significant differences between SC, HR and breath ration values for the three treatment groups. It is worth noting that the abdominal breathing activity appeared to be related to an increase of both mean SC and HR values. Typically increases in SC and HR values are generally consistent with physiological arousal [53], but it is also possible that the abdominal breathing activity was physically stimulating when compared to the overall sedentary nature of the experimental protocol, or that individual differences in physiological response could explain these differences.

This pilot study highlights potential relationships that may exist between breathing activities and participant SC, HR and breath ratio. It also may show potential relationships between RR and seven of the activities selected from the experimental activity sequence. In addition to examining these potential relationships, future studies should take into account other variables such as heart rate variability (HRV), individual stress levels and participant prior experience with breathing activities. Additionally, since physiological measures within an individual are not independent, multilevel modeling or hierarchical linear modeling statistical analysis approaches may be more appropriate for analyzing data from larger population samples.

References

1. Brumback, H.K.: Investigation of breath counting, abdominal breathing and physiological responses in relation to cognitive load. In: Schmorrow, Dylan D., Fidopiastis, Cali M. (eds.) AC 2017. LNCS (LNAI), vol. 10284, pp. 275–286. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_22
2. Abbott, D., Shirali, Y., Haws, J.K., Lack, C.W.: Biobehavioral assessment of the anxiety disorders: current progress and future directions. World J. Psychiatry. **7**, 133–147 (2017). https://doi.org/10.5498/wjp.v7.i3.133
3. Treskes, R.W., van der Velde, E.T., Barendse, R., Bruining, N.: Mobile health in cardiology: a review of currently available medical apps and equipment for remote monitoring. Expert Rev. Med. Devices **13**, 823–830 (2016). https://doi.org/10.1080/17434440.2016.1218277
4. Torrado, J.C., Gomez, J., Montoro, G.: Emotional self-regulation of individuals with autism spectrum disorders: Smartwatches for monitoring and interaction. Sensors **17**, 1–29 (2017). https://doi.org/10.3390/s17061359

5. Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., Shah, S.: AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In: Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, pp. 274–287. ACM, New York (2011). https://doi.org/10.1145/2070942.2070970
6. Kennedy, A.P., Epstein, D.H., Jobes, M.L., Agage, D., Tyburski, M., Phillips, K.A., Ali, A. A., Bari, R., Hossain, S.M., Hovsepian, K., Rahman, M.M., Ertin, E., Kumar, S., Preston, K. L.: Continuous in-the-field measurement of heart rate: Correlates of drug use, craving, stress, and mood in polydrug users. Drug Alcohol Depend. **151**, 159–166 (2015). https://doi.org/10.1016/j.drugalcdep.2015.03.024
7. Picard, R.W.: Automating the recognition of stress and emotion: from lab to real-world impact. Multimed. IEEE. **23**, 3–7 (2016). https://doi.org/10.1109/MMUL.2016.38
8. Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Roning, J., Forlizzi, J.F., Dey, A.K.: Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In: 2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), pp. 39–48. IEEE (2014). https://doi.org/10.1109/CCMB.2014.7020692
9. Bin, M.S., Khalifa, O.O., Saeed, R.A.: Real-time personalized stress detection from physiological signals. In: 2015 International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE), pp. 352–356 (2015). https://doi.org/10.1109/ICCNEEE.2015.7381390
10. Ming-Zher Poh, N.C., Swenson, R.W., Picard, R.W.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Trans. Biomed. Eng. **57**, 1243–1252 (2010). https://doi.org/10.1109/TBME.2009.2038487
11. Gravenhorst, F., Muaremi, A., Tröster, G., Arnrich, B., Gruenerbl, A.: Towards a mobile galvanic skin response measurement system for mentally disordered patients. In: Proceedings of the 8th International Conference on Body Area Networks, pp. 432–435. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels (2013). https://doi.org/10.4108/icst.bodynets.2013.253684
12. Kappeler-Setz, C., Gravenhorst, F., Schumm, J., Arnrich, B., Tröster, G.: Towards long term monitoring of electrodermal activity in daily life. Pers. Ubiquit. Comput. **17**, 261–271 (2011). https://doi.org/10.1007/s00779-011-0463-4
13. Al-Khalidi, F.Q., Saatchi, R., Burke, D., Elphick, H., Tan, S.: Respiration rate monitoring methods: a review. Pediatr. Pulmonol. **46**, 523–529 (2011). https://doi.org/10.1002/ppul.21416
14. Hernandez, J., Morris, R.R., Picard, R.W.: Call center stress recognition with person-specific models. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6974, pp. 125–134. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24600-5_16
15. Muaremi, A., Arnrich, B., Tröster, G.: Towards measuring stress with smartphones and wearable devices during workday and sleep. BioNanoScience **3**, 172–183 (2013). https://doi.org/10.1007/s12668-013-0089-2
16. Picard, R.W., Healey, J.: Affective wearables. Pers. Technol. **1**, 231–240 (1997). https://doi.org/10.1007/BF01682026
17. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans. Intell. Transp. Syst. **6**, 156–166 (2005). https://doi.org/10.1109/TITS.2005.848368
18. Zhai, J., Barreto, A.: Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In: Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1355–1358 (2006). https://doi.org/10.1109/IEMBS.2006.259421

19. Paul, G., Elam, B., Verhulst, S.: A longitudinal study of students' perceptions of using deep breathing meditation to reduce testing stresses. Teach. Learn. Med. **19**, 287–292 (2007). https://doi.org/10.1080/10401330701366754
20. American College of Health Association: American college health association-national college health assessment II: Reference group executive summary. American College of Health Association, Hanover (2017). http://www.acha-ncha.org/docs/NCHA-II_SPRING_2017_REFERENCE_GROUP_EXECUTIVE_SUMMARY.pdf
21. Feldt, R.C.: Development of a brief measure of college stress: the college student stress scale. Psychol. Rep. **102**, 855–860 (2008). https://doi.org/10.2466/pr0.102.3.855-860
22. Stallman, H.M., Hurst, C.P.: The university stress scale: measuring domains and extent of stress in university students. Aust. Psychol. **51**, 128–134 (2016). https://doi.org/10.1111/ap.12127
23. Bamber, M.D., Schneider, J.K.: Mindfulness-based meditation to decrease stress and anxiety in college students: a narrative synthesis of the research. Educ. Res. Rev. **18**, 1–32 (2016). https://doi.org/10.1016/j.edurev.2015.12.004
24. Lin, I.-M., Peper, E.: Psychophysiological patterns during cell phone text messaging: a preliminary study. Appl. Psychophysiol. Biofeedback. **34**, 53–57 (2009). https://doi.org/10.1007/s10484-009-9078-1
25. Peper, E., Harvey, R., Tylova, H.: Stress protocol for assessing computer-related disorders. Biofeedback **34**, 57–62 (2006)
26. Rosen, L., Carrier, L.M., Miller, A., Rokkum, J., Ruiz, A.: Sleeping with technology: Cognitive, affective, and technology usage predictors of sleep problems among college students. Sleep Health **2**, 49–56 (2016). https://doi.org/10.1016/j.sleh.2015.11.003
27. Edwards, L.: Combining biofeedback and mindfulness in education. Biofeedback **44**, 126–129 (2016). https://doi.org/10.5298/1081-5937-44.3.01
28. Fehring, R.J.: Effects of biofeedback-aided relaxation on the psychological stress symptoms of college students. Nurs. Res. **32**, 362–366 (1983)
29. Briz-Ponce, L., Juanes-Méndez, J.A., García-Peñalvo, F.J. (eds.): Handbook of research on mobile devices and applications in higher education settings. IGI Global, Hershey (2016)
30. Moss, D.: The house is crashing down on me: integrating mindfulness, breath training, and heart rate variability biofeedback for an anxiety disorder in a 71-year-old caregiver. Biofeedback **44**, 160–167 (2016). https://doi.org/10.5298/1081-5937-44.3.02
31. Wang, S.-Z., Li, S., Xu, X.-Y., Lin, G.-P., Shao, L., Zhao, Y., Wang, T.H.: Effect of slow abdominal breathing combined with biofeedback on blood pressure and heart rate variability in prehypertension. J. Altern. Complement. Med. **16**, 1039–1045 (2010). https://doi.org/10.1089/acm.2009.0577
32. Peper, E., Miceli, B., Harvey, R.: Educational model for self-healing: eliminating a chronic migraine with electromyography, autogenic training, posture, and mindfulness. Biofeedback **44**, 130–137 (2016). https://doi.org/10.5298/1081-5937-44.3.03
33. Prinsloo, G.E., Derman, W.E., Lambert, M.I., Rauch, H.G.L.: The effect of a single session of short duration biofeedback-induced deep breathing on measures of heart rate variability during laboratory-induced cognitive stress: a pilot study. Appl. Psychophysiol. Biofeedback **38**, 81–90 (2013). https://doi.org/10.1007/s10484-013-9210-0
34. Prinsloo, G.E., Rauch, H.G.L., Lambert, M.I., Muench, F., Noakes, T.D., Derman, W.E.: The effect of short duration heart rate variability (HRV) biofeedback on cognitive performance during laboratory induced cognitive stress. Appl. Cogn. Psychol. **25**, 792–801 (2011). https://doi.org/10.1002/acp.1750
35. Fried, R.: The Psychology and Physiology of Breathing: In Behavioral Medicine, Clinical Psychology, and Psychiatry. Springer, New York (1993). https://doi.org/10.1007/978-1-4899-1239-8

36. Timmons, B.H., Ley, R. (eds.): Behavioral and Psychological Approaches to Breathing Disorders. Plenum Press, New York (1994)
37. Chaitow, L., Gilbert, C., Bradley, D. (eds.): Recognizing and Treating Breathing Disorders: A Multidisciplinary Approach. Churchill Livingstone Elsevier (2014)
38. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.): Handbook of Psychophysiology. Cambridge University Press, Cambridge (2017)
39. Bhikkhu, Ā. (ed.): Mahāsatipaṭṭhānasuttaṁ (DN 22): The long discourse about the ways of attending to mindfulness (2011). http://www.ancient-buddhist-texts.net/Texts-and-Translations/Satipatthana/Satipatthana.pdf
40. Ali-Shah, O.: The Rules or Secrets of the Naqshbandi Order. Tractus Books, Reno (1998)
41. Eifring, H. (ed.): Meditation in Judaism, Christianity and Islam: Cultural Histories. Bloomsbury Academic, New York (2013)
42. Therā, N.M.: Satipaṭṭhāna sutta [MN 10: The Discourse on the Establishing of Mindfulness]. In: The Buddha and His Teachings. Buddhadhamma Foundation (1980). http://www.bps.lk/olib/bp/bp102s.pdf
43. Britton, W.B., Lepp, N.E., Niles, H.F., Rocha, T., Fisher, N., Gold, J.: A randomized controlled pilot trial of classroom-based mindfulness meditation compared to an active control condition in 6th grade children. J. Sch. Psychol. **52**, 263–278 (2014). https://doi.org/10.1016/j.jsp.2014.03.002
44. Hooker, K.E., Fodor, I.E.: Teaching mindfulness to children. Gestalt Rev. **12**, 75–91 (2008)
45. Sellakumar, G.K.: Effect of slow-deep breathing exercise to reduce anxiety among adolescent school students in a selected higher secondary school in Coimbatore. India. J. Psychol. Educ. Res. **23**, 54–72 (2015)
46. Terai, K., Shimo, T., Umezawa, A.: Slow diaphragmatic breathing as a relaxation skill for elementary school children: a psychophysiological assessment. Int. J. Psychophysiol. 229 (2014). https://doi.org/10.1016/j.ijpsycho.2014.08.897
47. Fonseca, D., Montero, J.A., Guenaga, M., Mentxaka, I.: Data analysis of coaching and advising in undergraduate students. an analytic approach. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2017. LNCS, vol. 10296, pp. 269–280. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58515-4_21
48. Dvořáková, K., Kishida, M., Li, J., Elavsky, S., Broderick, P.C., Agrusti, M.R., Greenberg, M.T.: Promoting healthy transition to college through mindfulness training with first-year college students: pilot randomized controlled trial. J. Am. Coll. Health **65**, 259–267 (2017). https://doi.org/10.1080/07448481.2017.1278605
49. Lichtenstein, B.: The mindfulness imperative: how the pedagogical principles of mindfulness provide the foundation for biofeedback. Biofeedback **44**, 121–125 (2016). https://doi.org/10.5298/1081-5937-44.3.07
50. Kabat-Zinn, J.: Mindfulness-based stress reduction (MBSR). Constr. Hum. Sci. **8**, 73–107 (2003)
51. Gold, E., Smith, A., Hopper, I., Herne, D., Tansey, G., Hulland, C.: Mindfulness-based stress reduction (MBSR) for primary school teachers. J. Child Fam. Stud. **19**, 184–189 (2010)
52. Goldin, P.R., Gross, J.J.: Effects of mindfulness-based stress reduction (MBSR) on emotion regulation in social anxiety disorder. Emotion **10**, 83–91 (2010). https://doi.org/10.1037/a0018441
53. Potter, R.F., Bolls, P.D.: Psychophysiological Measurement and Meaning: Cognitive and Emotional Processing of Media. Routledge, New York (2012)
54. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological Recording. Oxford University Press, Oxford (2000)

55. Peper, E., Groshans, G.H., Johnston, J., Harvey, R., Shaffer, F.: Calibrating respiratory strain gauges: what the numbers mean for monitoring respiration. Biofeedback **44**, 101–105 (2016). https://doi.org/10.5298/1081-5937-44.2.06
56. Stroop, J.R.: Studies of interference in serial verbal reactions. J. Exp. Psychol. Gen. **121**, 15 (1934/1992). https://doi.org/10.1037/0096-3445.121.1.15
57. MacLeod, C.M.: Half a century of research on the Stroop effect: An integrative review. Psychol. Bull. **109**, 163–203 (1991). https://doi.org/10.1037/0033-2909.109.2.163
58. Karthikeyan, P., Murugappan, M., Yaacob, S.: Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. J. Phys. Ther. Sci. **24**, 1341–1344 (2012). https://doi.org/10.1589/jpts.24.1341
59. Karthikeyan, P., Murugappan, M., Yaacob, S.: Analysis of Stroop color word test-based human stress detection using electrocardiography and heart rate variability signals. Arab. J. Sci. Eng. **39**, 1835–1847 (2014). https://doi.org/10.1007/s13369-013-0786-8
60. Karthikeyan, P., Murugappan, M., Yaacob, S.: A review on stress inducement stimuli for assessing human stress using physiological signals. In: 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, pp. 420–425 (2011). https://doi.org/10.1109/CSPA.2011.5759914
61. Wallén, N.H., Held, C., Rehnqvist, N., Hjemdahl, P.: Effects of mental and physical stress on platelet function in patients with stable angina pectoris and healthy controls. Eur. Heart J. **18**, 807–815 (1997)
62. Crabb, E.B., Franco, R.L., Caslin, H.L., Blanks, A.M., Bowen, M.K., Acevedo, E.O.: The effect of acute physical and mental stress on soluble cellular adhesion molecule concentration. Life Sci. **157**, 91–96 (2016). https://doi.org/10.1016/j.lfs.2016.05.042
63. Gwizdka, J.: Using Stroop task to assess cognitive load. In: Proceedings of the 28th Annual European Conference on Cognitive Ergonomics, pp. 219–222. ACM, New York (2010). https://doi.org/10.1145/1962300.1962345
64. Renaud, P., Blondin, J.-P.: The stress of Stroop performance: physiological and emotional responses to color–word interference, task pacing, and pacing speed. Int. J. Psychophysiol. **27**, 87–97 (1997). https://doi.org/10.1016/S0167-8760(97)00049-4
65. Gul, A., Humphreys, G.W.: Practice and colour-word integration in Stroop interference. Psicológica Rev. Metodol. Psicol. Exp. **36**, 37–67 (2015)
66. Yang, E.Z.: Stroop effect - an xhtml 1.0 strict Javascript based interactive program. http://ezyang.com/stroop/. Accessed 6 Feb 2017
67. BIOPAC Systems Incorporated: AcqKnowledge data acquisition and analysis software. BIOPAC Systems Incorporated, United States (2014). https://www.biopac.com/product/acqknowledge-software/
68. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**, 583–621 (1952)
69. IBM SPSS Statistics. International Business Machines Corporation (2017)

Human Performance Augmentation in Context: Using Artificial Intelligence to Deal with Variability—An Example from Narrative Influence

William D. Casebeer(✉), Matthias Ziegler, Amanda E. Kraft, Jason Poleski, and Bartlett Russell

Advanced Technology Laboratories, Lockheed Martin, Arlington, VA 22203, USA
william.d.casebeer@lmco.com

Abstract. Bringing together humans and machines in a performance-improving symbiosis requires giving our digital assistants, robots and other artificial teammates the ability to better understand the states of their human colleagues. In this paper, we discuss how technology can be used to assess human reactions to information, a critical technology development both for enabling the development of influence assessment tools, and for human-machine teaming. Developing technology suites to detect and exert influence is of paramount importance in a world where kinetic and non-kinetic effects interact to produce final outcomes in the national security domain. We discuss development of a comprehensive technology suite to allow the US and its Allies to detect and disrupt radicalization processes in multiple media; the suite is distinguished by its use of human-in-the-loop cognitive testing to allow rapid retailoring of information activity, and could give military personnel entirely new capabilities to understand and influence the information environment.

Keywords: Human machine teaming · Influence · Narrative
Physiological monitoring · Information operations

1 Introduction

Developing technology suites to detect and exert influence is of paramount importance in a world where kinetic and non-kinetic effects interact to produce final outcomes in the national security domain. Here, I discuss development of a comprehensive technology suite to allow the US and its Allies to detect and disrupt radicalization processes in multiple media; the suite is distinguished by its use of human-in-the-loop cognitive testing to allow rapid retailoring of information activity, and will give military personnel entirely new capabilities to understand and influence the information environment.

Violent non-state movements such as ISIL, al Qaeda, and others leverage cultural expertise and exquisite locally-grounded historical knowledge to form narratives and tell stories which exploit innocent bystanders and cultivate permissive operating environments in which to thrive; the same goes for state actors, such as Russia. Adversary

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 32–42, 2018.
https://doi.org/10.1007/978-3-319-91467-1_3

information operations can be effective at convincing their sometimes innocent targets to look the other way—or even actively support—terrorist tactics and strategies by providing people, money, moral and materiel support, or can be used to achieve strategic objectives such as undermining cultural conditions enabling democracies to thrive.

Detecting these ideologically-driven information operations is an important capability; the United States and its allies cannot respond to what we do not sense. More important, being able to formulate a holistic strategy for undercutting the efficacy of these operations is a critical part of a counter-terrorism and counter-radicalization strategy. This will involve developing tools and technologies to formulate and forecast the effect of counter-narratives, supporting information, and larger environmental factors on the future abilities of our adversaries. This could involve leveraging existing technologies, and tools which could be built relatively quickly, to equip the US with a comprehensive "counter-radicalization toolkit" to contest adversary information influence. This suite would allow the US to detect, analyze, and understand adversary information operations, and provide "human-in-the-loop" tools to assist in developing counter-narratives to influence the behavior of the audience in ways which will prevent them from being exploited by malignant violent non-state actors. Measures of performance and effectiveness will provide feedback to allow rapid calibration of a comprehensive counter-radicalization information campaign.

The proposed system accomplishes this by automating the analysis of multiple forms of media (broadcast, social, etc.), detecting emerging themes which enable violence to take root. Narrative templates connect the automated analysis of content with facts about local circumstance to build models which forecast future population and group-level behavior in light of the information being received and the surrounding environment. These drive a campaign planning tool, which allows the US and allies to shape the political and economic environment to minimize the chances of radicalization, and to build effective counter-narratives and alternate schema which trusted voices in the local community can use to change the information environment. Uniquely, the tool suite is connected to behavioral, psychological and physiological monitoring systems which allow rapid tailoring and pilot-testing of narratives in light of the expected audience, to boost the chance they will be heard and considered. This enables the US and its allies to speak truth to the power that violent non-state movements sometimes hold over innocent populations.

Technologies are available which are relatively mature which can contribute to this process, such as the Lockheed Martin Integrated Crisis Early Warning System (ICEWS) [1] and the Lockheed Martin Human Systems and Autonomy team's Cognitively-Aided Design and Evaluation (CADE) and related cognitive engineering processes, which can be leveraged to build this comprehensive counter-radicalization suite. Some technologies used in the construction of the system are exploratory, but with modest investment could be turned into operationally useful tools which the military—ranging from strategic planners to combatant commanders, to specialists in information support operations—can use to comprehensively defeat groups such as ISIL. This could take place quickly, allowing the technologies to be refined to give the US new capability to operate in the information and narrative domain by 2020.

2 Operational Opportunity

The Final Report of the 9/11 Commission spent a fair amount of time identifying and discussing the ideology of al Qaeda, and made strong recommendations to engage in the "struggle of ideas." Since that report, successive national strategy documents on counter-terrorism (CT) have arguably weakened the linkages between CT efforts and ideology and have focused primarily on kinetic actions. Further, the 9/11 Commission's report was very explicit about the nature and the definition of the ideology behind some violent non-state actors. Given that the process of radicalization has an information component, being able to understand and act within your adversary's information observe-orient-decide-act ("OODA") loop is a requirement for a comprehensive counter-radicalization strategy. Put differently, a grand counter-terrorism strategy would benefit from a comprehensive consideration of the stories terrorists tell; understanding the narratives which influence the genesis, growth, maturation and transformation of terrorist organizations will enable us to better fashion a strategy for undermining the efficacy of those narratives so as to deter, disrupt and defeat terrorist groups. More, recent developments in near-peer information operation awareness highlight how state actors leverage narrative formation and disruption to influence internal events elsewhere, as in the case of Russian interference in the NATO member nation and US political domains.

Such a "counter-narrative strategy" will have multiple components with layered asynchronous effects; while effective counter-stories will be difficult to coordinate and will involve multiple agents of action, their formulation is a necessary part of any comprehensive counter-terrorism effort. *Indeed, a failure on our part to come to grips with the narrative dimensions of the war on terrorism is a weakness already exploited by groups such as al Qaeda and ISIL; we can fully expect any adaptive adversary to act quickly to fill story gaps and exploit weaknesses in our narrative so as to ensure continued survival.* More than giving us another tool with which to confront terrorism, though, narrative considerations also allow us to better deal more generally with the emerging security threat of violent non-state actors and armed groups.

Why think that storytelling has anything to do with terrorism and counter-terrorism? Consider the psychological aspects of terrorism: there are multiple reasons why people choose to form or join organizations which use indiscriminant violence as a tactic to achieve their political objectives, all of them dealing at some point with *human psychology*. People feel alienated from their surroundings; they are denied political opportunity by the state; the state fails to provide basic necessities; they identify with those who advocate the use of violence; they are angered by excessive state force against political opponents; their essential needs are not being met; they feel deprived relative to peer groups elsewhere; and so on. These have all been offered as "root causes" of contentious politics in general, and terrorism in particular. Our purpose here is not to defend any particular position about root causes (indeed, some of those previously listed have been discredited as theories of terrorism), but instead merely to point out that all these causes have a proximate psychological mechanism—they exert influence by affecting the human mind/brain. If stories are part and parcel of human cognition, we would also then expect consequently that stories might affect how these causes play out to germinate, grow and sustain terrorism and radicalization [2]. Operators need to be

able to detect and analyze stories in progress, forecast their effects, formulate and enact alternate stories in a human-in-the-loop fashion, and assess the behavioral impact of their counter-narrative strategy. Our adversaries do this presently owing to their closeness to the cultures in which they operate; cultivating our own capability to do so will allow us to systematically disrupt their operations and leverage the softer elements of national power to prevent the exploitation of vulnerable populations.

3 Enabling Technologies

The technologies required to build this suite include the ability to sense, analyze and understand narrative information operations in multiple media, the ability to refine models forecasting group and population behavior in light of detected narratives quickly and with sensitivity to audience variability using cognitive and physiologic measures, and the ability to assess the behavioral impact of information operations.

Developments in existing technology suites—discussed below—and recent developments in the cognitive science of narrative and storytelling, serve as the backbone for this proposed system. It builds off well-established technologies (such as ICEWS), but incorporates novel physiologic and neurobiological sensors so as to provide a unique in the world human-in-the-narrative-loop counter-radicalization information operations test bed.

3.1 Proposed System

The proposed system integrates a two-pronged approach to analyzing information operations and their impact. The first chart details some of the existing and near-future technologies required to *detect* narrative information activity (using ICEWS Trending, Recognition and Assessment of Current Events or "iTRACE," a tool allowing you to detect event patterns in multiple media types), *predict* the impact the messaging might have on sentiment and behavior (the Social Network Opinion Dynamics and Analysis or "SNODA" tool, and the ICEWS Forecasting or "iCAST tool), and *evaluate* the actual impact on sentiment and behavior (using the ICEWS sentiment analysis or "iSENT" tool) [3]. Other systems could be used as well. This capability can then be connected to course of action development and analysis via the ICEWS environment in conjunction with electroencephalogram (EEG) signals—patterns of brain-generated electrical activity sensed on the top of the head—and other cognitive variables to quickly assay the impact of a revised narrative. This allows us to improve models of audience behavior in light of the change to the message or to the environment in which it is delivered. Figure 1 captures the information-related dimensions of the proposed system. Figure 2 captures the human-in-the-loop message prototyping dimensions.

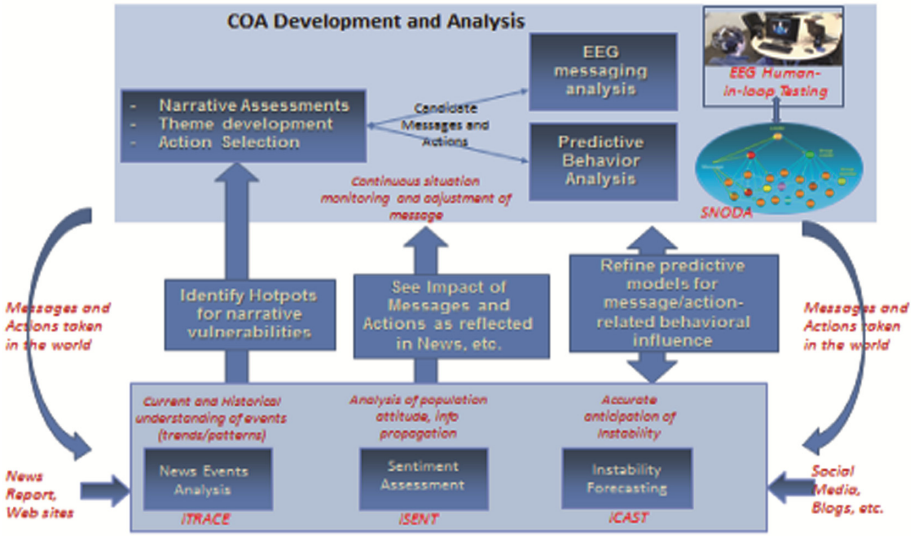

Fig. 1. The narrative information system

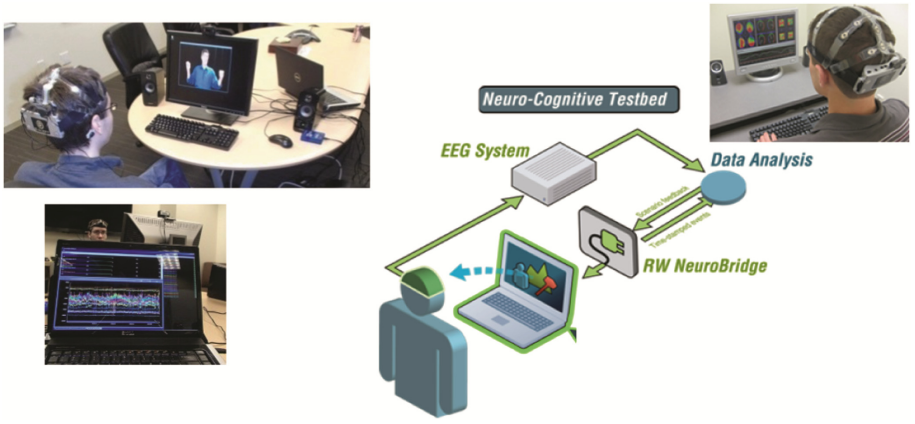

Fig. 2. Expanding the "EEG (electro-encephalogram) human-in-loop testing" block—a prototype narrative influence and message analysis test bed

3.2 System Capabilities

The system operates by combining the best computer science algorithms for parsing structured and semi-structured text from open sources to extract events and sentiment with models which forecast behavioral impact, and leverages work done by my lab and others in this area. These models are constantly improved by having representatives of the population one hopes to reach look at prototype messages in a closed-loop monitoring situation where their psychological and physiological reactions serve as proxies

for attention, engagement, arousal, empathy for characters, narrative transportation and immersion, and ultimately expected behavioral influence. Capabilities are discussed in more detail below in Sect. 3.4.

The technology suite would have the following general capabilities to:

(1) monitor and analyze multiple media types in real time,
(2) combine that analysis with other types of event data,
(3) automate extraction and analysis of narratives to allow sentiment forecasting,
(4) connect narrative analysis to social network analysis of populations and group,
(5) pilot test proposed information operations and counter-narratives with a human-in-the-loop, using the latest cognitive science and physiology,
(6) allow effective detection, analysis, forecasting, planning and execution of information operations.

3.3 Significance of Capabilities to Operational Opportunity

These capabilities enable military strategic planners, combatant commanders, military information support operations personnel, and others to understand the narrative dimensions of the information environment they will operate in and provide planning guidance necessary to allow rapid adjustment of messaging activity, improved mid-to-long-term adjustment of the environment of action via economic and political development, and an ability to understand the second and third-order effects of operations and adversary radicalizing narratives on the military operations environment (even in those rare cases when no particular information action can be taken).

In the military information support operations environment, this tool suite can provide capability that cuts across all aspects of the traditional operational cycle: planning, target audience analysis, series development, product development and design, approval, production/distribution/dissemination, and measures of effectiveness. Traditional tools related to counter-messaging can be brought to bear but in an environment which allows rapid retailoring of them to maximize their effectiveness.

3.4 Enabling Technology

Enabling technologies leveraged here include EEG devices and collection platforms used by companies such as Intific and others (such as the Human Systems and Autonomy lab), and from scientific developments stemming from work accomplished by the City College of New York (the Parra lab) [4], the University of Southern California (Damasio lab) [5], the Massachusetts Institute of Technology (Saxe lab) [6, 7], and others. This work has confirmed and extended relationships between story structure and content and detectable neural signals linked to behavior change. For example, principal components from the EEG signal correlate closely to viewer attention to a media stimulus and also predict whether the viewer will send a tweet about it [8] (Fig. 3).

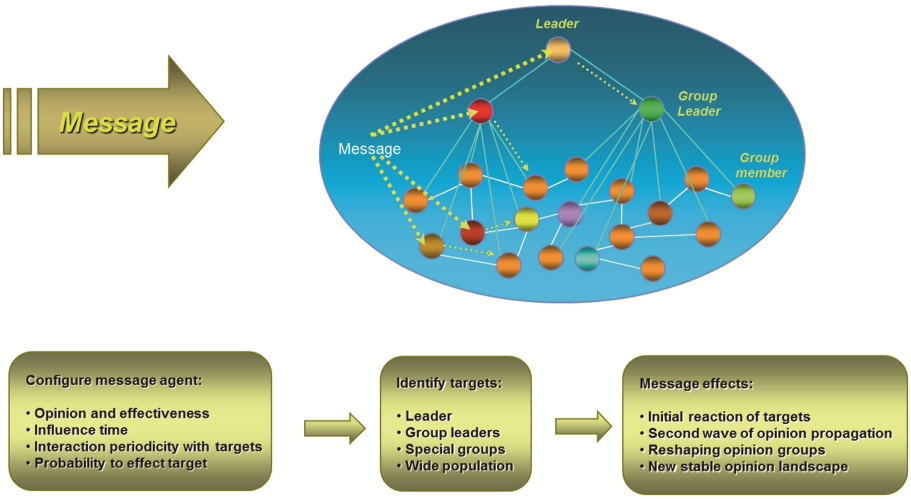

Fig. 3. Analyzing messages in social network influence context

In addition, the target audience will respond not only to messaging, but also to the actions that are taken by our military in the areas where the messaging is taking place. It will be important to make sure that the messages and actions together tell a coherent story. Understanding of how the target audiences responds to the messages and actions together can be analyzed in a "behavior-predictive agent-based model" that includes agent-based models of individuals and groups that are based on knowledge of their decision-making strategies designed and validated from inputs from the news, social media, social scientists, psychologists, and neuroscientists. The agent-based models can be combined in the LM ATL model interaction "backplane," allowing the agents to interact with models that represent their environment, such as whether they have electricity, food, access to water, etc., and the messaging models. The Social Network Opinion Dynamics & Analysis (SNODA) will analyze opinion propagation and stabilization in response to external influence campaigns or actions of the military. SNODA represents the network of influence relationships in a society and the opinions of individual members. Connections in the network encode the propensity for individual opinion shifts based on influences affecting each individual [9]. These existing models are primarily at the proof of concept level. However, as advances in technology and the sciences are used to improve the models, enabling the responses to messaging and actions to interact within a population will likely produce a more reliable result than models that produce these responses independently.

The events which feed into narrative templates and drive SNODA and iCAST predictive analyses come from iTRACE—it extracts event type, participants and intensity, locations, and times from unstructured open news sources. It provides a graphic display of events, trends and patterns with drill-down to underlying news stories. Event coding of news stories using Raytheon BBN's SERIF product as the primary event coder is one of the core technologies at the heart of the iTRACE capability. The stories come from English, Spanish and Portuguese language sources. To date, over 30 million news

stories processed and 20 million events have been extracted going back 20 years. This includes Factiva-aggregated news stories from over 6000 sources, plus Open Source Center feeds. The coder extracts events of the form of a "tuple" of (1) source actor, (2) event and (3) target actor, using established taxonomies and at a high ~80% accuracy. Once the events are coded the LM ATL geocoder, Lautenspot, is used to identify the location of the events. Most events can be correctly located to the country level, while some can be located at the province or even the city level. Each event is also assigned a hostility level between −10 and 10 based on the Goldstein scale where a 10 represents a cooperative event or cessation of hostilities and a −10 represents a very hostile (i.e. violent) event. These events and the Goldstein scores are used as source data by the SNODA and iCAST forecasting tool.

3.5 Maturity

A variety of technologies are brought together into this comprehensive suite. Depending on which piece of technology is under consideration, some capability exists that is already operationally fielded (for example, in the primary ICEWS system) [10]. Other capabilities—such as relationships between certain aspects of human physiology and likely narrative influence on behavior—are emerging findings from the basic sciences which are ripe to be incorporated into the technology suite. Pieces that are relatively immature, such as agent-based models linking narrative structure and content to expected propagation, can be matured relatively quickly.

The principal barriers to making the system usable are doctrinal and only secondarily technological. For instance, it is entirely possible to detect and analyze a story spreading in a particular form of social media, to model its likely effect on behavior, and then to propose and propagate an alternate narrative that has been stress-tested in the human-in-the-loop test bed. However, whether the results of this process can be used quickly are contingent on ensuring that operational commanders have the requisite authorities to quickly act in the information space abroad. In some cases, approval chains for the release of information can slow this process and render the technology not as effective as would otherwise be the case.

There is an industrial base here (primarily in assessing the impact of entertainment, and in informing business operations), and some of the work in the cognitive science laboratories mentioned earlier has used more familiar polling methodologies from this industry to test posited relationships between EEG monitoring and behavior. LM ATL and some of its personnel have been involved in both government and commercial settings in the development and testing of these technologies.

Many of the practitioners in this domain have military and information operations experience; for example, the author of this white paper is a former military officer with familiarity with the military planning process and who has worked with the military information support operations community in the information technology domain on previous projects, and LM ATL has experience in transitioning prototypes into operational use (as has already occurred with ICEWS).

3.6 Recommendations for Development

This system could emerge from prototype component development and integration to become fully operational with appropriate investments in (1) the narrative templates which will link sensed events to estimations of the impact of a particular narrative on a population, (2) the agent-based models which could undergird forecasting of narrative influence, and (3) continued investigation of and integration into the full system of neurobiological and physiological behavioral impact measures. The technologies will need to be tested in a controlled environment beginning with a demonstration, and then validated in an operational environment. This process will take several years, but the combined technology readiness level of the technologies—and the gaps that will need to be filled to develop an operational prototype—means that the right investment could assist in transitioning the technology from prototype to fielded system with demonstrated capability quickly.

4 Methods for Employing the Technology

The system could be fielded operationally for use in the military decision-making process, with forward-deployed components as well as reach-back to domestic piloting sites. It can support training exercises aimed at the military decision-making process, assisting staff development at training facilities such as those operated by the J-7 at Suffolk, VA, where social media analysis and operations are already tested, but not in a persistent fashion. It can be used at the strategic and operational levels by combatant commander staffs seeking quick intelligence preparation of the environment and rapid turns on the expected information effects of military operations, and by units such as Strategic Command's headquarters (charged with developing and deploying deterrence and influence frameworks). Most easily, it could quickly be integrated into all the existing processes used by groups such as the US Army's Military Information Support Operations Command at Fort Bragg, or the US Marine Corp's Information Operations Center at Quantico, who are already building and deploying information campaigns in support of US and coalition operations. The technology could also be usefully deployed to multinational coalition environments, such as the NATO Cyber Defense Centre of Excellence in Tallinn, Estonia.

The suite could also be deployed in other research environments, such as social media laboratories operated by the military at the Naval Postgraduate School, or even by national labs investigating influence and social media, such as Sandia National Laboratories. It would thus serve as a technical driver in supporting the larger whole-of-government exploration of deterrence, influence and information force projection.

Like almost all technologies, there are conversations to be had about ethical, legal and social issues. Existing legal and statutory authorities suffice for the system to be deployed in the environments just mentioned. To be used most effectively and in an agile fashion, information operation decisions will need to be pushed to the lowest levels possible, however. In general, there is a well-developed framework supporting the synchronization of traditional military operations and the information dimension (as in our core joint doctrine). Multiple analysts have already discussed the need for the US

military to continue investment in technologies which allow it to prevent violent non-state actor exploitation of the vulnerable (see, e.g., Casebeer [11]). The system does not need to be secret to be effective—the scientific findings that it relies on apply even when individuals understand that information influences their behavior. The development of the suite may even act as a deterrent to groups such as ISIL or the Russian Internet Research Agency who at present arguably think they have information dominance and can operate with impunity in the narrative sphere.

Equipping the US military and its allies with the technology required to engage and defeat ISIL and other violent non-state actors is challenging. The types of technologies discussed in the Technology Suite to Detect and Defeat Radicalization would provide us with an important tool that can be used to deter, disrupt and defeat our adversaries in the narrative and information spaces where they currently operate to radicalize individuals and cultivate permissive operating environments. It can be an important enabler for a comprehensive and effective counter-terrorism and counter-radicalization strategy and an important cultural stabilizer for democracies concerned to disrupt and deter attempts by other nation states to skew democratic deliberation and internal political events. Twenty-first century security challenges demand sophisticated and subtle approaches of the kind enabled by this technology. Its effective use in phase zero, one and two of conflict can save lives, prevent the need for costly kinetic operations, and work in synergy with the use of force when its application becomes a necessity [12].

References

1. For information on this system, see the ICEWS website at www.icews.com
2. Casebeer, W.D., Russell, J.A.: Storytelling and terrorism: towards a comprehensive 'counter-narrative strategy'. Strateg. Insights **IV**(3) (2005)
3. Malinchik, S.: Framework for modeling opinion dynamics influenced by targeted messages. In: The Second IEEE International Conference on Social Computing, Minneapolis, Minnesota, August 2010. http://www.atl.external.lmco.com/papers/1912.pdf
4. Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H., Parra, L.C.: Audience preferences are predicted by temporal reliability of neural processing. Nat. Commun. **5**, 29 (2014)
5. Araujo, H.F., Kaplan, J., Damasio, A.: Cortical midline structures and autobiographical-self processes: an activation-likelihood estimation meta-analysis. Front. Hum. Neurosci. **7**, 548 (2013)
6. Cikara, M., Bruneau, E., Van Bavel, J.J., Saxe, R.: Their pain gives us pleasure: how intergroup dynamics shape empathic failures and counter-empathic responses. J. Exp. Soc. Psychol. **55**, 110–125 (2014)
7. Bruneau, E., Dufour, N., Saxe, R.: How we know it hurts: item analysis of written narratives reveals distinct neural responses to others' physical pain and emotional suffering. PLoS ONE **8**, e63085 (2013)
8. Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H., Parra, L.C.: Audience preferences are predicted by temporal reliability of neural processing. Nat. Commun. **5**, 4567 (2014)
9. Malinchik, S., Rosenbluth, D.: Paradoxical dynamics of population opinion in response to influence of moderate leaders. In: IEEE Symposium Series on Computational Intelligence (SSCI 2011), Artificial Life, pp. 148–153, April 2011

10. O'Brien, S.P.: A multi-method approach for near real time conflict and crisis early warning. In: Subrahmanian, V.S. (ed.) Handbook of Computational Approaches to Counterterrorism. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-5311-6_18
11. Casebeer, W.D.: A neuroscience and national security normative framework for the twenty-first century. In: Giordano, J. (ed.) Neurotechnology in National Security and Defense: Practical Considerations, Neuroethical Concerns, 30 September 2014. Taylor and Francis (2014)
12. Thomas, T.S., Kiser, S.D., Casebeer, W.D.: Warlords Rising: Confronting Violent Non-State Actors. Lexington Books, New York (2005)

Human Machine Interactions: Velocity Considerations

Joseph Cottam(✉), Leslie M. Blaha, Kris Cook, and Mark Whiting

Pacific Northwest National Laboratory, Richland, WA 99352, USA
{joseph.cottam,leslie.blaha,kris.cook,mark.whiting}@pnnl.gov

Abstract. Measuring change is increasingly a computational task, but understanding change and its implications are fundamentally human challenges. Successful human/machine teams for streaming data analysis effectively balance data velocity with people's capacity to ingest, reason about, and act upon the data. Computational support is critical to aiding humans with finding what is needed when it is needed. This is particularly evident in supporting complex sensemaking, situation awareness, and decision making in streaming contexts. Herein, we conceptualize human/machine teams as interacting streams of data, generated from the interactions that are core to the human/machine team activity. These streams capture the relative velocities of the human and machine activities, which allows the machine to balance the capabilities of the two halves of the system. We review the known challenges in handling interacting streams that have been distilled in computational systems. And we use this perspective to understand some of the open challenges to designing effective human/machine systems that support the disparate velocities of humans and machines.

Keywords: Big data · Human-machine interaction
Interactive streaming analytics · Visual analytics

1 Introduction

Despite the availability of "big data," people remain effective processors of only small data. Human perceptual and cognitive capacities remain constant, including the effective limits on our working memory [1–3], limits on our attentional capacity [4,5], and often limited bandwidth information processing capacity [6–8]. People are also susceptible to a number of cognitive moderators that reduce our ability to process information quickly and without error, such as fatigue, stress, or cognitive overload [9]. We must rely on interactive interfaces to supply the functionality to balance the difference between small-data processing people and big-data processing machines.

The implications of this reliance on interactive interfaces to moderate the speeds of human and machine activity are two-fold. First, all the primary responsibility for processing and handling data falls to the machine. Machines can be

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 43–57, 2018.
https://doi.org/10.1007/978-3-319-91467-1_4

built to operate at the speed and capacity needed to handle the velocities of the data, the algorithms, and the user inputs. Machine intelligence, including complex algorithms and artificial intelligence, can be employed to determine how to execute processes and to provide solutions under dynamic system demands. Second, the design and engineering of those systems falls to humans. Human intelligence must determine the requirements of both the computing and human information processing elements, complete with user requirements, variable working environments, and changing goals.

The purpose of big data and computational resources is to be helpful to ourselves: we gather data and employ resources to solve human problems. With advancing machine technologies, the interaction between humans and machines has changed shape. Smaller, more efficient, and more powerful machines have both quantitatively and qualitatively changed what we try to compute on and the expected output of that computation. The development of *deep learning* is a recent example of this: it became computationally feasible to execute large-scale analysis (including higher speed networking, greater storage density, and improved compute power) that was previously unreachable. Quantitatively, more operations per second and larger data were economically available. Qualitatively, pervasive data collection resources provided richer data than had been available in the past. Collectively, these abilities are enabling advances in life-critical applications, such as self-driving cars and advanced health-care. Importantly, a major contributor to the success of many deep-learning efforts are the data gathered through new *interaction* patterns found in crowd-sourcing information (e.g., Mechanical Turk, Re-captcha) [10].

There are many conflicting definitions of "big data" but the majority share variants on three characteristics V's: volume, variety, and velocity [11,12]. (Other definitions add characteristics such as veracity, volatility, variability, validity, and value. It is beyond our scope to explore the necessity, correctness, or implications of each.) Each characteristic carries different constraints on human/machine teaming and system engineering. Velocity is of particular interest at present because it is a source of fragility in human/machine teams. Velocity of big data refers to the speed at which data is generated, recorded, or refreshed. Because machines are often involved in the production, data velocity is implicitly related to the speed at which machines can process some source data. Importantly, the producing machine and the analyzing machine are often distinct.

A human/machine team must also consider the velocity of the human: people's capacity to ingest, reason about, and act upon the data. Considering velocity from both perspectives, time is a shared but relatively inflexible medium. There is little we can do to change our experience with time: it does not wait or stretch or save for humans or for machines. However, humans and machines experience time and events over time on very different scales. Indeed, they are even measured in different units, with meaningful human activity measured in wall clock time in tens of milliseconds to years [13] and meaningful machine activity happening in system time down to the nanosecond. Consequently, apparent

velocity versus relative processing speed has significantly different implications for human/machine teams. Regardless of how fast things are happening or the measurement scales, "velocity" implies change over time, which means sequencing and distributions of events become meaningful in the context of time. Specifically, time and sequence restrict the analysis and force constraints on human reasoning.

Velocity, and particularly the difference in velocity between machines and humans, is an important consideration for the development of systems performing interactive streaming analytics at scale. Interactive streaming analytic systems, schematically diagrammed in Fig. 1, are human/machine systems that strive to enable human decision making on streams of data in (near) real-time. The goal is to produce correct, useful, and timely interpretations of the world, where a human's experience of the world is mitigated by the machine representations of the data streams. Note that when we refer to streams, we are referring to continuing generation and collection of data where the values are changing while being observed. Stream analytics steer data collection, summarize information, and manage events occurring in the streaming data. Such analytics may capture dynamic query updates and employ interactive interfaces to provide human-digestible information to support intelligent decision making. Such systems provide the joint human/machine intelligence needed to perform real-time systems monitoring (e.g., power grid control stations or flight control operations), continuous security screening, or mission command and control. Putting users into the streaming analytics process requires that interactive interfaces effectively match both the velocity of both human cognitive inference and interface interactions as well as the machine-derived velocity of updates to analysis and visual representations. Thus, interactive streaming analytics at scale requires an interface to mediate the difference between humans and machines to facilitate meaningful interactions.

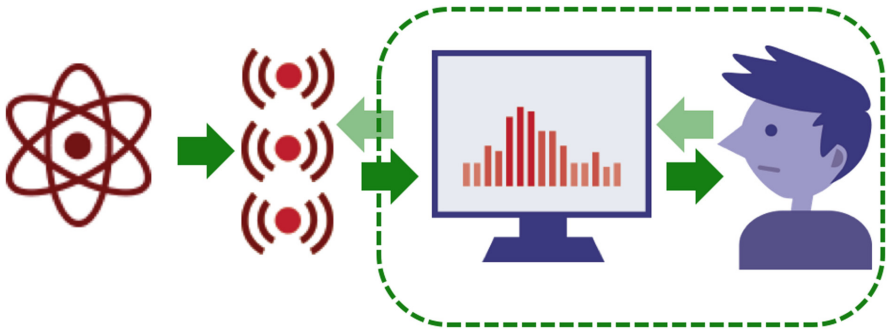

Fig. 1. Human/machine teaming for interactive streaming analytics at scale. The human/machine team is contained in the green box, with the interactive user interface mediating the differences in the velocity of the user activity (right side) and the velocity of data streaming in from the world (left side). (Color figure online)

Despite the need for interactive streaming systems in response to increasing data availability, we have rarely succeeded in developing effective solutions to this problem, let alone general solutions. One reason for a lack of general solutions is that we lack a complete understanding of the conceptual and design spaces and how they interact. This trade-off space captures the challenges of developing systems for large scale, streaming data and the related challenges and opportunities emerging from the cognitive capabilities of human operators working on streaming data. In this paper, we argue that designing and engineering for human/machine teaming in interactive streaming analytics or other dynamic operational environments is critical for finding effective solutions to balance the velocity of humans, the velocity of computers, and the velocity of the phenomena to which the human/machine must be responsive. We believe the key is emphasizing the interactions, because interaction data processing can be handed to the machine, alongside other data streams, and the humans can focus on designing effective interfaces to facilitate those interactions. From this perspective, human/machine interactions create interacting data streams, and leveraging known approaches for handling data streams can help us shape the questions about how to design effective human/machine teams to operate in dynamic, streaming conditions. Thus, we explore how reflecting on velocity shapes what designers need to consider about both the human and the machine to make an effective team.

2 Human/Machine Teams

The core of any team is interaction. Interactions provide the opportunity for team members to influence each other, and become a joint entity that is more than the sum of its parts. The information passed across a user interface (UI) is therefore crucial to understanding how a human/machine team acts. Fluid human/machine teaming is particularly critical for interactive streaming analytics systems that seek to put a human into the loop for real-time analysis [14].

Algorithmic, hardware, and cultural changes have enabled computer assistance in problem solving. Humans and machines form a team with distinct but complementary capabilities for analytical tasks. Humans are masters of context: capable of bridging gaps between disparate datasets and making heuristic judgments within datasets. In most cases, some vital information people work with cannot be easily encoded for computational reasoning. Humans are also able to act on information in ways distinct from computers, interfacing with the world outside of the analytical environment. The human side of the team brings the motivation for problem solving and is the ultimate arbiter of the goals to pursue.

Computer capabilities are often engineered to be complementary to human capabilities. Machines provide raw capacity, repeatability, and a certain flavor of efficiency. Computers handle large amounts of data at high speeds, they provide repeatable processes, and tools for verifiability and precision. In a streaming context, these machine capabilities are invaluable to human problem solving.

However, human and machine capabilities are not automatically cohesive. It takes attention to merge the two into a team. This is doubly true in streaming

context, because interactions need to focus on timely data analysis instead of learning system behaviors and functionality. It is desirable in this context to enable the system to simply observe the patterns of user interactions, and to use that data in meaningful ways to help shape the human/machine team into the cohesive team that can effectively perform the tasks of interest (e.g., [15,16]). The latter statement, however, encompasses a broad set of open challenges in how to use data in meaningful ways.

A number of advances in developing cohesive teams may be resolved by establishing better models for each teammate. If computers were better equipped with better human models, they would be able to more able to operate as a team member. Current computer-hosted models of humans are typically static and highly reactive (consider, most user interfaces essentially model likely "next" interactions or expected constraints on interactions). Improving these models could enable more proactive machine roles. Humans with better models of the machine can more effectively communicate with it and interpret the results. The interaction patterns between the humans and machines offer a key measurement of the interacting agents that can be used to establish and inform such models.

2.1 Human Input

Human inputs to a machine interface are naturally modeled as a stream. They are time varying, potentially unbounded, and not inherently repeatable [17,18]. Streaming data analysis systems that incorporate user input in a non-trivial way therefore must deal with the intersection of the user input with other data streams. These issues extend beyond general multi-stream issues discussed in Sect. 4 (though many of those issues should also be considered when dealing with human input).

Our interest is in human/machine teaming systems that have non-trivial user streams. For our purposes, "non-trivial" user inputs are those that influence how future items in other streams may be processed. In some systems, user inputs only inspect current items available in the user interface (UI). That inspection does not influence future system behaviors, unless mixed-initiative system approaches are engaged to indirectly influence future behavior through machine steering. Imposing a filter is perhaps the simplest non-trivial input: future items may or may not be displayed. This interaction is simple because it is deterministic and essentially algebraic in nature. More complex interactions include training examples through semi-supervised learning [19,20], semantic interactions [21], and interaction sequences that might indicate foraging or other exploratory analytic processes in provenance modeling. Conceptually the difference between trivial and non-trivial inputs can be roughly approximated thus: in a client-server system, a non-trivial user input is one that crosses from the client back to the server, while trivial inputs stay within the client [22].[1]

[1] This is a heuristic modeled on server-client designs. A sufficiently thin client will not be able to do simple inspection on its own, and a sufficiently capable client may do machine learning.

Practically, non-trivial user inputs may come in the form of keyboard actions, mouse clicks, touch inputs, or verbal communication, based on the design of the system. Although input details influence the exact velocity of the system, our stream-based model is agnostic to them because each will have a sequence and human-scale velocity.

2.2 Machine Input

Machines contribute multiple different kinds of input into a system. In addition to being integral in the data collection pipeline, machine actions around analytics and UI can also define a stream. Consider, for example, mixed-initiative computing systems. Tracking of user inputs is a hallmark of mixed-initiative systems for visual analytics [23]. In mixed-initiative systems, user interactions serve to steer and inform the machine analytics. Machine actions then make recommendations to the user for future interactions (an example of this recommendation workflow style is the Active Data Environment [24]). This approach creates a series of machine actions that influence the information and options presented back to the user. This is the machine interaction stream.

Semantic interactions were introduced to describe UI interactions that suggest particular interpretations to the machine because they occur in the context of a known interface metaphor [21,25]. Broadly, machines might leverage semantic interactions to understand the user's mental model and sensemaking process [26]. In many current mixed-initiative systems, the semantic interactions are interpreted as they occur through the course of analytic session; they are not necessarily revisited or reinterpreted, even if the analytic provenance is recorded. The machine reactions to semantic interactions are not usually directly analyzed, but they also reflect a component of the machine interaction stream. Conceptualizing machine interactions as a stream will capture all the key characteristics of the velocity of machine intelligence, which can be aligned with the velocity of human interactions and with the data ingestion and processing velocities.

3 (Partial) Streaming Solutions

Stream processing algorithms and systems have been an active area of research for decades. This existing work establishes a foundation for machine processing of streaming data at a variety of velocities. This section summarizes machine-processing oriented observations about streams.

3.1 Time & Sequence

A stream of values arrives over time, implying a sequence of values that arrive over time. Therefore, stream algorithms must work with changing values or a growing collection of values over time. This is distinct from traditional time-series analysis. In time-series analysis, time itself is a part of the data being analyzed. Often a full time series will be available at once, so the sequence of

values is not integral to the analysis strategy. In streaming analysis, time *may* be part of what is being analyzed; but sequence of values *must* be part of the analysis strategy (either to ensure it does not impact results or to employ it to improve results).

3.2 Sequence

A stream implies that values are presented over time. However, the presentation order does not need to correspond to their production order. Out-of-order appearance is common in many real world cases. A practical example is in the Transmission Control Protocol (TCP), which includes specific provision for out-of-order delivery of (streamed) packets. In a trivial case, production may be the iteration of values from database in reverse temporal order. As these two examples illustrate, a sequence may be irregular or regular and sequence deviations may be incidental or deliberate. Regardless of its form, origin and severity of sequence issues, if present, must be addressed.

3.3 Data Rates

Many streams have "too much" content to fully capture or practically store the entirety. In other words, the data rate may be too high to handle, either computationally or cognitively. Issues of frequency directly influence the scaling of a system computationally, and ability of humans to act on computational results. Choices must be made to determine sampling rates that are practical according to sensor limitations, that capture the meaningful events on the stream, and that enable the "right" amount of data to be analyzed and/or stored.

3.4 Cadence

Stream cadence may be highly regular or prone to bursts. Time-windowed stream volume may be variable as well. These have technical and interface implications.

On the technical side, cadence directly impacts resource allocation strategies. Regular cadence enables efficient resource management. Irregular cadences can be mitigated with buffering strategies (e.g., burst buffers on HPC machines, Cloud Flair for website requests) that smooth out the stream, trading variable volume for variable lag. Another option for handling variability is to simply "drop" stream content when it exceeds the design threshold (as is done in the ethernet networking protocol).

In the interface, stream cadence influences how updates are communicated. Whether irregular or regular, the cadence of updates itself may be part of the signal of interest. Displaying the distribution of updates can alleviate the cognitive burden of wondering about update frequency. The interface design should also consider the expected cadence range when implementing updates to minimize noise in the display.

3.5 Time Spans vs. Time Moments

The content of a stream may have drastically different *semantics*, even if the content is substantially similar. Returning to the sensor example, there are many valid configurations. A sensor may take measurements at given intervals and report each. It may take measurements at given intervals and only report those that differ from the prior. In the first case, the measurements are for moments in time, in the second reported measurements are for spans.

4 Crossing Streams

Many issues are compounded when multiple streams are brought together [22]. Stream velocities may differ in their rates, dimensionality, and variability. The contrast between different streams introduces many new characteristics, like cross-correlations or asynchrony, that must be dealt with in both analysis and interpretation.

4.1 Relative Skew

Audio and video are common streams encountered in a time-correlated fashion. YouTube, television, and films all rely on small temporal skews to maintain the illusion that the elements of the images are in fact the sources of the sounds. If the skew between the two streams becomes too large, the illusion of causality is broken. Whenever combining streams, issues of skew need to be considered. Do the streams travel similar paths, or is there an inherent lag between them? If they are produced together, do they travel together?

4.2 Multiple Resolution

Data from different sources often do not match in temporal resolution. This problem is often masked by a time-stamp convention that records all time as offset seconds (such as Unix timestamps). Regardless of the granularity of the measurement, the granularity of the time-stamp is driven by the accuracy of the available *clock*. Detecting that different resolutions are at play is the first problem.

The second problem is working with values recorded from different resolutions. For values with well-defined summarization properties, this can be straightforward operationally, but still requires users apply knowledge of context. For example, if one stream reports daily temperature and another reports hourly temperature from another location, what is the best way to mix the two? Maximum, minimum, averaging, random selection, and periodic selection can all be trivially applied to the more frequently reporting stream, but a key question is: which will provide the best comparison to the less frequent stream? Knowledge of how the slower stream is created likely dictates what is "best". Consequently, combinations are *operationally* simple while remaining *contextually* difficult.

4.3 Stream Interdependence

Streams may not be independent. Consider a UI for an industrial control application. The display shows the results of a stream of a data, capturing the current state of the system. The user inputs are a stream back to the system that influences future values. In other words, human inputs are in response to events on task streams. They can also influence what happens later on that task stream.

5 Implications for Design of Effective Human/Machine Teams

We have conceptualized human/machine teams as systems designed for non-trivial interactions at the core, and that those non-trivial interactions produce streams of data. Thus, we have a way to capture the velocity of human and machine behaviors, as well as the relative impact of those velocities on each other. All human and machine data streams can be analyzed by the machine, meeting its responsibility for handling all the data. We next consider some of the system behavior considerations that are known to influence the effectiveness of human/machine teams. In light of adopting the perspective of human/machine teams as interacting data streams, we elucidate some implications for these factors, particularly with respect to the management of differing velocities. And we identify open research questions for the human-centric system design process.

5.1 Different Working Speeds

The working speed difference between humans and machines is an obvious point of both conflict and complementary abilities. If every shift of machine attention were presented to a human to interact with, the machine would be left largely idle. Human relative slowness provides a mediating reservoir of attention. Thus, system designs can use the slower speed of humans to aggregate and synthesize machine activities and outputs before presenting to the user. Using the interacting data streams to be aware of the human attention availability and engagement, the machine can manage this timing adaptively.

5.2 Communication and Reconciliation

How do team members communicate goals, findings, and the need for status updates with each other? There are issues of interruption, non-computationally represented information, and changing priorities to be considered. In verbal communication between people, this often seems straightforward. Yet, between humans and machines, the communication barriers often seem insurmountable, despite advances in spoken-communication systems (like Siri and Alexa). Multimodal and bi-directional communication is enabled through interacting streams. As the world and streaming data evolve over time, both the human or machine may communicate a new context. The interacting stream capture the evolving mental models or state of the agents. But a key open question remains: how does the team's shared state evolve?

5.3 Workflows and Task Allocation

Division of labor between human and machine intelligent agents is informed by a number of system attributes, including the nature of the tasks and the specific capabilities of the computational resources. Interacting streams capture fluctuations in working cadence, as well as information about what the machine and user are working on through semantic and mixed-initiative interactions. This enables adaptive and dynamic task allocations. It may also enable effective multitasking in humans, who often think they are good multi-taskers but rarely are [27,28]. Capturing the relative machine and human cadences further raises opportunities to adaptively and effectively interrupt current activities. That is, the streams may inform the time points at which can each teammate might usefully provide input to the other, and in what form in which that interruption might be most influential.

5.4 Trust, Verification, and Uncertainty

A number of open questions exist around how to ensure that users trust and rely on the machine intelligence and autonomous system elements. How do you present information at the right level of detail to be appropriately considered? How do you develop an appropriate appreciation in the human of the machine's abilities, without introducing more work than the machine or human is able to assume? What are the levers and limits of interrogation that a human can employ against a machine? What does it mean for a machine to be certain? How much of a model of the human can the machine effectively employ in building the relationship? How complete a model does the human need of the machine? Designs based on interacting streams may help to address a number of these questions, and extensive research is needed to understand the relevant of design options that facilitate human and machine trust and reliance.

5.5 Perceived Autonomy

We apply computational machinery to problems humans want to solve. However, machine capabilities are expanding and with them comes an expectation of greater autonomy, meaning adaptive system behaviors to manage information and operate on data without direct human supervision. For many existing interactive systems, machine tasking for non-routine tasks is largely under the control of the human operator. If interaction streams can inform models of user intent and need, how well can we identify ways in which the machine make take the lead? How much work is *initiated* by the machine? How many resources can it apply to a task before asking for human input?

Autonomy deals with the adaptive ability to initiate lines of inquiry (autonomy in the large) and the adaptive ability to execute tasks without additional input (autonomy in the small). Both forms carry benefits and pitfalls, and have implications for the velocity of the related machine behaviors. Autonomy in the large provides the appearance of an able assistant (like Jarvis from Iron Man or

the Doctor from Star Trek's Voyager) or a potential replacement. In a more contemporary context, contextually aware advisors such as spelling/grammar check, auto-complete in web-search or the Active Data Environment [24] demonstrate a level of machine autonomy. In potentially autonomous machines, how much work "on their own" is the owner comfortable with? How much do they want to direct themselves? Is it just "spare cycles" that the machine has for its own inquiries, or can it use a baseline percentage?

In the small, machines that require constant attention can be an annoyance and impediment rather than a benefit. Indeed, poor interaction design can lead to a lack of autonomy for all team members as each blocks the other's progress. However, when context dictates the proper resolution of ambiguities, it may be better to receive inputs than to simply guess. Humor around auto-correct is derived from failures when there is too much autonomy in the small, while dialog box fatigue that leads to everything being blindly accepted and installed is a failure of too little autonomy.

5.6 Influence

Machine autonomy implies the active metaphor for instruction is not command, but influence. This influence can be inferred from the interaction of the user by having the machine interpret the interactions in the context of computational models of user intent. This is an indirect means of influence. At the opposite end of the spectrum is more direct influence, such as allocating computational resources to specific tasks or placing explicit priority markers. These are methods for a human to influence a machine's behavior. Machine influence of human behavior may be observed in the order that options are presented in and in the intrusiveness of interface elements. The salient aspect of influence is that it is indirect and that it preserves choice in when and how actions are taken.

6 Case Study: Test Harness

Considering interaction as a core consideration of human/machine for streaming data analysis leads to different considerations in system design than otherwise. Our experience designing and implementing test infrastructure for interactive streaming analytics is illustrative. Although human-in-the-loop data analytics are desirable in many settings, humans contribute variable behaviors to the team. This variability introduces novel challenges to system testing and evaluation.

In traditional software testing, algorithmic correctness and overall robustness are the principle consideration. For algorithmic correctness, systems are tested to see if inputs produce the expected outputs (or for stochastic systems, an output in an acceptable range acceptably often). A system state may be supplied artificially through mocks or stubs [29] to simulate the context developed over longer executions. Acceptable system robustness is assessed through some form of stress testing, checking for resource usage and the ability to remain operational (or degrade gracefully). Inputs in either case are typically supplied by the system

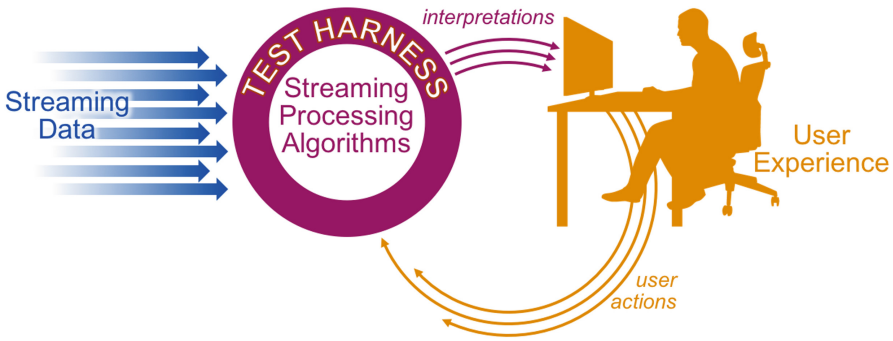

Fig. 2. System-level context for test harness.

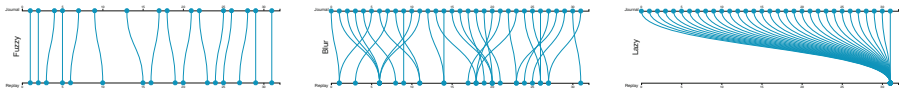

Fig. 3. Three test patterns. Fuzzy (left) tests for minor changes in timing. Blur (center) tests for changes in sequence. Lazy (right) tests if the user input needed to be provided mid-stream at all.

designer, possibly with guidance from a potential future user. Other properties that are commonly tested for include resource utilization, system responsiveness, data security or feature regression. Each of these properties is valuable in a useful system, but the mock/stub/etc. architecture often glosses over the very interactions that are central to the human/machine teaming experience.

In contrast, our experience developing a test harness for interactive human/machine teaming has led us to propose some additional properties that should be assessed. These properties include:

Timing sensitivity: How sensitive to exact interaction timing are results?
Sequence sensitivity: How sensitive to exact interaction order are results?
Input criticalities: Which interactions are essential to a result and which can be omitted/ignored?
Output inflections: Does the output vary smoothly as inputs change, or does it exhibit inflection points? Where are the inflection points?

These interaction-focused properties rest on the interactions that occur when the sequence and timing of user input are considered. In other words, they arise from treating user input as a *stream* that interacts with other input streams.

An architectural overview of our stream analytics test harness is shown in Fig. 2. The test harness wraps an analytic system, recording user inputs with a vector time stamp [30] derived from other input streams. The user inputs are the manifestation of interaction points in the combined human/machine teaming system. The recording is timed relative to the other input streams to enable greater flexibility in the testing. (More details can be found in a prior

publication [22].) With this architecture, modeling and simulation-based tests for timing and sequence can be conducted with user inputs as the seeds of the test. Figure 3 illustrates a few such tests. Each test entails structured change of recorded input. The change may affect sequence, timing, exact values, presence/absence, or a combination of these things. By presenting slightly changed inputs and comparing the results to the original results, velocity-relevant system properties can be efficient explored. This also provides a way for user inputs to be incorporated into the software development cycle in a meaningful way, without requiring constant user input or insisting on unnatural user consistency of action.

7 Conclusion

Human/machine teams are of growing importance to data analysis. Effectively forging a human/machine team requires attention to the capabilities and limitations of each. This is especially evident for streaming data analysis because the data velocity is experienced substantially differently for the human versus the machine. However, there is much work to build on to start constructing effective human/machine team interfaces. Focusing on the interactions instead of the state kept in either side of the team provides a useful perspective on this problem.

Acknowledgments. This effort was sponsored by the Analysis in Motion Initiative at the Pacific Northwest National Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Baddeley, A.: Working memory: looking back and looking forward. Nat. Rev. Neurosci. **4**(10), 829–839 (2003)
2. Cowan, N.: The magical mystery four: how is working memory capacity limited, and why? Curr. Dir. Psychol. Sci. **19**(1), 51–57 (2010)
3. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81–97 (1956)
4. Cowan, N., Elliott, E.M., Saults, J.S., Morey, C.C., Mattox, S., Hismjatullina, A., Conway, A.R.: On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. Cogn. Psychol. **51**(1), 42–100 (2005)
5. Posner, M.I., Snyder, C.R., Davidson, B.J.: Attention and the detection of signals. J. Exp. Psychol.: Gen. **109**(2), 160–174 (1980)
6. Wickens, C.D.: The structure of attentional resources. Attention Perform. VIII **8**, 239–257 (1980)
7. Blaha, L.M.: An examination of task demands on the elicited processing capacity. In: Little, D.R., Altieri, N., Fifić, M., Yang, C.T. (eds.) Systems Factorial Technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms. Academic Press, London (2017)

8. Heathcote, A., Coleman, J.R., Eidels, A., Watson, J.M., Houpt, J., Strayer, D.L.: Working memorys workload capacity. Mem. Cogn. **43**(7), 973–989 (2015)
9. Gluck, K.A., Gunzelmann, G.: Computational process modeling and cognitive stressors: background and prospects for application in cognitive engineering. In: Lee, J.D., Kirk, A. (eds.) The Oxford Handbook of Cognitive Engineering, pp. 424–432. Oxford University Press, Oxford (2013)
10. Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing Annotations for Visual Object Detection. In: AAAI Human Computation Workshop (2012)
11. Gartner Group: Pattern-based strategy: Getting value from big data (2011)
12. McAfee, A., Brynjolfsson, E., Davenport, T.H., et al.: Big data: the management revolution. Harvard Bus. Rev. **90**(10), 60–68 (2012)
13. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1994)
14. Dasgupta, A., Arendt, D.L., Franklin, L.R., Wong, P.C., Cook, K.A.: Human factors in streaming data analysis: Challenges and opportunities for information visualization. Computer Graphics Forum **in publication** (2017)
15. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: the role of humans in interactive machine learning. AI Mag. **35**(4), 105–120 (2014)
16. Jasper, R.J., Blaha, L.M.: Interface metaphors for interactive machine learning. In: Proceedings of Human-Computer Interaction International: Augmented Cognition 2017, Vancouver, Canada, July 2017
17. Luce, R.D.: Response Times: Their Role in Inferring Elementary Mental Organization, vol. 8. Oxford University Press, Oxford (1986)
18. Townsend, J.T., Ashby, F.G.: Stochastic Modeling of Elementary Psychological Processes. CUP Archive, Cambridge (1983)
19. Amershi, S., Fogarty, J., Weld, D.: Regroup: Interactive machine learning for on-demand group creation in social networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 21–30. ACM (2012)
20. Arendt, D., Komurlu, C., Blaha, L.M.: CHISSL: A human-machine collaboration space for unsupervised learning. In: Proceedings of Human-Computer Interaction International: Augmented Cognition 2017, Vancouver, Canada (July 2017)
21. Endert, A., Fiaux, P., North, C.: Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2879–2888 (2012)
22. Cottam, J.A., Blaha, L., Zarzhitsky, D., Thomas, M., Skomski, E.: Crossing the streams: Fuzz testing with user input. In: 2017 IEEE International Conference on Big Data (Big Data). (Dec 2017) 4362–4371
23. Horvitz, E.: Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM (1999) 159–166
24. Cook, K., Cramer, N., Israel, D., Wolverton, M., Bruce, J., Burtner, R., Endert, A.: Mixed-initiative visual analytics using task-driven recommendations. In: 2015 IEEE Conference on Visual Analytics Science and Technology (VAST). (Oct 2015) 9–16
25. Felger, W., Schröder, F.: The visualization input pipeline-enabling semantic interaction in scientific visualization. In: Computer Graphics Forum. Volume 11, Wiley Online Library (1992) 139–151
26. Gotz, D., Zhou, M.X.: Characterizing users' visual analytic activity for insight provenance. Information Visualization **8**(1), 42–55 (2009)
27. Salvucci, D.D., Taatgen, N.A.: The multitasking mind. Oxford University Press (2010)

28. Watson, J.M., Strayer, D.L.: Supertaskers: profiles in extraordinary multitasking ability. Psychonomic Bulletin & Review **17**(4), 479–485 (2010)
29. Freeman, S., Pryce, N.: Growing Object-Oriented Software, Guided by Tests. 1st edn. Addison-Wesley Professional (2009)
30. Lamport, L.: Time, clocks, and the ordering of events in a distributed system. Commun. ACM **21**(7), 558–565 (1978)

Strengthening Health and Improving Emotional Defenses (SHIELD)

Seth Elkin-Frankston(✉), Arthur Wollocko, and James Niehaus

Charles River Analytics, 02138 Cambridge, MA, USA
selkinfrankston@cra.com

Abstract. Relaxation techniques such as deep breathing, and meditation can be used to gain more control of how individuals respond to stressful situations. While these techniques are becoming increasingly mainstream, there is still a stigma that can deter some users. Unfortunately, these populations stand to gain the most from developing these psychological tools. We set out to develop a mobile application to make relaxation training more appealing and approachable for the targeted population, which we believe is critical in order to gain wide scale usage. The Department of Defense has devoted substantial resources to developing stress prevention and resilience programs to combat the effects of stress; however, there is limited evidence to justify the cost and scope of current programs. We aimed to develop a low-cost, evidence-based mobile application tailored for the Marine Corps. Our solution, Strengthening Health and Improving Emotional Defenses (SHIELD), is designed to be a comprehensive approach based on the latest evidence-based strategies to train Marines to develop psychological resilience and promote healthy responses to adverse and stressful events. The overall SHIELD program is designed to promote gradual, self-paced practice, allowing Marines to complete training on a schedule that works for them.

Keywords: Resiliency · Psychological health · Psychological flexibility
Mindfulness · Self-regulation · Self-awareness

1 Objective and Significance

1.1 Problem Description

The psychological stress experienced by Marines can have negative consequences that reach beyond the individual; it affects job performance, personal relationships, and families. While most individuals can successfully respond to adverse situations, some may need help developing these essential skills at some point during their military tenure and when returning to civilian life. The inability to cope with chronic and acute day-to-day stressors, such as separation from loved ones or adjusting to the physical and mental demands of the Marine Corps, can leave individuals vulnerable to the harmful effects of stress, such as substance abuse or behavioral misconduct [1]. Teaching evidence-based strategies to promote psychological resilience—that is, the ability to adapt to stressful situations [2]—before exposure to stress can mitigate its costly and often harmful long-term effects [3, 4]. A number of programs have been

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 58–66, 2018.
https://doi.org/10.1007/978-3-319-91467-1_5

developed to strengthen psychological resilience, including BattleMind [5] (now known as Resilience Training), the Army's Comprehensive Soldier Fitness (CSF) program [6], and the Marine Corps' Combat Operational Stress Control [7]. However, these programs are costly, time consuming, and resource intensive [8]. The limited body of empirical evidence within military populations and the lack of a standard definition of effectiveness makes comparing programs difficult [9], and the inability to identify the effective components of a given program limits the overall utility of any concerted effort designed to develop psychological flexibility. If empirical evidence about independent components is readily available, users can focus on only the most effective components while ignoring those that are less effective, thereby optimizing the program for all individuals. Low-cost physiological sensors can also be used to provide objective feedback to further optimize and tailor any training program.

The Marine Corps requires a cost-effective, evidence-based psychological flexibility program that can be adaptively integrated into a variety of training approaches. This program must produce behavioral and physiological data to verify both short- and long-term effectiveness. A successful curriculum for training psychological flexibility to mitigate stress effects on Marines must meet three primary requirements:

1. The training program must be **driven by evidence and well-grounded scientific theory, while minimizing costly resource requirements**. The inclusion of evidence-based approaches increases the likelihood that the program will be effective. Current programs used by the armed forces often require a large staff of program managers and subject matter experts [9]; to reduce costs, the Marine Corps needs a program with inherently low operating costs and minimal resource requirements.
2. The program must **collect and record stress levels using available sensor technologies to demonstrate efficacy**. Sensor technologies are continually being improved and developed, and are rapidly proliferating, making it difficult to know what equipment will be available. A successful approach should make opportunistic use of sensors that are available in given environments (including the sensors on a Marine's personal equipment).
3. The program must **flexibly integrate into existing Marine Corps training and exercise regimes**. A low-cost, flexible curriculum is critical to meet this final challenge. A Marine's day-to-day activities are highly scheduled and constrained; with a flexible curriculum, Marines can independently practice resilience training without requiring instructor guidance or external resources.

1.2 Technical Approach

To meet these three requirements, we aimed to design and demonstrate a psychological flexibility program for Strengthening Health and Improving Emotional Defenses (SHIELD). SHIELD is a comprehensive approach based on the latest evidence-based strategies to train psychological flexibility and promote healthy responses to adverse and stressful events. To illustrate the technical merit, innovation, and soundness of our approach, we developed a prototype mobile application that delivers curriculum

modules, measures trainee progress, and interfaces with commercially available activity trackers to access physiological data.

Our technical approach addressed the three requirements defined above. First, the SHIELD training program must be driven by evidence and well-grounded scientific theory, while minimizing costly resource requirements. A successful approach should be evidence-based, highly adaptive, and accessible. We outlined a training and education-based curriculum to increase psychological flexibility using components from multiple evidence-based approaches (e.g., Mindfulness Based Stress Reduction (MBSR), relaxation response training, education about the physiology of stress, and yoga). These and other similar techniques emphasize mind-body awareness; they target and reduce stress with demonstrated efficacy in a diverse array of populations [10–15]. Our curriculum design includes techniques that minimize the amount of training and external resources required. We employed evidence-based practices from several existing stress reduction programs so that SHIELD is effective, low-cost, and requires little to no additional resources beyond those items typically available to Marines.

Second, to collect and record stress levels using available sensor technologies to demonstrate efficacy, we combined behavioral measures and body sensors that are already built into or easily integrated with personal mobile smartphone devices. This approach leverages the recently adopted Marine Corps Commercial Mobile Device Strategy, which allows Marines to carry personal mobile devices, a policy otherwise known as "Bring Your Own Device" (BYOD). Consumer-grade sensors that Marines are likely to carry provide high quality data that are comparable to more expensive technologies [16, 17]. In addition to the cost benefits of relying on BYOD sensors, Marines may be more likely to use wearable sensors in combination with the SHIELD intervention because they are more comfortable with wearable sensors they select as opposed to wearing a piece of required issued equipment. We developed a process to gather objective measures of efficacy that include the frequency and severity of disciplinary infractions, patterns of misconduct, and self-reported measures combined with low-cost embedded sensors to continuously track and model levels of stress using classification algorithms.

Third, to develop a curriculum that flexibly integrates into existing Marine Corps training and exercise regimes, our training program is delivered within a mobile smartphone application designed to motivate Marines to set and meet their own goals (without significant external oversight), and provide a fast and easy mechanism for Marines, instructors, and commanding officers to monitor program participation and progress. We focused on activity-based stress-reduction and stress coping training exercises delivered via a mobile platform to ensure flexible integration into the Marine Corps schedule. We developed and integrated a complete training module into the prototype mobile application and implemented user input and behavior logging capabilities. Additional modules focus on breathing exercises, mindfulness meditation, yoga, relaxation exercises, and awareness of the impact of stress on the mind and body. This focus will ensure a balance between effectiveness and usability.

2 Methods and Results

Our approach addresses three primary challenges. First, to develop a program that is driven by evidence and grounded scientific theory, while minimizing costly resource requirements. We define a training and education-based curriculum to increase psychological flexibility using components from multiple evidence-based approaches. Selected techniques emphasize mind-body awareness and target and reduce stress with demonstrated efficacy in a diverse array of populations [10]. Second, to collect and record stress levels using available sensor technologies to demonstrate feasibility, we combine behavioral measures and body sensors that are already built into or easily integrated with personal mobile smartphone devices. This approach leverages the recently adopted Marine Corps policy that allows Marines to carry personal mobile devices, an initiative otherwise known as "Bring Your Own Device". This is particularly useful because consumer-grade sensors that Marines carry provide high quality data that are comparable to more expensive technologies. Lastly, to develop a curriculum that flexibly integrates into existing Marine Corps training and exercise regimes, our training program is delivered by a mobile smartphone application designed to motivate Marines to set and meet their own goals (without significant external oversight), and provide a fast and easy mechanism for Marines, instructors, and commanding officers to monitor program participation and progress. Our approach focuses on activity-based stress-reduction and stress coping training exercises delivered via a mobile platform to ensure flexible integration into the Marine Corps schedule. Modules under development focus on breathing exercises, mindfulness meditation, relaxation exercises, and awareness of the impact of stress on the mind and body.

2.1 Results

We evaluated components of the SHIELD program (e.g., curriculum modules, user interfaces (UIs), descriptive language, assessment measures, and wearable sensor integration). Components were evaluated on a weekly basis as part of our internal review meetings, which were attended by Charles River and all subject matter experts. The validation and review process was driven by three primary concerns: (1) modifying content to be in line with the Marine Corps milieu; (2) modifying and optimizing content to be delivered via a mobile application; and (3) ensuring the effectiveness and reliability of the SHIELD program.

Curriculum Design and Requirements Analysis. Our goal was to define and analyze the core components of the SHIELD curriculum. Our aim was to (1) identify the limitations of current training programs; (2) analyze relevant Marine training environments; and (3) identify actionable items to design and optimize the effectiveness of the SHIELD program. We worked internally with subject matter experts to identify specific requirements for training psychological flexibility. Following a review of relevant Marine training environments (e.g., physical training, recreation and class-room instruction), we learned that the curriculum and content must be comprehensive and self-contained, as well as remaining consistent across language,

readability, accessibility, organization, and assessment within all aspects of the SHIELD program.

We reviewed current prevention and stress resilience training programs available to civilian and military populations. Some of these programs are widely applied, such as the Operational Stress Control and Readiness (OSCAR) program, others are narrowly focused, such as the Mindfulness-Based Mind Fitness Training (MMFT) program. Lessons from these and other programs [9] helped to shape and guide the overall design of the SHIELD program. For example, we learned that programs with large time commitments or required classroom training were less well-tolerated than programs that were less structured [9]. We also learned that a program must be evaluated over time. For these reasons, we designed SHIELD so it does not need classroom or in-person instructional training, but does include the ability to quantify and track progress over time.

Practices include exercises to develop self-awareness, such as simple breath practices (tactical breathing) and mindfulness (focus) exercises, as well as self-regulation practices, including body-scan, and heartbeat awareness (see Fig. 1). The included exercises were selected to due to their simplicity and effectiveness, even when self-guided.

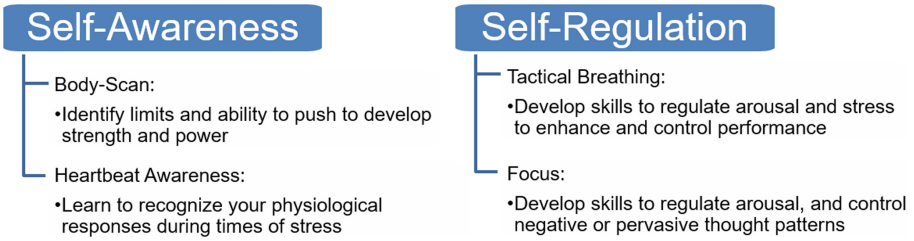

Fig. 1. The SHIELD curriculum includes exercises to train self-awareness and self-regulation

While the design can be highly structured based on individual need, we understand the need to develop a solution that is flexible enough to meet individual needs or to integrate into large-scale programs. Therefore, we designed the curriculum content to be selected à la carte to build customized programs or to be completed in full.

Mobile Application Development. We designed a mobile application to integrate sensor technologies, evaluate stress levels, deliver curriculum content, and provide an intuitive feedback and evaluation mechanism. The mobile application allows Marines to interact with the SHIELD program through a mobile application that guides them through each exercise and training modules (see Fig. 2). The mobile application also introduces the Marine to the SHIELD program, providing both a general overview and detailed instructions, as well as an interface to administer pre- and post-assessment measures.

The SHIELD mobile application design integrates sensor technologies, evaluates stress levels, delivers curriculum content, and provides an intuitive feedback and evaluation mechanism for Marines (see Fig. 3).

Fig. 2. SHIELD mobile application

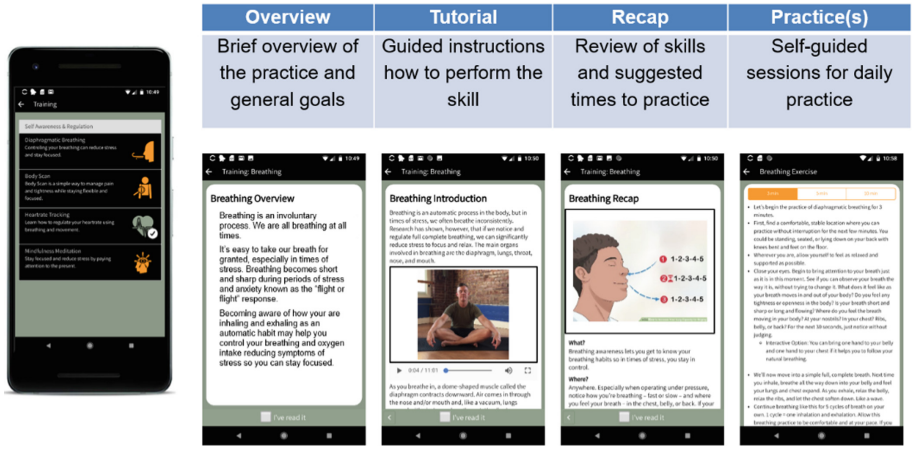

Fig. 3. Example of overview and guided training for breathing exercise

A key challenge and development risk for any remote application is connectivity with a central server so that information can be downloaded, uploaded, and communicated across the system in a reliable manner. To support this effort, we rely on a robust, cross-platform, mobile application communication framework to support essential types of communications on a back-end cloud server with intermittent connectivity. The framework is designed to maintain privacy and security of user data, both at rest and in transit. Besides the use of standard SSL for data encryption between the mobile application and back-end server, it requires mobile apps to be authorized and authenticated to connect to the server end point for communication. This HIPAA-compliant framework ensures only authorized mobile applications can communicate and push user data (so fake data cannot be added or overwrite other application data), and only authenticated mobile applications pull user data (so a user cannot see another user's data). All this is done without the user needing to remember user IDs that they can potentially share (the application is uniquely authorized and is not tied to

user ID). The user needs to remember only a PIN that opens the encrypted application. Sharing the pin will not allow another app to get the user's data.

The three types of communication supported by the framework include:

- **Mobile application pulls information from a website:** Whenever an application is opened by the user, or when a specific widget or UI element is enabled by the user, the pull functions that must be performed by the applications are registered with the framework. Each pull function essentially maps to a server-side end point that returns data in an open data-interchange format. Since the framework knows the encrypted authorization code, it uses the authorization to pull the response from the server end point and push the data to the application layer. The framework can also execute scripts (that change the state of UI elements) based on the statistics returned by the server (e.g., count or value of a data element). The framework can be configured to respond with appropriate messages upon communication failure, including connection failure. The framework includes a handshake mechanism to ensure reliability of message receipt, delivery guarantees, and to avoid message duplication.
- **Mobile application pushes information to a website:** This is similar to the first type of communication, except the request to push data includes authorization as well as the payload. However, associated with push activity is a network-aware message queue framework that ensures that a message created for an end point successfully reaches the endpoint, even though there may not be network connectivity at the time the message was created. The framework achieves eventual consistency and works fine with intermittent connectivity.
- **The mobile device receives a push or remote notification:** In many mobile applications, a server-side cloud generates data targeted for a mobile application. The mobile OS allows apps to pull the data only when they are opened. One example is email and social media apps. Leveraging on push notification systems provided by Apple and Google, the mobile communication framework allows the server to send mobile-device-targeted custom notifications, which when opened, execute appropriate pulls and calls appropriate application layer hooks or listeners to return the required data.

3 Discussion

Charles River Analytics set out to demonstrate the feasibility of a portable psychological flexibility program. Our approach was motivated by three primary goals: (1) develop a program that is driven by evidence and well-grounded in scientific theory, while minimizing costly resource requirements; (2) support capabilities to collect and record stress levels using COTS sensor technologies; and (3) design a program that can flexibly integrate into existing Marine Corps training and exercise regimes. We demonstrate an approach to develop a platform for introducing and teaching relaxation techniques that support psychological resilience to audiences that may otherwise be reluctant to engage in such practices. While relaxation techniques such as meditation and mindfulness training are becoming increasingly popular amongst the general

population, we recognize that previously existing biases may deter its usage in certain audiences. Our initial evaluation provides a methodology for delivering relaxation techniques in a manner that is more approachable to groups who may otherwise be resistant, such as the Marines.

The operational Navy concept of the SHIELD system is that Marines will access the SHIELD application by downloading the app from Government app stores or Google Play. Marines can then engage with the application and individual modules whenever their schedule permits. No module is designed to require more than 5 or 10 min to complete. The entire curriculum can be completed in as little as six weeks, or Marines can choose to focus only on specific modules that suit their needs. Each week of modules focuses on a specific topic (e.g., Yoga, breathing exercises). While each week contains a number of specific modules, there are also companion tools that the Marine can select to help practice specific skills (e.g., Three Part Breath). Other operational uses include officers ensuring their Marines are fit for duty. For example, officers overseeing several Marines can use the data analysis and prediction capabilities within SHIELD to leverage physiological data gathered during normal daily activities, which are automatically inserted into SHIELD's stress classification models. These models can be used to make mission-critical activity or return-to-duty assessments on each Marine, and ensure the Marine has access to the support they require.

The *future naval relevance* of the SHIELD system is to provide the Navy with a low-cost, non-invasive sensing, assessment, and treatment option for training psychological flexibility and enhancing the resilience and performance of Marines. SHIELD accomplishes this in several ways: (1) the design of appropriate curriculum material; (2) the integration and capitalization on features of commercially available, ubiquitous sensor platforms; (3) the analysis of sensor data to enable prediction of stress; (4) the unobtrusive delivery of these capabilities through a mobile application.

References

1. MacManus, D., Rona, R., Dickson, H., Somaini, G., Fear, N., Wessely, S.: Aggressive and violent behavior among military personnel deployed to Iraq and Afghanistan: prevalence and link with deployment and combat exposure. Epidemiol. Rev. **37**, 196–212 (2015)
2. Jenson, J.M., Fraser, M.W.: A risk and resilience framework for child, youth, and family policy. In: Jenson, J.M., Fraser, M.W. (eds.) Social Policy for Children and Families: A Risk and Resilience Perspective. SAGE (2006)
3. Charney, D.S.: Psychobiological mechanisms of resilience and vulnerability: implications for successful adaptation to extreme stress. Am. J. Psychiatry **161**, 195–216 (2004)
4. Mealer, M., Conrad, D., Evans, J., Jooste, K., Solyntjes, J., Rothbaum, B., Moss, M.: Feasibility and acceptability of a resilience training program for intensive care unit nurses. Am. J. Crit. Care **23**, e97–e105 (2014)
5. Adler, A.B., Bliese, P.D., McGurk, D., Hoge, C.W., Castro, C.A.: Battlemind debriefing and battlemind training as early interventions with soldiers returning from Iraq: randomization by platoon. J. Consult. Clin. Psychol. **77**, 928–940 (2009)
6. Cornum, R., Matthews, M.D., Seligman, M.E.P.: Comprehensive soldier fitness: building resilience in a challenging institutional context. Am. Psychol. **66**, 4–9 (2011)

7. Corps, U.M., Navy, U.S.: Combat and operational stress control. MCRP 6-11C/NTTP 1-15M (2010)
8. Blakeley, K., Jansen, D.J.: Post-traumatic stress disorder and other mental health problems in the military: oversight issues for congress (2013)
9. Meredith, L.S.: Promoting psychological resilience in the U.S. military. RAND **1**, 2 (2011)
10. Crawford, C., Wallerstedt, D.B., Khorsan, R., Clausen, S.S., Jonas, W.B., Walter, J.A.G.: A systematic review of biopsychosocial training programs for the self-management of emotional stress: potential applications for the military. Evid. Based Complement. Alternat. Med. **2013**, 1–23 (2013)
11. Gaab, J., Sonderegger, L., Scherrer, S., Ehlert, U.: Psychoneuroendocrine effects of cognitive-behavioral stress management in a naturalistic setting–a randomized controlled trial. Psychoneuroendocrinology **31**, 428–438 (2006)
12. Hölzel, B.K., Carmody, J., Evans, K.C., Hoge, E.A., Dusek, J.A., Morgan, L., Pitman, R.K., Lazar, S.W.: Stress reduction correlates with structural changes in the amygdala. Soc. Cogn. Affect. Neurosci. **5**, 11–17 (2010)
13. Kim, S.H., Schneider, S.M., Bevans, M., Kravitz, L., Mermier, C., Qualls, C., Burge, M.R.: PTSD symptom reduction with mindfulness-based stretching and deep breathing exercise: randomized controlled clinical trial of efficacy. J. Clin. Endocrinol. Metab. **98**, 2984–2992 (2013)
14. Tang, Y.-Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., Yu, Q., Sui, D., Rothbart, M.K., Fan, M., Posner, M.I.: Short-term meditation training improves attention and self-regulation. Proc. Natl. Acad. Sci. **104**, 17152–17156 (2007)
15. Williams, K.A., Kolar, M.M., Reger, B.E., Pearson, J.C.: Evaluation of a wellness-based mindfulness stress reduction intervention: a controlled trial. Am. J. Health Promot. AJHP. **15**, 422–432 (2001)
16. Tao, W., Liu, T., Zheng, R., Feng, H.: Gait analysis using wearable. Sensors. **12**, 2255–2283 (2012)
17. Vanderlei, L.C.M., Silva, R.A., Pastre, C.M., Azevedo, F.M., Godoy, M.F.: Comparison of the Polar S810i monitor and the ECG for the analysis of heart rate variability in the time and frequency domains. Braz. J. Med. Biol. Res. **41**, 854–859 (2008)

Assessment of Wearable Tactile System: Perception, Learning, and Recall

Linda R. Elliott[1,2(✉)], Bruce J. P. Mortimer[1,2], Rodger A. Pettitt[1,2], and Robert E. Wooldridge[1,2]

[1] Army Research Laboratory, Fort Benning, GA, USA
linda.r.elliott.civ@mail.mil
[2] Engineering Acoustics, Inc., Casselberry, FL, USA

Abstract. Previous research investigated concepts of tactile salience and core variables mediating effects on human perception and learning, resulting in validation of independent scaled ratings of tactile salience. This approach provides an integrated and systematic approach to assess effectiveness of tactile displays. We report an initial series of comparative tests of various multi-tactor cues, or tactions. Tactions were developed to vary in temporal sequencing and amplitude. In the first experiment 8 tactions were used; a follow-up investigation used 12. In this report we summarize results, with a focus on experiment methods association with measurement of tactile salience, ease of learning, and ease of recall.

Keywords: Army robotic systems · Tactile cues · Army tactile systems
Multimodal systems

1 Introduction

The development of vibrating tactors has focused on physiological characteristics when optimizing systems for human perception [1–3]. Neuropsychological research has focused on cognitive and neural correlates of tactile perception and memory [4]. An understanding of both the physiological and neurophysiological requirements is needed for the development of advanced human-in-the-loop, tactile based systems.

Vibrotactile cues can provide user information ranging from simple alerts for attention management (e.g., cell phone vibrations) to direction, spatial orientation, and more complex communications [5–9]. Quantitative meta-analyses of over 40 empirical studies showed significant positive impacts of tactile cueing on operational workload and performance, particularly when workload and attentional demands are high and/or when tactile cues are added to augment visual cues [5]. Complex vibrotactile cues driven from dynamic activation of multiple tactors have been developed to be intuitively understood, with little or no training [10–12]. Through these studies, tactile systems have demonstrated several key advantages when added to dismount Soldier navigation and/or communication systems, including human- robot interaction systems (e.g., communications from other Soldiers and also from robotic sensors).

In this report, we describe efforts to further investigate attributes of multi-tactor taction cues used to communicate a variety of alerting messages, with regard to

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 67–77, 2018.
https://doi.org/10.1007/978-3-319-91467-1_6

perceptions of salience, ease of learning and recognition, and recall. We build upon previous investigations that found differences in response time and accuracy based on characteristics of tactors, such as amplitude and gain, and identified tactor engineering characteristics most likely to be perceived and recognized [13]. In this report we focus on additional aspects of multi-tactor taction differences, with regard to temporal sequencing and complexity of tactor signaling. Results described here are based on two studies, a preliminary exploration using 8 tactions, and a second study refined by results of the first experiment, that used 12 tactions.

Tactors and Tactile Belt. For this effort, we utilized 2 types of tactors developed to optimize human tactile perception, the Engineering Acoustics Inc (EAI) C-3 tactor and the EAI EMR tactor, as shown in Fig. 1. The EMR produces about 0.7 mm displacement amplitude, with an operating frequency around 80–120 Hz. The EMR uses rotational motors that are suspended in a unique actuator configuration. The C-3 tactor is similar in performance to the C-2, but is lighter and has a smaller diameter. The C-3 utilizes a unique engineering approach proven to be particularly salient under strenuous movement [13, 14]. The contact with the skin is from the predominant moving mass, driving the skin with perpendicular sinusoidal movement that is independent of the loading on the housing [14]. The tactors are mounted in two rows (one row of 8 EMR tactors and one row of 8 C-3 tactors) within a belt form factor (Fig. 1).

Fig. 1. EAI EMR and C-3 tactor transducers (left to right), and 16-tactor tactile belt

Tactions. Tactions refer to the tactile patterns that are generated on a tactile array with characteristic features such as; spatial location, movement, temporal (pulse, rhythm and meter), frequency and intensity. Tactions are felt by the user, interpreted and associated with meaning. In the first study, eight tactions were created through using visual graphics software to easily create, save, and modify taction characteristics. A subset of the initial tactions were developed and used in previous studies using Soldier subjects and were found to be easy to perceive and interpret [15]. For the second study, four additional taction swere developed.

Tactions for our study were developed to vary systematically along two dimensions: (a) temporal sequencing and (b) complexity of tactor signaling characteristics. Regarding temporal sequencing, we followed Barber et al. [16] for the definition of static versus dynamic patterns. Static tactions present a constant pattern using the same tactors in a repetitive fashion. Dynamic tactions using used a sequenced presentation on different tactile locations that provides a sensation of movement across the tactors.

We defined the level of complexity as Standard or Complex. Standard tactions use combinations of tone burst pulsating vibrotactile patterns as described in the review by Sarter and Jones [1]. These tone burst patterns are typically single frequency and can be pulse length modulated as described by Brewster and Brown [17]. Complex tactions use amplitude and/or frequency sweeps and/or short pulsatile sequences to create somatosensory illusion experiences (usually associated with the perception of movements). Examples would include various illusions such as the cutaneous rabbit [18], paint-brush illusion [19] and phi (motion; [20]).

We describe one example of each of the four kinds of tactions, using a screen shot of the taction creation software, which describes which tactor is activated. Tactors 1–8 are EMR tactors arranged sequentially in one row of the belt, tactors 9–16 describe C-3 tactors arranged similarly in a second row on the belt. Tactor 1 and 9 are thus located on the belly of the participant. In experiment 2, we used 12 tactions, four of each category.

Standard/Static. Standard/static tactions were typified by static "pulsing" of tactors that did not vary in signal characteristics. They were simple and repetitive. As an example, the NBC (nuclear biological chemical threat) taction comprises a dynamic sequence of alternating pulses (between the back left front right). It was implemented on C-3, tactors 9–16, as portrayed in Fig. 2. Each tactor pulse in the sequence comprises a 250 Hz tone-burst at the maximum displacement.

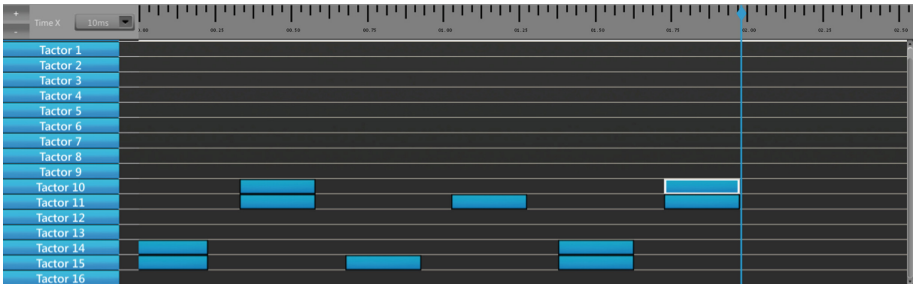

Fig. 2. NBC "standard/static" taction using the C-3 (tactors 9–16)

Standard/Dynamic. Standard/dynamic tactions used standard tactor signal characteristics, with a dynamic temporal sequencing on multiple tactor locations. As an example, the adapted Rally taction comprises a dynamic sequence of pulses starting in the center (belly) and moving clockwise around the body that is repeated twice. It was implemented on C-3, tactions 9–16, as portrayed in Fig. 8. Note that this is a slightly different implementation of Rally from previous experiments; tactors were somewhat overlapping in duration (Fig. 3).

Complex Static. Complex/static tactions were "static" in the sense that the pulse stimuli were presented on fixed tactors in the belt array. However, each tactor pulse was "complex" in that the amplitude or gain was ramped linearly. As an example, the IED taction utilized 9 simultaneous pulses. It uses both the EMR and C-3 tactors in the

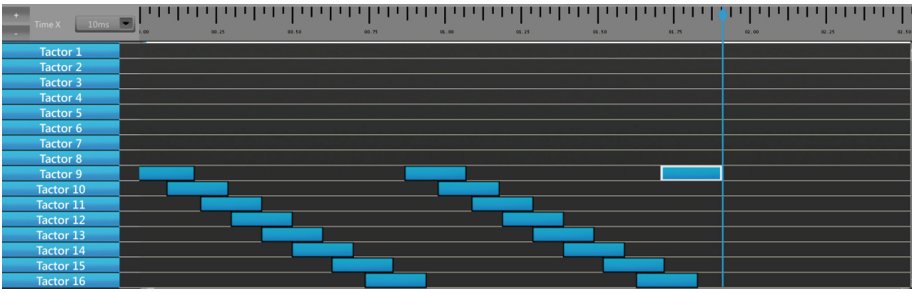

Fig. 3. Rally "standard/dynamic" taction utilizing a pulse sequence on the C-3 (tactors 9–16)

taction. Specifically, each tactor was pulsed on for 1,500-ms tone-burst duration while the gain was linearly varied from maximum to zero (254-1 gain) (see Fig. 4). Thus, this tactions would be felt as an initial strong burst that "levels out" (e.g., IED).

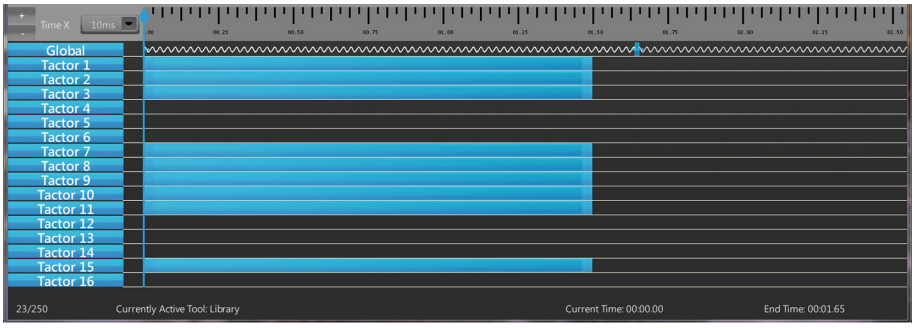

Fig. 4. IED "complex/static" taction utilizing EMR (tactors 1–8) and C-3 (tactors 9–16)

Complex Dynamic. Complex/dynamic tactions used tactors with complex characteristics (i.e. "ramped" characteristics of gain or amplitude), along with dynamic temporal sequencing. For example, the Move Up taction comprises of sequenced ramps on two (or four) tactors in adjacent rows tactors. Thus, the C-3 and EMR tactors are used and ramped simultaneously. This taction pattern is dynamic following the tactile "paintbrush" illusion [19] and provides a sensation of back and forth movement over the front torso (Fig. 5).

2 Method

Overview of the Assessment. While two experiments were conducted, we focus on the second, which was refined based on experiment 1 results. In experiment 1, participants easily learned eight tactions, allowing us to add 4 tactions in experiment 2, ensuring we had three tactions of each type (12 tactions). Individual tactions were also

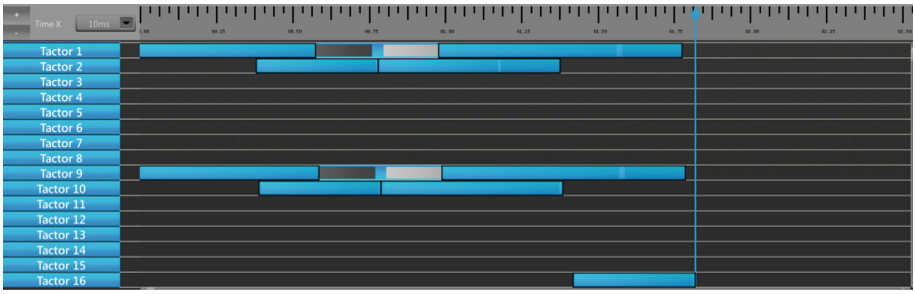

Fig. 5. MoveUp "complex/dynamic" taction that utilizes both the EMR (tactors 1–8) and C-3 (tactors 9–16)

refined in response to Soldier feedback from experiment 1; for example, shortening the overall length of a particularly lengthy taction, to be more similar in length to the other tactions, and creating more distinction with regard to the Rally taction.

Twenty Soldiers assigned to operational and training units located at Fort Benning, GA participated in the five-day assessment. The average of age of the Soldiers was 27.25 years. Soldiers ranked from E-4 to E-5, with 70% of the Soldiers being E-4. 40% of Soldiers reported an average of 3.38 years for time in service, while 55% reported less than one year of time in service. Soldiers were informed of the nature and purpose of the investigation and given the opportunity to opt out with no repercussions. They provided responses to demographic questionnaires, were assigned a roster number, which corresponded to a counterbalanced experiment design where each Soldier (a) provided ratings of tactile salience on each taction, (b) experienced training on taction meaning, then participated in two performance trials, one where s/he was standing stationary, and one where s/he moved along a 2×4 beam placed on the floor in a slightly raised square pattern. After a three-hour break, each Soldier participated in two more performance trials, to explore any effects on recall of taction meanings.

Tactile Salience: Training and Measurement. Each Soldier was given an oral and hands-on training of the tactile system. They donned the belt and wore headphones that emitted pink noise, to eliminate any audio cues associated with tactions. The experimenter activated each taction in turn, to give the recipient an overall familiarity of all tactions.

Soldiers were given the following instructions: "We will be presenting you with 12 different patterns of tactile signals. We will let you feel each of them first, to give you an idea of what each one feels like, and how they differ. Then, we will give you the signals one at a time, and ask you to give each one a rating, from one to five, that indicates how strongly, or easily, you think each one can be felt." They were presented with a poster describing the 5 point scale for salience, ranging from 1 = weak, blurred, faint, vague to 5 = noticeable, distinct, strong, salient. A previous investigation established the reliability of this rating-based approach to measurement, compared to more traditional forced-choice methodology [21]. Each Soldier provided ratings of salience for each taction. Each taction was presented twice.

Taction Meaning: Training and Ease of Learning. After ratings of salience were collected, each Soldier was trained on the meaning of each taction. The tactions were trained first in sets of three. Each set included all tactions of a particular type (e.g., standard static). Three tactions were presented, with meanings. The instructor would repeat each taction of this set, in random order, until the Soldier labeled each taction correctly, three times in a row. After all 12 tactions were presented to the Soldier in this way, the instructor would repeat each of the 12 tactions, in counterbalanced order, until the Soldier was able to correctly identify each taction three times in a row. The instructor documented the number of times each taction was repeated to achieve requisite performance.

Performance Trials. After each Soldier was fully trained on the meaning of each taction, they participated in two performance trials in the morning, and two performance trials about three hours later. They did not get refresher training before the afternoon sessions and were asked not to discuss their experience with each other. An experimenter accompanied them during this break time. Sessions were counterbalanced according to Table 1.

Table 1. Assignment of soldiers to conditions. The stationary test condition refers to tactions being presented with the participant standing, and moving refers to one where s/he was required to move along a 2 × 4 beam placed on the floor in a slightly raised square pattern.

Roster	Morning		Afternoon	
1, 5, 9, 13, 17	Stationary A	Moving B	Moving C	Stationary D
2, 6, 10, 14. 18	Stationary B	Moving A	Stationary D	Moving C
3, 7, 11, 15, 19	Moving C	Stationary D	Moving A	Stationary B
4, 8, 12, 16, 20	Moving D	Stationary C	Stationary B	Moving A

A, B, C, and D refer to 4 counterbalanced orders of taction presentation

3 Results

3.1 Salience

As the figure below shows, both main factors (Static vs. Dynamic, Standard vs. Complex) and the interaction term had main effects on ratings of salience. Repeated measures analyses of variance show a significant main effect for the static versus dynamic variable ($F, 1, 19 = 6.67$, $p < 0.02$, $\eta\rho^2 = 0.26$) and for the standard versus complex variable ($F\ 1, 19 = 56.18$, $p < 0.001$, $\eta\rho^2 = 0.75$) and for the interaction term ($F\ 1,\ 19 = 97.37$, $p < .001$, $\eta\rho^2 = 0.83l$). Effect sizes as calculated by partial eta squared ($\eta\rho^2$) were high. Results indicate that while static tactions were higher in salience, compared to dynamic tactions; this difference was significantly larger for complex tactions (Fig. 6).

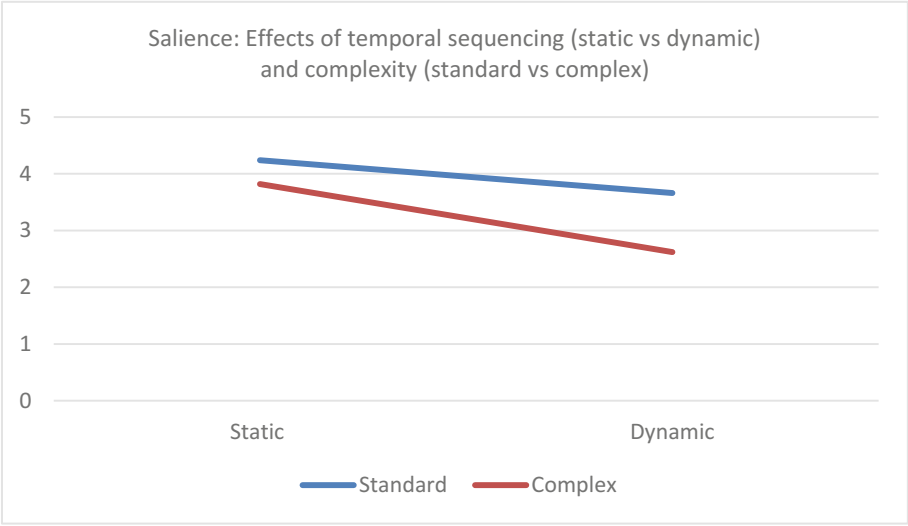

Fig. 6. Salience: simple vs. complex tactions: static vs. dynamic

3.2 Ease of Learning

Soldiers provided very high ratings of training effectiveness, ranging from means of 5.20 to 5.80 (7pt. scale). Each time the Soldier participant made an error, the instructor noted the error, and the taction that it was mistaken for, and communicated the correct taction meaning. This process was repeated, going through each of the twelve tactions, until the Soldier correctly identified each taction three times in a row. Tactions associated with highest total error during this learning process were "Target Detected" (standard/dynamic; 13 errors), "Move up" (complex/dynamic; 12 errors), "Wheel spin" (standard/dynamic; 5 errors), "Disperse" (complex/dynamic; 5 errors), and "Freeze" (5 errors). Aside from "Freeze", these tactions were dynamic. The mean number of repetitions to learn the tactions to criterion performance ranged from 2.00 to 2.90. Some Soldiers learned all tactions very easily, while some had difficulty with most tactions, requiring four to five repetitions of all 12.

3.3 Performance and Recall

Figure 7 provides mean performance scores for each taction, by time of day. There were few differences in recall accuracy due to time of day; mean accuracy for tactions remained high, for tactions that were associated with high accuracy in the AM. There was some decline for standard dynamic tactions (Target Detected, Rally, Wheel spin). Repeated measures ANOVA examining three factors and interactions showed a significant effect due to Static vs. Dynamic factor ($F\ 1, 19 = 30.16$, $p < 0.00$, $\eta\rho^2 = 0.61$), and a significant interaction between Standard vs Complex factor and AM vs. PM ($F\ 1, 19 = 4.13$, $p = 0.05$, $\eta\rho^2 = 0.13$).

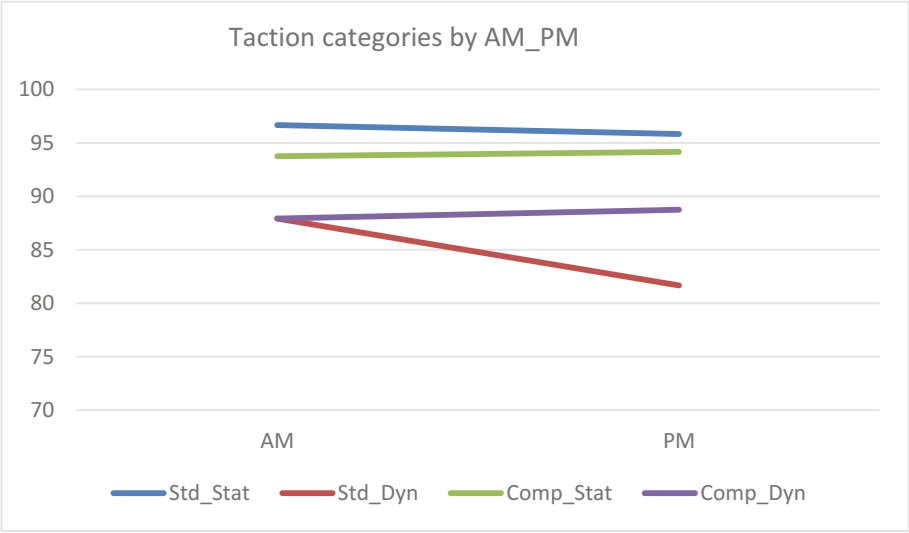

Fig. 7. Taction categories by AM_PM

3.4 Performance and Movement

Figure 8 provides mean accuracy scores for each taction category, by participant movement test condition. The movement test conditions were counterbalanced against order. Values were quite similar across movement conditions, with the exception of Wheelspin, which differed from 87.5% (stationary) to 71.3% (balance beam). Repeated measures ANOVA showed a significant effect for Static vs. Dynamic factor ($F\ 1,\ 19 = 37.74$, $p < 0.0001$, $\eta\rho^2 = 0.66$), but not for Standard vs. Complex ($F\ 1,\ 19 = 0.33$, $p = 0.57$, $\eta\rho^2 = 0.02$), or Stationary vs. Movement ($F\ 1,\ 19 = 2.41$, $p = 0.14$, $\eta\rho^2 = 0.11$). There was a significant interaction for Static vs. Dynamic and Stationary-Movement, showing that the difference in movement condition had an effect depending on whether the taction was Static vs. Dynamic ($F\ 1,\ 19 = 6.65$, $P < 0.02$, $\eta\rho^2 = 0.26$). Other interactions were not significant.

Subjective Feedback. Soldier ratings were overall positive, reporting high ratings for system comfort and fit. Ratings also indicated the tactions were easy to learn and recognize. Mean ratings were generally high for ease of perceiving the tactions, in general, and while moving. When asked which tactions were easiest to learn and remember, they listed "Rally" (n = 16, standard/dynamic) and "Point Right" (n = 13, standard/static). Most difficult tactions were listed as "Target Detected" (n = 14, standard/dynamic), "Move Up" (n = 12, complex/dynamic) and "Wheel spin" (n = 10, standard/dynamic). Post-session ratings of "ease of recognizing each cue", based on a 7pt. scale where 7 = "extremely easy to recognize" were highest for "Point Right" (mean = 7.00) and "Rally" (mean = 6.68). While Rally is a dynamic taction, and thus predicted to be less easily learned, it is also a taction that directly emulates the Army hand and arm signal for Rally, when the hand is raised overhead and is rotated in a

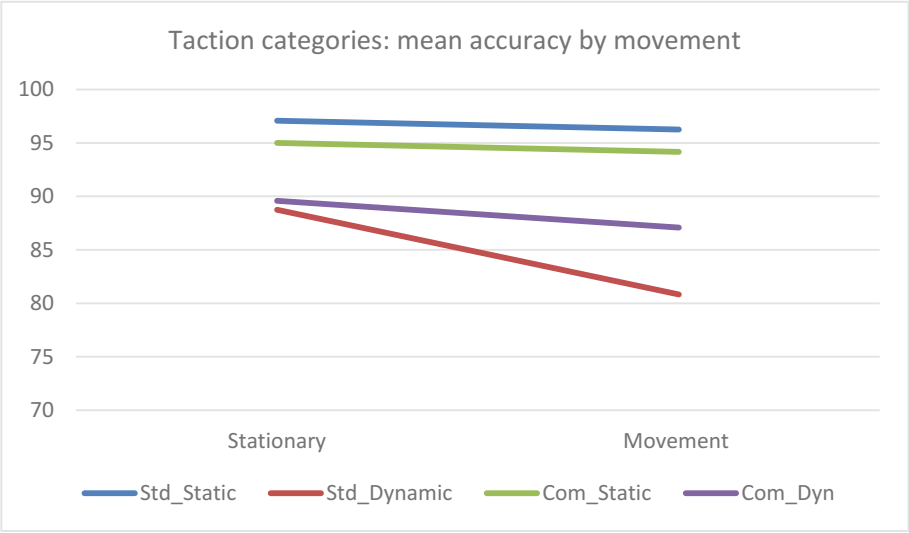

Fig. 8. Mean accuracy scores for each taction category, by participant movement test condition.

circular motion. This "link" to an existing and familiar cue is likely more easily learned and remembered. A follow-up study will examine the effects of longer time durations between initial and subsequent performance.

The mean rating (7pt. scale) for "operational relevance" was relatively high (mean = 5.41, SD = 1.12). Soldiers indicated that the system, when developed to be combat ready (e.g., secure network to Army systems, rugged, etc.), would be useful for situations requiring noise disciple and when visibility is low (e.g., night operations, dense vegetation, smoke, etc.). Suggestions were offered regarding form, fit, power usage, and additional capabilities [21, 22].

4 Summary and Conclusions

Salience. Consistent with results from Experiment 1, mean ratings were higher for standard tactions compared to complex, and higher for static tactions compared to dynamic. However, Experiment 2 results, based upon a more powerful experiment design, describe a significant interaction where complex tactions were more negatively affected by dynamic characteristics. The trend suggests that simple repetitive tactions are perceived as more salient.

Ease of Learning. Averaging across taction categories, standard static tactions also appear to be easiest to learn, having the lowest rate of error (total = 7), followed by complex static (total = 10), standard dynamic (total = 19) and complex dynamic (total = 20). Results follow the same trend as for salience: results suggest standard static tactions are more easily learned.

Performance. Overall mean accuracy, averaged over time and movement condition, showed that Soldiers learned 9 of 12 tactions with over 90% accuracy (92–99%). Soldiers had most difficulty with "Move Up" (complex/dynamic; 78%), "Wheel spin" (standard/dynamic), and "Target Detected" (standard/dynamic). ANOVA showed differences in performance due to the static vs dynamic factor were significant.

Recall. Comparison of AM versus PM performance showed little overall degradation in performance. However, analyses of taction categories showed significant decline in performance for more dynamic tactions, and for an interaction effect, such that standard dynamic tactions were most negatively affected by time.

Movement. Comparison of stationary and balance beam conditions also showed little overall degradation in performance, with the exception of standard dynamic tactions, such that standard dynamic tactions were significantly lower in the movement condition.

Results show consistent trends in favor of standard/static tactions, in terms of ease of perception (i.e., salience), learning, recognition, and recall. These results serve to better guide developers of tactile cueing displays, when developing tactions for non-directional alerts. Results also suggest the importance of familiarity of a taction, when the perceived taction could be "linked" to an existing concept. In this way, the Rally taction, while dynamic, was more easily learned and recalled, compared to other dynamic tactions. The rally taction directly emulated the circular Soldier hand and arm signal for "Rally".

Results also showed that participants could easily learn up to twelve tactions in a relatively short period of time. Experiment 1, which used eight tactions, had very little variance in learning or performance. While more variance was associated with twelve tactions, there were some participants who easily learned the 12 tactions with little repetition and 100% accuracy in performance. These results will inform subsequent investigations of taction characteristics and individual differences.

References

1. Jones, L., Sarter, N.: Tactile displays: Guidance for their design and application. Hum. Factors **50**, 90–111 (2008)
2. Mortimer, B., Zets, G., Cholewiak, R.: Vibrotactile transduction and transducers. J. Acoust. Soc. Am. **121**, 2970 (2007)
3. Mortimer, B., Zets, G., Mort, G., Shovan, C.: Implementing effective tactile symbology for orientation and navigation. In: Jacko, J.A. (ed.) HCI 2011. LNCS, vol. 6763, pp. 321–328. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21616-9_36
4. Gallace, A., Spence, C.: The cognitive and neural correlates of tactile memory. Psychol. Bull. **135**(3), 380–406 (2009)
5. Elliott, L., Coovert, M., Redden, R.: A summary review of meta-analysis of tactile and visual displays. In: Proceedings of the 13th International Conference on Human-Computer Interaction, San Diego, CA, June 2009
6. Elliott, L.R., Mortimer, B., Hartnett-Pomranky, G., Zets, G., Mort, G.: Augmenting soldier situation awareness and navigation through tactile cueing. In: Yamamoto, S. (ed.) HIMI 2015. LNCS, vol. 9172, pp. 345–353. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20612-7_33

7. Rupert, A.H.: An instrumentation solution for reducing spatial disorientation mishaps. IEEE Eng. Med. Biol. **19**, 71–80 (2000)
8. Rupert, A.H.: Tactile situation awareness system: proprioceptive prostheses for sensory deficiencies. Aviat. Space Environ. Med. **71**(9), A92–A99 (2000)
9. Van Erp, J.: Tactile Displays for Navigation and Orientation: Perception and Behavior. Mostert and van Onderen, Leiden (2007)
10. Brill, C., Gilson, R.: Tactile technology for covert communications. In: Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society (2006)
11. Lylykangas, J., Surakka, V., Rantala, J., Raisamo, R.: Intuitiveness of vibrotactile speed regulation cues. ACM Trans. Appl. Percept. **10**(4), 1–15 (2013)
12. Roady, T., Ferris, T.: An analysis of static, dynamic, and salutatory vibrotactile stimuli to inform the design of efficient haptic communication systems. In: Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society (2012)
13. Pettitt, R., Redden, E., Carstens, C.: Comparison of army hand and arm signals to a covert tactile communication system in a dynamic environment. Army Research Laboratory (US), Aberdeen Proving Ground (MD). Technical report No. ARL-TR-3838, August 2006. http://www.arl.army.mil/arlreports/2006/ARL-TR-3838.pdf
14. Mortimer, B., Zets, G., Cholewiak, R.: Vibrotactile transduction and transducers. J. Acoust. Soc. Am. **121**, 2970 (2007)
15. Gilson, R., Redden, E., Elliott, L.: Remote tactile displays for future soldiers. Army Research Laboratory (US), Aberdeen Proving Ground (MD). Technical report No. ARL-SR-0152 (2007). http://www.arl.army.mil/arlreports/2007/ARL-SR-0152.pdf
16. Barber, D.J., Reinerman-Jones, L.E., Matthews, G.: Toward a tactile language for human–robot interaction. Hum. Factors **57**(3), 471–490 (2014)
17. Brewster, S.A., Brown, L.M.: Tactons: structured tactile messages for non-visual information display. In: ACS Conferences in Research and Practice in Information Technology Australasian User Interface Conference 2004, 18–22 January 2004, Dunedin, New Zealand, vol. 28, pp. 15–23 (2004)
18. Geldard, F.A., Sherrick, C.E.: The cutaneous "rabbit": a perceptual illusion. Science **178** (4057), 178–179 (1972)
19. Israr, A., Poupyrev, I.: Tactile brush: drawing on skin with a tactile grid display. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2011)
20. Burtt, H.E.: Tactual Illusions of Movements. J. Exp. Psychol. **2**, 371–385 (1917)
21. Elliott, L., Mortimer, B., Pomranky-Hartnett, G., Pettitt, R., Rapozo, F., Rapozo, A., Wooldridge, R.: Tactile cues: taction characteristics, saliences, ease of learning, and recall. Army Research Laboratory (US), Aberdeen Proving Ground (MD). Technical report in review, August 2006
22. Mortimer, B.J.P., Elliott, L.R.: Identifying errors in tactile displays and best practice usage guidelines. In: Chen, J. (ed.) Advances in Human Factors in Robots and Unmanned Systems. AISC, vol. 595, pp. 226–235. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60384-1_22

Visualization of Network Security Data by Haptic

Manabu Ishihara[1(✉)] and Taiki Kanayama[2]

[1] Department of Innovative Electrical and Electronic Engineering, National Institute of Technology, Oyama College, Oyama 323-0806, Japan
ishihara@oyama-ct.ac.jp
[2] Advanced Course of Electrical and Computer Engineering, National Institute of Technology, Oyama College, Oyama, Japan

Abstract. Antivirus software technology is a security technology for analysis of networks and detection of information leakage and virus software. Computers are designed to process the detected viruses in the background. Here we report our design, construction, and investigation of a prototype system to sense virus-induced traffic anomalies in a network and react haptically, as a form of haptic visualization security technology that can provide physical experience of anomalies generated when viruses occur, in training beginners, as well as other applications.

Keywords: Haptic device · Antivirus software · Visualization technologies

1 Introduction

All the many methods of cyberattacks can present a threat to the underpinnings of society. Antivirus software is a type of security technology that functions in an invisible world (i.e., the background) to analyze operations, discover anomalies, and eliminate the related viruses. Visualization of security technology [1] enables intuitive understanding of virus-induced anomalies by rendering virus-peculiar behavior recognizable to humans. A typical example is the NIRVANA-Kai system developed by the National Institute of Information and Communications Technology [2] to monitor cybersecurity. The visualizations can also be used for education and training of non-specialists. However, relatively few studies have been reported on performing security technology visualization, haptization, and other modes of conversion to sensory perception.

Here we describe our development and investigation of a system for intuitive representation by haptic display of DDoS, using IP address analysis [3] with network security in the background state and assuming an attack on the Web.

2 Equipment Used

The SensAble PHANToM Omni haptic presentation device (hereafter, Omni) shown in Fig. 1 was used in this study. The system control computer was constructed with a 3.40 GHz 4 GB RAM Intel Core i7-2600 CPU, with Windows 7 Professional as the

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 78–93, 2018.
https://doi.org/10.1007/978-3-319-91467-1_7

OS. An Imada DPS-5 digital force gauge was used as a torque meter to measure the actual force presented by the Omni. Microsoft Visual C++ 2008 was used as the development environment. Open Haptics was used as the library file for control of the Omni and WinPcap for control of packet capture.

Fig. 1. PHANToM Omni.

3 System Overview

The system performs the four steps of (1) packet capture, (2) analysis, (3) drawing, and (4) haptic presentation. It automatically performs packet capture on startup, searches the packet for IP address and time to live (TTL), and displays images for the packets on the screen. Finally, it presents the reaction to the user. The user operates Omni to touch the virtual objects on the screen and sense the traffic amount response on the hand. IP packets ordinarily reach their destination after passing through routers less than 30 times, but some have unusually long TTL values. Many of those are generated by special software, and the long TTL value may indicate occurrence of an anomalous communication. Yamada et al. have described a method of detecting malicious communications from their TTL values [3], and in our proposed system, we apply it to packet analysis for detection of malicious communication.

4 Force Threshold Measurement Experiment [13]

4.1 Method

Various studies relating to the difference necessary to distinguish between two presented forces, such as the elucidation by Weinstein and Weber [4, 5] on the relation between reference force and rate of change, have been reported, but the relation for a force-sensing haptic device such as that of the present study has remained unclear.

We therefore investigate the level of recognizable difference in this study, in which the force values are expressed in terms of the Omni "force levels" of 0.0 to 1.0, with 1.0 as the maximum force that can be presented by Omni and 0.0 as the level when no force is presented. The maximum force presented by the Omni is given as 3.0 kg-m/s^2 (3.0 N), and a force level of 1.0 thus represents a force of about 3.0 N.

In operating Omni, the participant grasps the Omni stylus (It's mean pen part) and uses it to compare the size of the forces in the left and right halves of the virtual space on the screen shown in Fig. 2. The reference stimulus is displayed on one side and the

comparison stimulus (4 types) on the other, in random combinations. The comparison stimulus force level is presented in increments of 0.2 in the range 0.0 to 0.8 and the reference stimulus is the center value 0.4. The participant compares the left and right force sizes and chooses between the three choices of "both about the same", "stronger on the left", and "stronger on the right", and the results are analyzed [6].

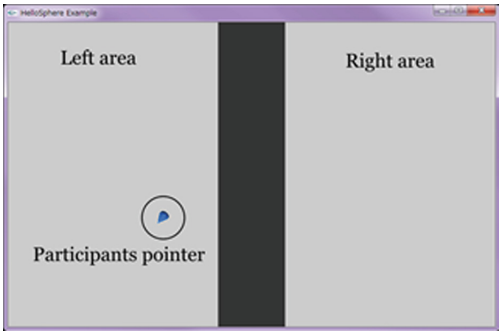

Fig. 2. Force threshold measurement screen.

4.2 Results

Table 1 shows the experimental results for 14 male participants aged 18 to 21, with a as the reference stimulus and xi as the comparison stimulus. Figure 3 shows these results in terms of maximum likelihood, and Table 2 shows the obtained parameter values.

Figure 3 shows a graphical representation of the results with the data points plotted along the horizontal axis for the presented comparison stimulus and along the vertical axis for the probability distribution (determination ratio), and thus the determination probabilities for the parameter values.

Table 1. Force threshold measurement results.

Reference stimulus a: 0.4; comparison stimulus xi in 0.2 increments			
xi	a < xi	a ≈ xi	a > xi
0.0	0	0	14
0.2	0	0	14
0.4	1	7	6
0.6	13	1	0
0.8	14	0	0

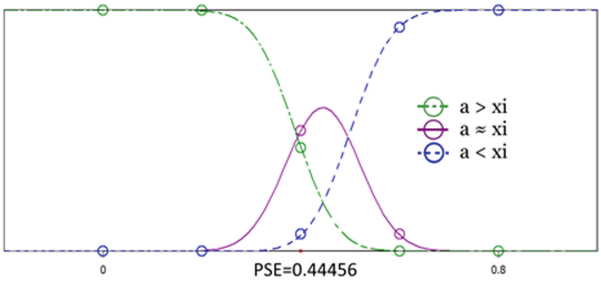

Fig. 3. Result of maximum likelihood method.

Table 2. Force threshold parameters.

μ	σ	Crit.	DL $Z_{0.75}$	Asc. DL μ+ $Z_{0.75}$	Dsc. DL μ− $Z_{0.75}$
0.44	0.067	0.056	0.045	0.485	0.395

μ: mean; σ: standard deviation; Crit: criterion;
DL: difference threshold;
Asc.: ascending; Dsc. descending.

As this shows, nearly all the participants recognized a difference in force size in all cases except where the reference stimulus in Table 1 was presented on both sides. The analysis results indicate the determination criterion c was 0.056, and thus that a difference of 0.2 between two stimuli is sufficient for good discrimination between them, so we used increments of 0.2 in the constructed system and related the resulting values to the differences in traffic amount for their recognition.

5 System Prototype

5.1 Calculation of Traffic Volume

The system prototype presents a reaction force corresponding to the results of the IP packet analysis and traffic amount, with the packet amount calculated as follows.

We measure the total outflow n (bytes) of the traffic each minute and calculate the mean traffic amount in that minute. We first investigate the IP address recorded in the IP header of the arriving packet. If this address has already been recorded, then we add the packet length in the IP header to the packet buffer of that IP address. If not, then we record this new IP address and initialize the packet buffer with the packet length. This operation is repeated for 1 min. At the end of 1 min, we add the previous packet buffer mean multiplied by the number of data n and the newest packet buffer, divide by n + 1, and take the result as the new average. We next divide the newest packet buffer by the average and change the presented force by the resulting value. The calculation algorithm is as follows.

Sn: sum from 0 to n; $\bar{x}$: mean from 0 to n; $\bar{x}_{n+1}$: mean from 0 to n + 1.

$$S_n = \sum_0^n x_i$$

$$\bar{x}_n = \frac{S_n}{n} = \frac{\sum_0^n x_i}{n}$$

$$\bar{x}_{n+1} = \frac{S_{n+1}}{n+1} = \frac{\sum_0^n x_i + x_{n+1}}{n+1} = \frac{n \times \bar{x}_n + x_{n+1}}{n+1}$$

5.2 Traffic Display System

The operating screen in the prototype system is as shown in Fig. 4, with the personal computer that controls the system being represented at the center and lines representing LANs extending radially from the computer with the IP address of each nearby. The blue triangular pyramid is the pointer, which is moved freely by the user. In this figure, the system is performing automatic packet capture.

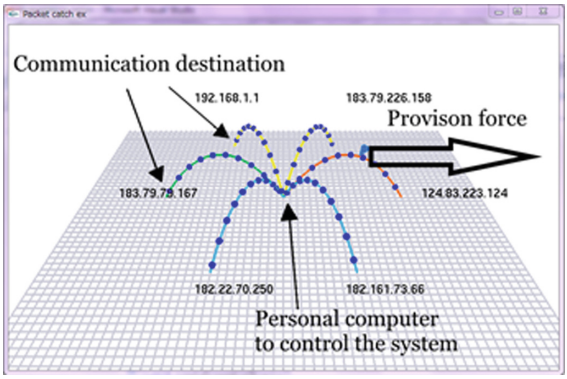

Fig. 4. Prototype system operating screen. (Color figure online)

The traffic amount on each line can be sensed by the user as a force by touching the line with the Omni stylus while depressing the stylus button. The line color changes from blue through yellow to red with increasing traffic volume, thus reflecting the emotional color meaning [7] in a manner analogous to temperature change. The system shows the line of a packet with 30 or more hops in red, as an anomalous packet. The relation between the assessed traffic amount, the line color, and the force level is shown in Table 3.

The system algorithm flow begins with packet capture, followed by reading of the IP address, packet length, and TTL from the IP header and then by determination of whether that IP address has already been recorded from the dataset holding recorded IP addresses, total traffic amounts, and hop numbers. If it has, then that packet length is added to the total traffic amount, and if it has not, then a new dataset is constructed, the IP address is recorded, and the packet length is added to the total traffic amount. The hop number is next calculated from the TTL value and if it exceeds 30, then 1 is added to the hop number in the dataset. This is repeated, and at 1 min, the top 6 cases of total traffic amount are displayed in descending order. If the hop number of the dataset is 1 or more, then the hop number of 30 or more in Table 3 is applied. The dataset is written to the text file as a log and then initialized. The process is repeated thereafter.

Table 3. Traffic assessment, line color, and force level.

Assessment	Line color	Force level
≤ 10	Blue	0.0
11–20	Pale blue	0.2
21–30	Green	0.4
31–40	Yellow	0.6
41–50	Orange	0.8
≥ 51	Red	1.0
Hops: ≥ 30	Red	1.0

In the next step, the system displays the traffic amounts on a two-dimensional day–time plane as shown in Fig. 5, with the colors in the figure changing with the traffic amount as shown in Table 3. This change to an algorithm with a two-dimensional day–time display facilitates the detection of anomalous behavior by comparison with the most recent traffic amount.

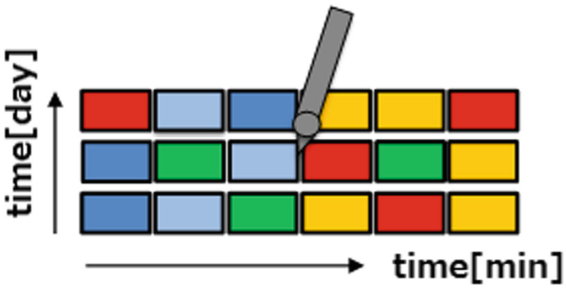

Fig. 5. Day–time plane.

5.3 Improvement and Attendant Modification

Figure 6 shows the improved version of the system operation screen. In this improved version, the computer is again located in the center of the screen, but blocks are used to show the traffic amount at given times. The blue cone is the pointer moved by the user. The IP address is shown near the trailing edge of the block.

In this way, each block provides a summary of the traffic amount for a given time and enables comparison with the previous and subsequent times, and the user can feel the force of a given block by touching it with the stylus. The block stacking for each day enables comparison of the time on a given day with the same time on the previous two days.

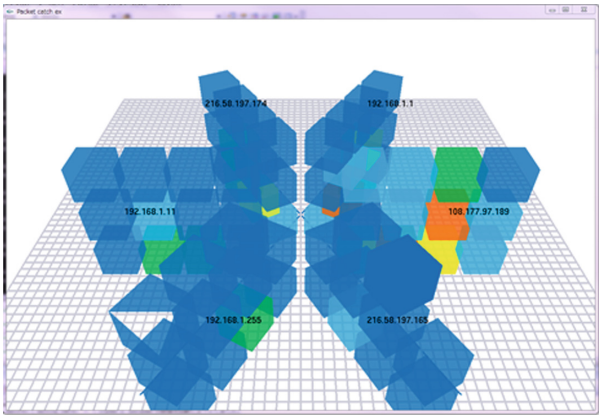

Fig. 6. Improved system operation screen.

The use of blocks instead of lines to present the haptic sensation posed a new problem that required resolution before adoption of the improved version. If the force was presented instantaneously when the stylus touched the block, then in many cases the force was sufficient to cause immediate recoil from the block and disappearance of the presented force, resulting in an instantly vanishing force sensation.

To resolve this problem, we modified the force presentation as weak near the block surface, increasing with proximity to the block center, and reaching full strength for that block near its center. For maximum ease of use, we investigated the optimum proportion of the block for this force change, by measuring the threshold relative to the block proportion.

In this measurement, we had the participant grasp the Omni stylus and use it to touch two blocks in a virtual space, and compared block proportions containing the force change. Figure 7 shows the program execution screen in this experiment. The reference stimulus was presented in the block on one side. In the other block, four comparison stimuli and five reference stimuli were presented at random. In the change portion, changes in presented force ratio of 1/4, 1/3, 1/2, 2/3, and 3/4 between the block surface and the maximum value were compared. The value 1/2 was taken as the

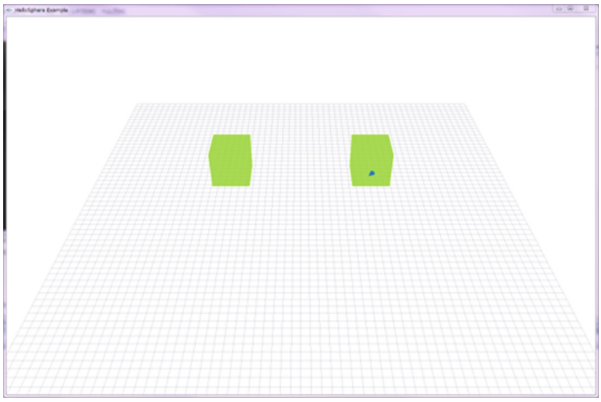

Fig. 7. Screen for execution of the change portion measurement program.

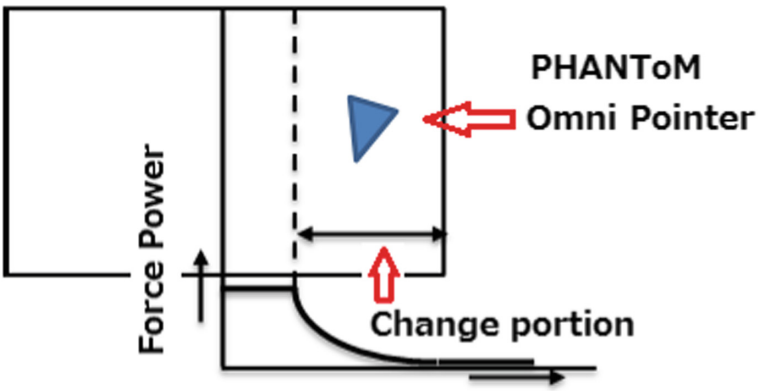

Fig. 8. Relation between block and change portion.

reference stimulus and the others were taken as comparison stimuli. Figure 8 shows the relation between the block and the change portion. We first confirmed that differences in the comparison targets could be perceived and then had each participant compare the left and right changes and select one of the two-choice responses: "the left change is sharper" or "the right change is sharper". The participants were three men in aged 18 to 21 with 5 iterations for each participant.

5.4 Results of Change Portion Threshold Measurement

Table 4 shows the measurement results, in terms of the number of answers given. As in Sect. 4.2, a is the reference stimulus and xi is the comparison stimulus. Figure 9 shows the maximum likelihood derived from these results, with the horizontal axis representing the stimulus strength and the vertical axis representing the number of correct answers.

Table 4. Results of determination of proportion threshold in change portion.

Reference stimulus a = 1/2		
Comparison stimulus xi	xi > a	xi < a
1/4	3	12
1/3	0	15
1/2	5	10
2/3	12	3
3/4	12	3

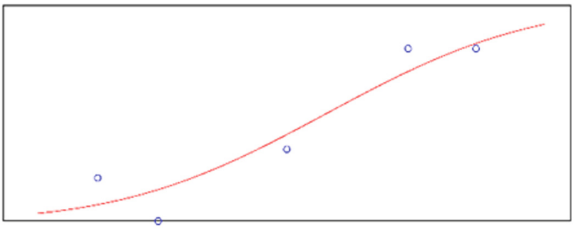

Fig. 9. Result of maximum likelihood method.

The reference stimulus was 1/2 and the point of subjective equality (PSE) was 0.55, thus showing only a slight error in the PSE relative to the reference stimulus. The range of stimuli deemed equivalent to the PSE was 0.41 to 0.69. Stimuli outside this range will therefore be deemed nonequivalent.

5.5 Optimum Decision by Analytic Hierarchy Process (AHP)

The results of this experiment indicate that a difference of 0.2 or more is sufficient for recognition of the difference in the change portion. To select the optimum value, we therefore performed pairwise comparison of differences increasing or decreasing in increments of 0.2 centered on the reference stimulus of 1/2 and thus comprising 0.3, 0.5, and 0.7, with decision by AHP.

Although AHP is generally used for ambiguous decisions, it is also used for decisions on human sensation amount and haptic sensation [8–10], which are considered mainly as ranking of determinations, and it was applied in this experiment in the same light. Two blocks were presented, and the stimuli were compared pairwise. Each participant then responded to the question of which was easier to use and to what degree on a scale of 1 to 9.

The participants were three men (A, B, and C) in aged 18 to 21, and the results shown in Table 4 were obtained. Determination was made by AHP based on the criterion index (CI). A determination is used if the CI is 0.1 or less, or for practical operation, 0.15 or less. We accordingly concluded that the changes can be ranked for each difference of 0.2, and here regarded the application of a value of 0.2 or more as appropriate for ranking sensation amount and decided to perform the design with 0.5 as the difference. Table 5 shows the results.

Table 5. Results of optimum decision by AHP.

Comparison stimulus	Importance		
	A	B	C
0.3	0.098	0.379	0.304
0.5	0.715	0.508	0.575
0.7	0.187	0.113	0.121
C.I.	0.001	0.082	0.109

5.6 Conversion of Force Level

5.6.1 Investigation of Presented Frictional Force

To date, the presented stimuli in experiments have been considered mainly as frictional forces, which might be regarded as not representative of the normal force. The PHANToM operating sensation and observations during the experiment, however, provide experiential evidence of close mutuality between the sensation of force pushing down on the plane of virtual space and the sensation of frictional resistance. In contrast, the experiments to date have indicated that a difference existed between the value settings on the PHANToM when treated as a coefficient of friction and the coefficient of friction actually presented by the PHANToM.

No description relating to this difference is given in the control program OpenHaptics™ toolkit v3.0 reference documents, and there is no detailed specification of parameters of the friction sensation presentation other than that "0 represents no presentation and 1 is the maximum value that can be presented by the machine." To clarify the relation between these set values and the values actually presented, we therefore investigated the coefficient of friction values in the actual presentation by the PHANToM.

5.6.2 Experimental Method

As shown schematically in Fig. 10, we immobilized part of the PHANToM arm, to simulate the PHANToM operation by the participant, by suspending a weight from its stylus and connecting it to a force gauge with kite string (No. 8). The force gauge was used to pull the stylus at a constant speed, thus reproducing the participant's experimental operation of the PHANToM and the resulting frictional force sensation of the participant, as well as to measure the frictional force. The suspended weight itself was 0.2 kg and the total weight including the 0.029 kg of the material used to immobilize the arm and the 0.042 kg of the stylus component on the arm was 0.271 kg.

To determine the relation between the value setting on the PHANToM (the "setting") and the frictional force obtained from the force gauge measurement, we measured the normal force of the weight with the configuration shown in Fig. 10 and calculated the coefficient of friction from that value.

The measurements were made with settings for 0.0 to 1.0 in increments of 0.1 on the PHANToM, with reference to the PHANToM specification of the definition in the Sensable OpenHaptics™ API Reference Manual of 0 as the setting with no force presented and 1 as the maximum possible force presentation of the machine.

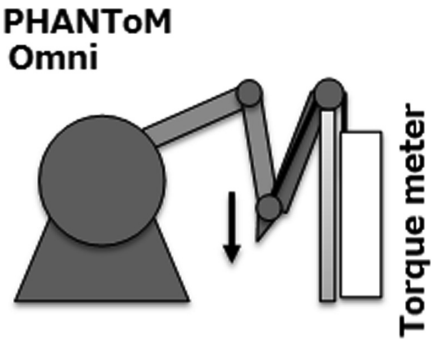

Fig. 10. Configuration in force level conversion experiment.

The measurement was performed 12 times at each setting but the maximum and minimum obtained values were excluded as possible measurement errors, and the results of 10 measurements for each setting were therefore used in the calculation. It was also considered that the minimum measured value of 0.0 might not be correct at the minimum setting at which no force is presented, and that measurement for 1.0, involving the maximum force that can be presented by the PHANToM, might lead to system breakage.

In the calculation from these measurement results, we applied the formula

$$\text{frictional force} = \text{coefficient of friction} \times \text{normal force},$$

to investigate the actual values handled by the haptic device.

5.6.3 Experimental Results

The measurement results are shown in Table 6 and graphically in Fig. 11. As noted above, the setting values are those that can be entered on the PHANToM, and the measurement results are the mean values of the measurements with the force gauge excluding the maximum and minimum measured values, and the coefficient of friction was calculated from the resulting values and the suspended weight.

Table 6. Frictional forces presented by the PHANToM.

Setting	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Mean of measured values [N]	0.1650	0.1450	0.1213	0.1163	0.0963	0.0838	0.0738	0.0763	0.063
Standard deviation σ	0.0135	0.0175	0.0188	0.0133	0.0142	0.0072	0.0258	0.0503	0.0249
Calculated coefficient of friction	0.06189	0.05439	0.04548	0.04360	0.03610	0.03141	0.02766	0.02860	0.02344

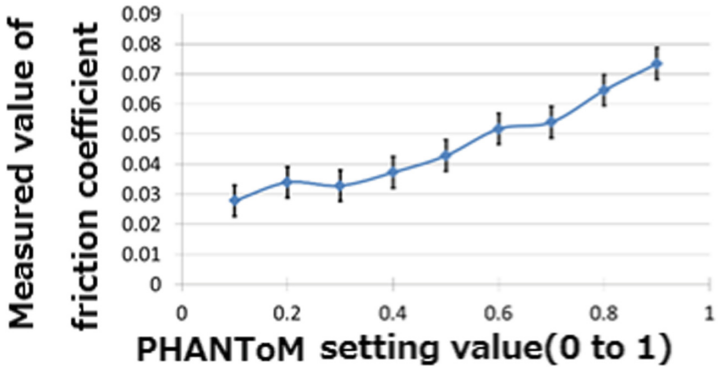

Fig. 11. Relation between settings and coefficients of friction.

These results show that the relation between the settings and the coefficients of friction is not linear, and we therefore applied approximation (Fig. 12).

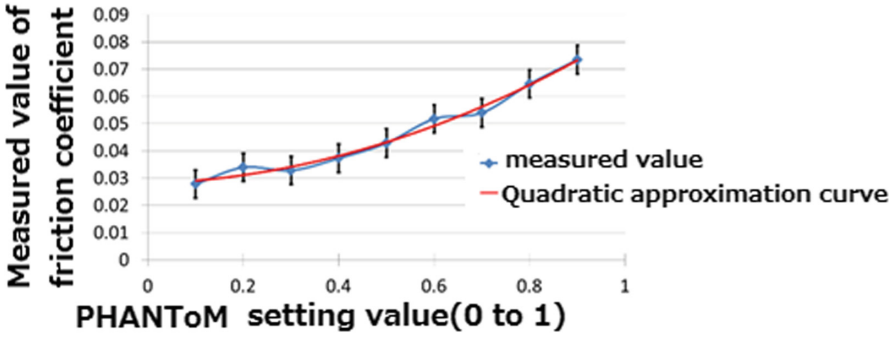

Fig. 12. Measurement results and quadratic approximation.

Comparison of the approximation and the measurement results show that the approximation curve fits within the range of measurement error, and thus that the approximation is correct. The equation obtained by the approximation is

$$M = 0.042p^2 + 0.0045p + 0.0236, \quad (1)$$

where M represents the value of the coefficient of friction actually presented and p represents the values of 0.0 to 1.0 with the setting entries on the PHANToM. With the experiment described above performed using the parameter F′ defined as the product of p and 7.9, the maximum value that can be presented by the PHANToM DeskTop, the relation $p = F'/7.9$ can therefore be applied to use the actually presented coefficient of friction as the variable.

5.6.4 Investigation

These experimental results show the relation between frictional sensation in the presentation by the PHANToM DeskTop and the setting in the application. From this, by applying Eq. (1) to the experimental results therefore enables investigation of the human sensation induced with the haptic device using values closer to the actual values. It is difficult for humans to recognize changes in frictional sensation with stimulus increments of 0.4 N or less, which by Eq. (1) corresponds to an actual coefficient of friction of about 0.23. Observations during the experiment showed that the force (normal force) applied by the humans pushing on the surface to perceive a frictional sensation using the haptic device was about 1.5 N. This is near the pen pressure generally applied by healthy individuals, and thus that no large difference from pen pressure occurs even when pressing on a virtual space. Taken together, the results showed that a change of 0.3 N or more is necessary for sensing a change in the frictional force presented by the haptic device.

5.6.5 Application to the Proposed System

To facilitate the use of various haptic devices, we converted the presented force to newtons (N) in the MKS unit system, a common unit system. In the experimental method for this purpose, the device [11] was installed in the configuration shown schematically in Fig. 10, force changes were applied in increments of 0.1 between 0 to 1.0 by the Omni control program variable with 12 measurement repetitions, and the mean value was calculated from the measurement results exclusive of the maximum and the minimum.

The experimental results are shown in Fig. 13. In the third-order approximation, the error remained within 5%, suggesting that the haptic force presented by the Omni indirect drive is nonlinear due to displacements generated by the dive, and thus follows a characteristic curve resembling a third-order approximation.

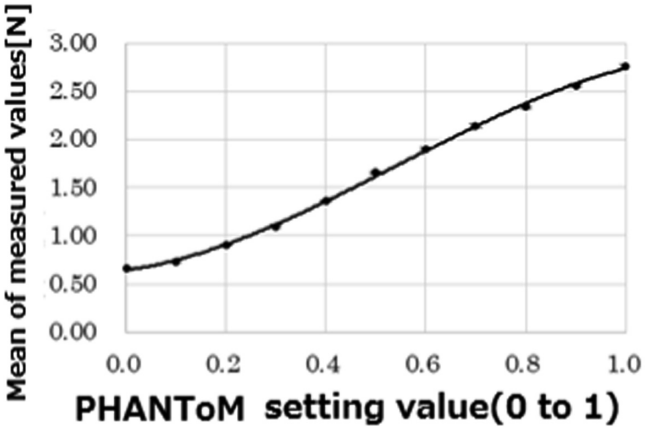

Fig. 13. Relation between force and force level.

6 System Specification

We applied these experimental results to redefine the system specification. As modified thereby, the system performs packet capture automatically after start-up, writes the top 6 cases of IP addresses having the largest traffic amount in 1 min in the order of descending traffic amount, and draws blocks colored in correspondence with those amounts. When the user touches a block with the pointer, it presents the force corresponding to that traffic amount. The process is then repeated. The relation between the traffic amount and the presented force is determined by comparison with the average from the time of system start-up and the size of the divergence, as in the Nakajima et al. method of network anomaly detection [12]. Table 7 shows the relation between the traffic amount, block color, and force level after modification of the prototype system.

The modifications of the prototype system are essentially as follows.

- Blocks are used instead of lines to present force.
- A new block is added once every 1 min instead of renewing the force represented by each line every 10 s.

With a block used to present the haptic sensation, in touching it, the pointer would become partially or wholly invisible to the user if the block were completely impenetrable. We therefore instead used semi-impenetrable blocks to make the pointer portion behind the block visible as shown in Fig. 14 and thus facilitate understanding of its location by the user.

Table 7. Traffic amount, color, and force level relation.

Cumulative traffic amount average	Color	Force level
<0.25	Blue	0.1
0.25–0.75	Pale blue	0.2
0.75–1.25	Yellow	0.4
1.25–1.75	Orange	0.6
>1.75	Red	0.8

Fig. 14. Before and after penetrability modification.

With the maxim force level of 1.0, which is the maximum possible on Omni, operation could not be performed and it was therefore excluded. A slight force presentation was used for the minimum value, since a minimum value of 0.0 would represent an empty space in which no force can be applied.

A basic reason for use of Omni in this system is that it allows the use of a mouse to zoom and rotate the display in the operation, which facilitates positioning of the reference base on the screen.

7 Conclusions

The results indicate this system can show communication amounts and malicious packets intuitively via Omni and that, with the improved version, it can be effectively used as whitelist filtering software through comparison of communications by their day and time based on human haptic sensing. The method can be used to pass only communications approved in advance and block all others, but as applied here, it can also focus on communication amount and block any communication with an amount exceeding a basic standard. In this way, it is expected to provide a tool for general users of ordinary personal computers to learn the need for security technology and understand its importance, required reliability, and other essential aspects, as well as to aid the search for defensive methods in regard to the Internet.

The experimental results also showed that third-order curve force levels and the MKS unit system can be adopted in the system and thus that we have been able to construct a system that can run with haptic devices other than Omni and therefore used more widely.

We plan to add functions enabling classification not only by IP address but also by port number, and switching between transmission and receiving amounts.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 17K00504.

References

1. Kaneko, H.: Integrated analysis malware from geographic visualization. In: Proceedings of the Computer Symposium 2011 (2011). Inf. Process. Soc. Jpn. **2011**(3), 179–184 (2011). (in Japanese)
2. NICT: NIRVANA KAI. http://www.nict.go.jp/cyber/research.html, October 2017
3. Yamada, R., Tobe, K., Goto, S.: Discriminating malicious packets using TTL in the IP header. IEICE Technical report, vol. 111, no. 469, pp. 235–240, March 2012. (in Japanese)
4. Weber, E.H., Ross, H.E., Murray D.J.: Trans.: De subtilitate tactu, pp. 19–135. Academic Press for Experimental Psychology Society (1834/1978)
5. Weinstein, S., Kenshalo, D.R.: Intensive and extensive aspects of tactile sensitivity as function of body part, sex, and laterality, pp. 195–222 (1968)
6. Okamoto, Y.: Keiryou Shinrigaku. Baifukan Press, Tokyo (2006). (in Japanese)
7. Inami, M., Kuriyama, T., Abee, M.: Motion and color(3). Dep. Bull. Pap. Shimane Univ. **28**(3), 35–50 (1994). (in Japanese)
8. Ishihara, M., Shirataki, J.: Relationship of acoustic sound to distance. Trans. Jpn. Soc. Mech. Eng. C **60**(580), 4211–4215 (1994). (in Japanese)
9. Saaty, T.L.: The Analytic Hierarchy Process, pp. 17–21. McGraw Hill, New York City (1980)

10. Ishihara, M., Negishi, N.: Effect of feedback force delays on the operation of haptic displays. IEEJ Trans. Electr. Electron. Eng. **3**(1), 151–153 (2005). https://doi.org/10.1002/tee.20247
11. Harada, H., Rahok., S.A., Suzuki, S., Ishihara, M.: Experiment of representing roughness with haptic devices: In: JSME Conference on Robotics and Mechatronics (Robomec) 2015, 2A2-X04(1-2) (2015). (in Japanese)
12. Nakajima, A., Shigematsu, K., Mizutani, M., Takeda, K., Murai, J.: Implementation of network sonification system. In: Proceedings of the National Conference on Information Processing Society of Japan, pp. 485–486 (2011). (in Japanese)
13. Ishihara, M., Kanayama, T.: Visualization technologies of information security support system using haptic devices. In: Tryfonas, T. (ed.) HAS 2017. LNCS, vol. 10292, pp. 329–338. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58460-7_23

Using Scenarios to Validate Requirements Through the Use of Eye-Tracking in Prototyping

Tia Larsen-Calcano[1], Omar Ochoa[2(✉)], and Richard Simonson[1]

[1] Department of Human Factors Psychology, Norflok, USA
{larsenct, simonsrl}@my.erau.edu
[2] Department of Electrical, Computer, Software and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
ochoao@erau.edu

Abstract. Research has shown that eliciting and capturing the correct behavior of systems reduces the number of defects that a system contains. A requirements engineer will model the functions of the system to gain a comprehensive understanding of the system in question. Engineers must verify the model for correctness by either having another engineer review it or build a prototype and validate with a stakeholder. However, research has shown that this form of verification can be ineffective because looking at an existing model can be suggestive and stump the development of new ideas. This paper provides an automated technique that can be used as an unbiased review of use case scenarios. Using the prototype and a scenario, a stakeholder can be guided through the use case scenario demonstrating where they expect to find the next step. Instead of measuring how many mistakes the individual makes or how long it takes them to complete the scenario, the purpose of going through the scenario with the prototype can be to capture where the individual is actually looking. Analysis of that information can be used to identify missing requirements such as interaction steps that should have alternative sequences, or determining problems with the flow of actions. The proposed approach can be used to both facilitate the further elicitation of requirements using eye-tracking techniques and validate prototypes in tandem with use case scenarios.

Keywords: Requirements · Validation · Eye-tracking

1 Introduction

Before an idea becomes a system or even a list of requirement specifications, an acting requirements engineer captures an initial description as set forth by the stakeholder. A stakeholder is defined as an individual that either affects or affects the system success [1]. Starting with the initial set of raw requirements, the engineer begins the long process of requirements analysis by modeling eliciting further requirements. It is important to emphasize the importance of this step in the development process, as a system cannot just be built with just an initial description. Implementing a system without a thorough analysis of the priorities, wants and needs of the stakeholder, will

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 94–104, 2018.
https://doi.org/10.1007/978-3-319-91467-1_8

lead to the development team spending time building an incorrect system that the stakeholder will not accept.

Certain challenging issues are associated with requirements engineering. Requirements engineers may not have the necessary domain knowledge to understand what the stakeholder wants. Stakeholders may make implicit domain assumptions and fail to communicate those to the requirements engineer. The stakeholder may not fully understand the desired behavior of the system. If there are multiple stakeholders, an engineer may run into the problem of conflicting priorities.

Research has shown that it is more costly to correct defects later in the lifecycle of a system than eliciting and capturing the correct requirements earlier [2]. Before writing the requirements, a requirements engineer will model the behavior of the system to lead to a comprehensive understanding of the system in question. These models can come in all many different forms including a use case diagram, state chart, data flow diagram, etc., all serving their own role in defining different perspectives of a system. Currently, the only research done with eye tracking measures and UML diagrams has been a couple of studies using people at different levels of experience and measuring their workload as they interact with an existing diagram. These studies have showed that the layout of the UML diagrams can have an impact on readability thus making it less time consuming and cheaper to work with [3].

After creating a model of the system, engineers must verify the model for correctness by either having another engineer review it or build a prototype and validate with a stakeholder. This stage of development is where it is important for models to be in a readable layout so that other engineers can easily understand what is being captured. However, research has shown that this form of verification can be ineffective because looking at an existing model can be suggestive and stump the development of new ideas as suggestive memory has been proven especially in the field of law [4]. Therefore, the use of an objective measure can be more comprehensive. Thus, the purpose of this study is to define a new approach to verify use case scenarios in a way that is objective (i.e. non-human and automatic) and a technique that has proven to provide useful non-subjective information to user experience experts has been eye tracking.

This paper provides a technique that can be used as an unbiased review of use case scenarios. Once a requirements engineer creates a use case diagram modelling the relationships of different entities that interact with a system, the use case scenarios are built detailing every step that entities can make and the response the system gives. Based off the scenarios and information from other models, a prototype is built with low or high fidelity. Using the prototype and a scenario, a stakeholder or non-member of the design team can be guided through the use case model demonstrating where they expect to find the next step. Instead of measuring how many mistakes the individual makes or how long it takes them to complete the scenario, the purpose of going through the scenario with the prototype can be to capture where the individual is actually looking. Analysis of that information can be used to further elicit requirements such as identifying steps that should have alternative sequences or determining problems with the flow of the steps or major problems with the design of the prototype from the very beginning. The proposed approach can be used to both facilitate the further elicitation of requirements using eye-tracking techniques and validate prototypes in tandem with use case scenarios.

2 Background

Eye tracking is a method that has been around for over 100 years. The beginning of eye tracking started with a standard camera with users going frame by frame to document the movements of the eye. Nowadays, eye trackers have adapted to capture a high rate of frames and automatically analyze data and present it in charts.

2.1 Eye Tracker Data Collection

An eye tracker collects information about where a participant is looking in an area by tracking the scanpath, movement and fixation. A scanpath for the purpose of this study is defined as the sequence of eye movements across a page, and a fixation is defined as "when a gaze remains within a small area for a given time" [5]. The method by which it collects this information depends on the type of eye tracker used, but the information that it collects is standard across all devices. The raw data that an eye tracker collects consists of the eye movement of the participant as well as the size of the pupil. The eye movement is translated to an x/y coordinate on the screen while the fidelity of the eye trackers determines how much movement is collected [6]. The higher the fidelity of the eye tracker the higher the sample rate by which the eye tracker collects. Typical sampling rate ranges from 30 Hz – 250 Hz although the range of 50 Hz to 60 Hz is the norm for usability studies [6]. It is important to determine which sampling rate will be used for a study, as higher sampling rates will out put more data and require more data reduction when reviewing participant data. The data reduction and post processing of the data depends on the eye tracker used, data can be visualized as heat or focus maps, gaze plots, and aggregated gaze plots. Heat maps compile groups of eye tracking gaze plots and visualize them into a visual light spectrum from red, representing a heavy traffic gaze area, to blue, representing a light traffic gaze area [7]. Gaze plots are a visualization of a scanpath for a user [8], and aggregated gaze plots are scan patterns aggregated from many users viewing the same visual stimulus [9]. Each visualization provides specific in-formation about a scanpath and fixation. However, these visual-izations are mean-ingless unless backed with a specific research question.

2.2 Usability Studies Usage

Eye tracking is a popular method to supplement usability testing, however it should only be used when appropriate. The Neilson Norman group recommends that one should only use eye tracking after about one hundred rounds of regular usability testing [10]. The reasoning behind this belief is that eye tracking requires specific research questions to gather meaningful data and requires supplemental usability testing methods to make sense of the gaze plot data. In addition, the financial, labor [11], and time obligations are significantly greater than the majority of the usability testing methods. The two most popular usability testing methods to supplement eye tracking in usability studies are think aloud protocols, where a participant is asked to "verbalize whatever crosses their mind during the task performance" [12].

The approach presented in the paper uses the video-based combined corneal reflection eye tracking method supplemented by a retrospective think aloud protocol.

The video-based combined corneal reflection method is one of the most widely used because it offers point of regard tracking, where the eye movement is distinguishable from the head movement of a participant, without any intrusive head stabilizing techniques. The approach was supplemented with a retrospective interview to prevent any confirmation bias that may have occurred during data analysis as well to extrapolate possible reasons behind the participants thought process and actions.

2.3 Use Case Model and Scenarios

A use case model is a representation of the desired behavior in a system the interactions with the surrounding environment [13]. Use case models are composed of *actors* and *use cases*, surrounded by a system boundary drawn as a box around the system. *Use cases* represent an abstract view of the system and its interactions with the external entities. These external entities, which can be human users, organizations or other systems, are depicted as *actors*. Lines associated between *actors* and *use cases* represent interactions of value to the *actors* by the system. The use case model facilitates the elicitation of functional requirements by providing an abstraction of the main uses of a software system and the *actor* that will interact with the system, enabling a way for requirement engineers to analyze and identify the functional needs. This assists in the analysis of what information these actors will provide or receive from the system thus highlighting requirements about interfaces. Additional relationships are used such as *includes* and *extends* to model relationship between the *use cases*, which signify use case inclusion and optional extension to use cases, respectively.

A *scenario* describes the specific exchange of information as a flow of interactions between the actors associated with the use case and the system. The steps include alternative flows that can be taken while executing the main flow. The in-depth analysis of a use case happens when writing a scenario, the use case and actor provide a general abstraction of what the actor comes to the system to do, but the requirements of what the system must rise up from writing the scenario text. Each *use case* can have one or more *scenarios* as documentation and because they are simple text, stakeholders can use the *scenarios* to validate the elicited behavior. In addition, developers can use *scenarios* to build other analysis models, create UI prototypes, walkthrough design and implementations, and generate test cases.

3 Related Work

In addition to all the current uses of eye tracking as defined above, eye tracking is being applied across multiple disciplines to find all the different applications. For example, a lab just recently, in 2016, proposed a relationship between baseline pupil size and intelligence along with other labs that have done research with applications that include an engineer's understanding of a technical drawing [14, 15]. Eye tracking is becoming more common even though it is not the most reliable measure mainly because eye tracking is mainly nonintrusive. Most current measures used for things like situational awareness and workload are measures that interrupt the task and require verbal or written information like the Situational Awareness Global Assessment Technique (SAGAT) and

National Aeronautics and Space Administration Task Load Index (NASA-TLX) [16, 17]. There have been a couple studies done with eye tracking and UML models. However, unlike the focus of this study where the design is specific to a specific model, the previous studies were fixed on the overall design qualities to reduce workload [18–21].

3.1 Model Validation

Validating the models created is an important step in the process of defining the correct requirements for a system. Models capture the current understanding of the expected behavior of the system, thus a model needs to be as correct as possible to mitigate specifying incorrect requirements. Traditional approaches in the validation of models include the application of reviews, specifically two types of reviews can be used, inspections and walkthroughs. A review aims at examining the model created, looking for flaws, inconsistencies, or omissions that can be corrected, and usually done by requirements engineers and stakeholders [22]. An inspection is a formal review that follows a well-defined set of steps, in which each participant is assigned a specific role in the review [23]. A walkthrough is an informal type of review, i.e. it does not follow a pre-define process and participants do not have a specific role [24]. These techniques have been shown to be effective at detecting defects in requirements specifications [25–28].

4 Approach

As stated in Sect. 2, use case diagrams are a specific type of UML model that maps out the different actors and their interactions with the system. The visualization of these interactions contributes to the development of the use case scenarios by defining functionality that needs to be accessible for a user to be satisfied with the use. Use case scenarios are what drives the development of a prototype especially in a computer system and are what the design team goes through when demonstrating a prototype to a stakeholder. For this study, the scenario and prototype are based off a system that was defined and designed for a graduate-level requirements class modeled and specified. The students were offered a description for the desired program named Sharing and Discovering Semantic Use Case Scenarios (SADSUCS); below is an excerpt of that description:

> "SADSUCS is a research project that requires a web-based interface that will allow teammates to efficiently share and discover use case diagrams and scenarios. SADSUCS must be a user-friendly tool to encourage its adoption. Users should be able to create models of systems following a common use case diagram syntax. The system should generate "scenario templates" to be populated with missing information by the user. It is envisioned that the creation of the models will be done by the users' dragging and dropping components in and out of a virtual workspace. The system must store projects in the cloud, along with each project's history of most recent snapshots, each saved every time the user clicks 'save' or every 5 minutes if the model has changed. The app will allow users to undo their actions."

The use case scenario and prototype were derived from those completed in the class. A sample scenario used for the study is shown below:

```
Use Case: View Public Use Case Diagram
Description: A user wishes to find and view a public use
case diagram. The user is not part of the project.
Actors: User, Cloud Management Database
Precondition: The user has an account with the system.
The use case diagram is not private and can be viewed.
Trigger Condition: The user has launched the application.
Steps:
1.  The system displays the login page prompting the user
for a username and a password.
2.  The user enters a username and a password and selects
to login.
3.  The system successfully verifies the entered password
for the given username by comparing it with the password
stored in the Cloud Management Database (ALT 1).
4.  The system displays popular diagrams and an option to
search for use cases.
5.  The user enters use case keyword information in the
search field and selects the search option.
6.  The system queries the Cloud Management Database us-
ing the keywords provided.
7.  The Cloud Management Database returns a list of
matching use cases (ALT 2).
8.  The system displays the list of matching use cases.
9.  The user selects a use case diagram to view from the
list.
10. The system displays the selected use case diagram.
11. The user views and navigates the use case diagram.
12. End of use case.

 ALT 1: Step 3: The system fails to verify the entered
 username and password.
 Step 3.1: The system displays an error message stating
 that the username and password were not recognized.
 Step 3.2: Return to step 1.

 ALT 2: Step 7: The Cloud Management Database returns no
 matching use cases.
 Step 7.1: The system displays a message stating that no
 use cases matched the criteria.
 Step 7.2: Return to step 4.
```

The prototype was built using the elicited use case scenario using the wireframing program, Axure [29]. The program used in this study was derived from the ones built in the class. Sample screenshots of the prototype can be found in the section below accompanied by the gaze plot data.

4.1 Eye-Tracking Study Setup

The individual selected to part of this study was an undergraduate female between the ages of 18 to 25 that did not require the use on contacts or eyeglasses. The participant sat in front of the Tobii eye tracking equipped computer at Embry-Riddle Aeronautical University's Usability Lab. After explaining the purpose of the pilot study, to validate a framework using the eye tracking information, the eye tracker was calibrated to the participant. The participant was given the task to find a public UML diagram and allowed to navigate the prototype with little guidance. Guidance and answers were provided when the participant did not know how to continue or had a question about how to proceed; however, the participant was encouraged to navigate the system in self-guided approach.

5 Results and Observations

The eye-tracking test combined with the retrospective think aloud provided insight into what the participant was thinking and doing during their completion of the task. The home page provides little information about the site, and by reviewing the scanpath of the participant, it becomes apparent that the main area of interest is the paragraph text, contact, and about feature. The reasoning behind this scanpath is explained by the participant in the interview, where they stated to "not know what the purpose of the site is" and did not know how to continue. The task of logging in to the site is then considered a failure, as they needed assistance in continuing.

The eye tracking analysis of the dashboard page is broken up into two phases. The first is the participants natural scan for information in scan plots one through ten, and a second phase of when the participant asked to be reminded of the task in scan plots eleven through twenty. In the interview, the participant noted that they were confused as "nothing on the page said diagram." This was deemed as a failure of the task to find public use diagrams.

The public projects page has a strong indication of an area of interest in the use cases panel. Seventeen out of the twenty-three (74%) total gaze points are focused in this area (see Fig. 1). When asked in the retrospective interview why they were focused on the area they responded with the fact that they were told to search for a use case diagram, however after being unable to interact with any of them they the participant decided to move to the search function. This is considered a pass of the task to search for a public use diagram, and a failure on part of the task for not being specific enough.

The final page the eye tracker captured was the use case diagram associated with this task. Again, an obvious area of interest was identified and is located in the use case diagram where twenty-eight of the thirty-seven (76%) of the gaze points were located.

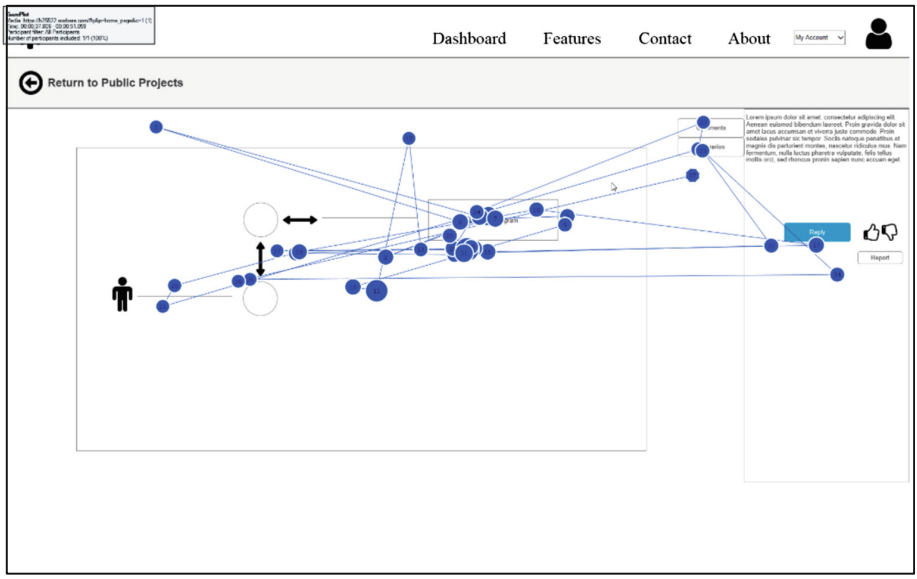

Fig. 1. Use case panel

In the interview, the participant suggested they were trying to figure out if this was the end of the task or if they had to continue through the site.

5.1 Modification Based off Results

Using the information gathered from the eye tracker, the use case scenario can be revised to provide a more comprehensive use case scenario. In effect, this was a new alternative way to elicit requirements about the system. For example, the screenshot below was part of the prototype that was used along with an overlay of the gazeplot gathered by the participant. This specific screen involves steps 4 and 5 as follows (Fig. 2):

```
4.  The system displays popular diagrams and an option to
search for use cases.
5.  The user enters use case keyword information in the
search field and selects the search option.
```

Using the data from the gazeplot above, after the system displays the diagrams, the user looks at the different use cases, the dashboard option, the create a new project option, and browsing the public projects option. Using this information, there should be 3 different alternative use cases that should be added to step 5 as followed:

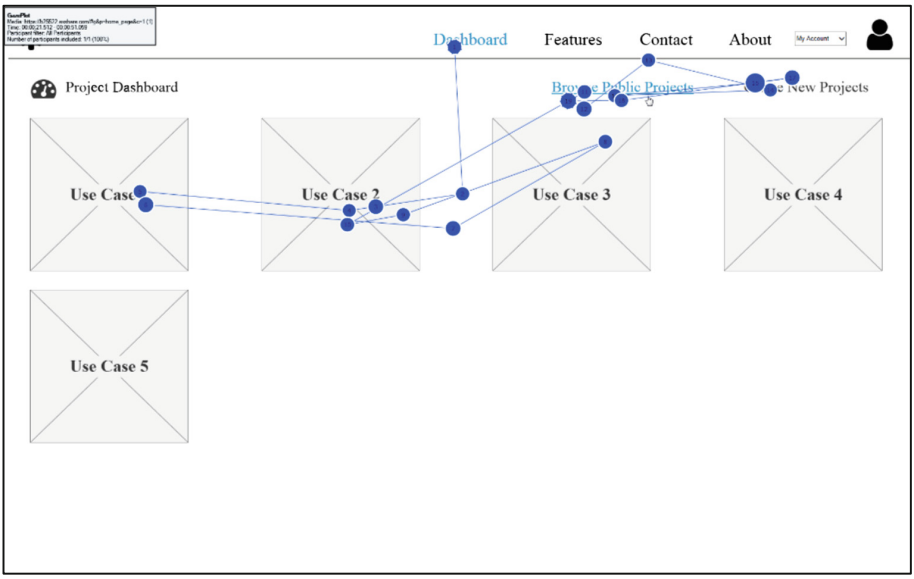

Fig. 2. User dashboard

```
ALT 1: Step 5: User selects a use case.
Step 5.1: Continue with step 10.
ALT 2: Step 5: User selects the create a new project op-
tion.
Step 5.1: The system displays a prompt for the user to
enter information relevant to a new project.
Step 5.2: The user enters information in the prompt.
Step 5.3: The user selects a submit option.
Step 5.4: Continue with step 10.
```

As a result, the proposed approach was used to further elicit requirements using eye-tracking techniques. The participant acted as a validation for the scenario, which was implemented indirectly by the prototype. With the data gathered from the eye-tracker, this technique was successfully used to modify the use case scenario and provide requirement validation on the modeled system, identifying missing functionality. The initial results show promise that this technique can be used in use case scenario validating through eye tracking of prototypes.

6 Future Work

The purpose of this study was to provide a description of a new approach to validate requirements via scenarios and conduct a pilot study. It is important to note that only one participant was guided through the prototype and issues can be found with the data

that was collected. For example, since there was only one data set, things like losses of calibration could not be accounted for. Thus, more research should to develop a framework or set of standards that a design team could use with ease without understanding completely the concept of eye tracking and user-based testing.

Mouse tracking is another usability testing method that can be used both independently and supplementary to eye tracking. Research indicates that a participant's gaze leads the movement of their mouse movements [30], and it is a common practice to use mouse tracking when eye tracking is not a viable option. These methods are most useful when used in combination, as in usability the click and clickstream of a user's navigation through an interface [31]. The incorporation of mouse tracking into the proposed approach could provide more accurate information, however it might provide an average user with too much information and thus possibly making it.

References

1. Freeman, R.E.: Strategic management: a stakeholder theory. J. Manage. Stud. **39**(1), 1–21 (1984)
2. Nuseibeh, B., Easterbrook, S.: Requirements engineering: a roadmap. In: Proceedings of the Conference on the Future of Software Engineering. ACM, Limerick (2000)
3. Eichelberger, H., Schmid, K.: Guidelines on the aesthetic quality of UML class diagrams. Inf. Softw. Technol. **51**(12), 1686–1698 (2009)
4. Brigham, J.C., Maass, A., Snyder, L.D., Spaulding, K.: Accuracy of eyewitness identification in a field setting. J. Pers. Soc. Psychol. **42**(4), 673 (1982)
5. Harezlak, K.: Eye movement dynamics during imposed fixations. J. Inf. Sci. **384**, 249–262 (2017)
6. Goldberg, J.H., Wichansky, A.M.: Eye tracking in usability evaluation: a practitioner's guide. In: Hyönä, J., Radach, R., Deubel, H. (eds.) The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research. North-Holland, Amsterdam (2003)
7. Drusch, G., Bastien, J.M., Dinet, J.: From gaze plots to eye fixation patterns using a clustering method based on Hausdorff distances. In: Proceedings of the 23rd Conference on l'Interaction Homme-Machine. ACM, Sophia Antipolis (2011)
8. Chen, L., Pearl, P.: Users' eye gaze pattern in organization-based recommender interfaces. In: Proceedings of the 16th International Conference on Intelligent User Interfaces. ACM, Palo Alto (2011)
9. Hembrooke, H., Feusner, M., Gay, G.: Averaging scan patterns and what they can tell us. In: Proceedings of the 2006 Symposium on Eye Tracking Research & Applications. ACM, San Diego (2006)
10. Pernice, K., Nielson, J.: How to Conduct Eyetracking Studies. Nielson Norman Group, Fremont (2009)
11. Khachatryan, H., Rihn, A.L.: Eye-tracking methodology and applications in consumer research. IFAS Extension, University of Florida (2014)
12. Gambier, Y., Doorslaer, L.V.: Handbook of Translation Studies. John Benjamins Publishing Company, Amsterdam/Philadelphia (1984)
13. Cockburn, A.: Writing Effective Use Cases, 1st edn. Addison-Wesley Professional, Boston (2000)

14. Matthiesen, S., Meboldt, M., Ruckpaul, A., Mussgnug, M.: Eye tracking, a method for engineering design research on engineers' behavior while analyzing technical systems. In: Proceedings of the International Conference on Engineering Design, pp. 277–286 (2013)
15. Tsukahara, J.S., Harrison, T.L., Engle, R.W.: The relationship between baseline pupil size and intelligence. Cogn. Psychol. **91**, 109–123 (2016)
16. Endsley, M.R.: Situation awareness global assessment technique (SAGAT). In: Aerospace Electronics Conference, pp. 789–795 (1988)
17. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in Psychology, pp. 139–183 (1988)
18. Bednarik, R., Tukiainen, M.: An eye-tracking methodology for characterizing program comprehension processes. In: Proceedings of the 2006 Symposium on Eye Tracking Research & Applications (2006)
19. Sharif, B., Maletic, J.I.: An eye tracking study on the effects of layout in understanding the role of design patterns. In: IEEE International Conference on Software Maintenance, pp. 1–10 (2010)
20. Sharif, B., Salem, W.: On the use of eye tracking in software traceability, pp. 67–70 (2011)
21. Yusuf, S., Kagdi, H., Maletic, J.I.: Assessing the comprehension of UML class diagrams via eye tracking. In: IEEE International Conference on Program Comprehension, pp. 113–122 (2007)
22. Sommerville, I.: Software Engineering, 9th edn. Addison-Wesley, Boston (1996)
23. Brown, N.: High-level best practices-what hot companies are doing to stay ahead and how DoD programs can benefit. Crosstalk (1999)
24. Melo, W., Shull, F., Travassos, G.H.: Software review guidelines. COPPE/UFRJ Systems Engineering and Computer Science Program Technical report ES-556/01 (2001)
25. Boehm, B.: Verifying and validation software requirements and design specifications. IEEE Softw. **1**(1), 75–88 (1984)
26. Potts, C., Takahashi, K., Anton, A.I.: Inquiry-based requirements analysis. IEEE Softw. **11** (2), 21–32 (1994)
27. Caroll, J.M.: Scenario-based design: envisioning work and technology in system development, 1st edn. Wiley, New York (1995)
28. Hsia, P., Samuel, J., Gao, J., Kung, D., Toyoshima, Y., Chen, C.: Formal approach to scenario analysis. IEEE Softw. **11**(2), 33–41 (1994)
29. Axure Homepage. https://www.axure.com/
30. Liebling, D.J., Dumais, S.T.: Gaze and mouse coordination in everyday work. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publicaiton, pp. 1141–1150 (2014)
31. Mao, J., Liu, Y., Zhang, M., Ma, S.: Estimating credibility of user clicks with mouse movement and eye-tracking information. In: Zong, C., Nie, J.Y., Zhao, D., Feng, Y. (eds.) Natural Language Processing and Chinese Computing. Communications in Computer and Information Science, vol. 496, pp. 263–274. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45924-9_24

Measuring Focused Attention Using Fixation Inner-Density

Wen Liu, Soussan Djamasbi(✉), Andrew C. Trapp, and Mina Shojaeizadeh

User Experience and Decision Making Research Laboratory, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA
{wliu3, djamasbi, atrapp, minashojaei}@wpi.edu

Abstract. Examining user reactions via the unobtrusive method of eye tracking is becoming increasingly popular in user experience studies. A major focus of this type of research is accurately capturing user attention to stimuli, which is typically established by translating raw eye movement signals into fixations, that is, ocular events characterized by relatively stable gaze over a specific stimulus. Grounded in the argument that inner-density of gaze points within a fixation represents focused attention, a recent study has developed the fixation-inner-density (FID) methodology, which identifies fixations based on the compactness of individual gaze points. In this study we compare the FID filter with a widely used method of fixation identification, namely the I-VT filter. To do so we use a set of measures that investigate the distribution of gaze points at a micro-level, that is, the patterns of individual gaze points within each fixation. Our results show that in general fixations identified by the FID filter are significantly denser and more compact around their fixation center. They are also more likely to have randomly distributed gaze points within the square box that spatially bounds a fixation. Our results also show that fixation duration is significantly different between the two methods. Because fixation is a major unit of analysis in behavioral studies and fixation duration is a major representation of the intensity of attention, awareness, and effort, our results suggest that the FID filter is likely to increase the sensitivity of such eye tracking investigations into behavior.

Keywords: Eye tracking · Fixation identification · Fixation-inner-density
Fixation micro-patterns

1 Introduction

The study of eye movements in user experience research is becoming increasingly popular because eye tracking technology enables capturing the focus of a person's gaze on a visual display at any given time. Human gaze serves as a reliable indicator of attention because it represents effort in maintaining the eyes relatively steady to take foveal snapshots of an object for subsequent processing by the brain [1]. Hence, extracting relatively stable gaze points that are near in both spatial and temporal proximity, that is translating the raw gaze data into fixations, is essential in many eye tracking studies [2, 3]. One primary method for identifying fixations in a stream of raw

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 105–116, 2018.
https://doi.org/10.1007/978-3-319-91467-1_9

eye movement data is the Velocity-Threshold Identification (I-VT) algorithm. The I-VT filter uses a fixed velocity threshold to identify whether individual gaze points qualify as a fixation point, or a saccade point.

Because a fixation is the collection of gaze points that are near to one another in both time and proximity, a denser collection of gaze points within a fixation represents higher level of focused attention, and thus higher level of cognitive processing [4]. Thus, a recent study [5] proposes a new way to group gaze points into fixations based on their inner-density property. Similar to the I-VT filter, this new Fixation Inner-density (FID) filter first uses a velocity threshold to identify a candidate set of gaze points that are slow enough to form a fixation. It then uses optimization-based techniques to identify a densest fixation of gaze points among all candidate points. Identifying fixations using the FID filter naturally eliminates those gaze points that are near to tolerance settings. *How* gaze points are dispersed in a fixation affects fixation metrics such as the duration and center location, and there is evidence that the FID filter reduces the possibility of skewing these metrics [5].

In this paper we translate raw gaze data into fixation using the I-VT and FID filters. We demonstrate that fixations processed by the FID filter are superior in terms of three key fixation *micro-patterns* than those that are processed by the I-VT filter. First, they are *denser*. Second, the extent to which points are dispersed within a fixation is *smaller*. Third, the points within a fixation are more likely to be uniformly distributed. This investigation is important because the compactness and the patterns of distribution of gaze points can directly affect fixation metrics, such as fixation duration and fixation center position, that are commonly used in eye-tracking studies to assess viewing behavior. This study is the first to investigate such fixation *micro-patterns* or properties of the distribution of gaze points within an individual fixation.

2 Background

Raw gaze data is a sequence of (x, y, t) triplets, where (x, y) represents the measured location of user gaze, and t is the time stamp. Common sampling rates in eye trackers range from 30 Hz to 1,000 Hz, and a gaze sequence can easily contain tens of thousands of triplets. Gaze data is often categorized into two common types: fixations and saccades. Fixations are pauses over informative regions during eye movement; in gaze data, a fixation is where gaze point triplets aggregate together. Fixation identification methods cluster those intensive gaze points into fixations to present focused attention and cognitive effort in eye tracking research [4].

One popular fixation identification algorithm is the I-VT filter. It identifies fixations by gaze point *velocity*. If the velocity exceeds the predefined threshold V, the corresponding gaze point is identified as a saccade, otherwise it is categorized as a fixation point. I-VT filter is efficient and practical; however, it has the drawback of ignoring the information about the spatial arrangement of individual gaze points within a distinct fixation. Some fixation metrics can express the distribution of points within a fixation. One such metric is fixation inner-density, which was introduced by [4] and further refined in [5]. Fixation inner-density represents user focus, and [4] has validated that fixation inner-density is correlated with normalized fixation duration and average pupil

dilation variation during fixation. The FID filter uses optimization-based techniques to optimize for inner-density, which means that it selects a set of candidate gaze points that guarantees there is no better set with respect to the objective function of maximizing fixation inner-density. Fixation inner-density improves upon previous fixation identification methods because it combines both the temporal and the spatial aspects of the fixation into a single metric that evaluates the compactness of a fixation.

As the problem of fixation identification is a type of time-series clustering, it shares the commonality that interpreting clustering results is somewhat subjective in nature. Hence, the choice of an appropriate metric will directly affect the formation of the clusters. While density and dispersion properties can be measured in various ways, they are inherently *positively related* to the number of gaze points in a fixation, and *negatively related* to the area occupied by the constituent points. We next discuss some important metrics to evaluate density and dispersion properties within fixations.

3 Methodology

We consider two representative ways of measuring fixation inner-density, both of which are advocated in [5]. Suppose a fixation identification algorithm locates fixations in a gaze data sequence with T gaze points. For any given fixation f, let n_f denotes the count of points inside f, and let i, j be any two points in f. We denote the Euclidean distance between i and j as d_{ij}, the minimum area box that spatially bounds the fixation as A_{sq}, and the minimum area rectangle box that spatially bounds the fixation as A_{rt}. The first density metric (D_1) is the average pairwise distance between points within a fixation.

$$D_1 = \frac{\sum_{i=1}^{n_f} \sum_{j=i+1}^{n_f} d_{ij}}{\binom{n_f}{2}}. \tag{1}$$

The second density metric (D_2) is the minimum area square bounding box surrounding the fixation divided by the number of fixation points it contains:

$$D_2 = \frac{A_{sq}}{n_f}. \tag{2}$$

For both the D_1 and D_2 density metrics, small values imply greater density. A third metric, Standard Distance (*SD*), measures the dispersion of gaze points around the fixation center. *SD* is a common metric in the Geographic Information System (GIS) literature, that evaluates how points are distributed around the fixation center [6]. Similar to standard deviation, *SD* quantifies the dispersion of a set of data values. Hence, the *SD* score is a summary statistic representing the compactness of point distribution. Smaller *SD* values correspond to gaze points that are more concentrated around the center $\left(\overline{X}_f, \overline{Y}_f\right)$ of fixation f, expressed as (Fig. 1):

$$\overline{X}_f = \frac{\sum_{i=1}^{n_f} x_i}{n_f}, \overline{Y}_f = \frac{\sum_{i=1}^{n_f} y_i}{n_f}. \tag{3}$$

The standard distance of fixation f, SD_f, is:

$$SD_f = \sqrt{\frac{\sum_{i=1}^{n_f}(x_i - \overline{X}_f)^2}{n_f} + \frac{\sum_{i=1}^{n_f}(y_i - \overline{y}_f)^2}{n_f}}. \tag{4}$$

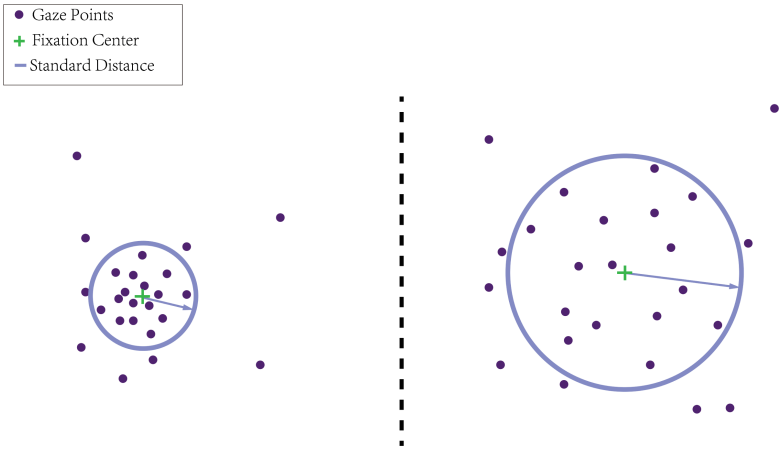

Fig. 1. An illustrative depiction of standard distance, *SD*. When considering an identical number of gaze points, *SD* is smaller when points are more compactly distributed around the center (left); when they are more dispersed, *SD* becomes larger (right).

Spatial pattern analysis can also be examined in measuring the fixation gaze point distribution pattern. The Average Nearest Neighbor (ANN) [6] is used to measure the degree to which fixation gaze points are *clustered*, versus *randomly distributed*, within a fixation bounding area. A fixation resulting from focused gaze toward a single area of interest would tend to exhibit a more uniformly distributed pattern, with greater *ANN* values. The *ANN* ratio is calculated as the average distance between each point and its nearest neighbor, divided by the expected average distance between points if a random pattern is assumed. *ANN* values greater than one imply that the fixation gaze points are *dispersed*; as this ratio decreases, fixation gaze points increasingly exhibit clustering (Fig. 2).

The four metrics D_1, D_2, *SD* and *ANN* will be used to evaluate three aspects of inner fixation patterns: fixation inner-density, fixation points dispersion, and their distribution. We expect fixations identified with the FID filter to be denser and more uniformly distributed than those identified with the I-VT filter. Our density assertion, which stems from the method of fixation identification, helps to test whether the FID filter does indeed more accurately group individual gaze points into focused attention.

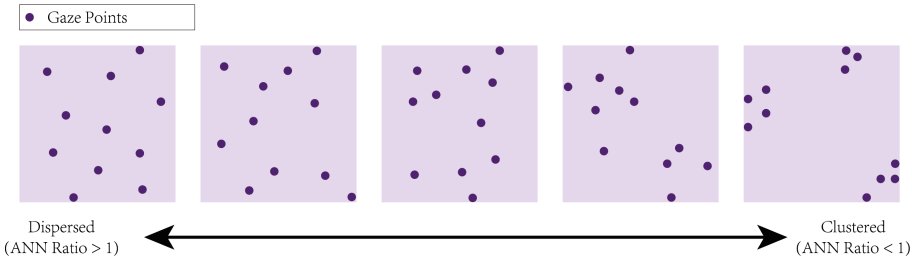

Fig. 2. Illustrating the *ANN* ratio as the distribution of gaze points change within an identical minimum square bounding box.

Our assertion that gaze points identified with the FID filter are more randomly distributed stems from the argument that if a fixation is compact, that is it has high inner-density, it is more likely to have a more uniform distribution around its center.

In addition to the above assertions, we also examine the impact of FID and I-VT filters on fixation duration and center location.

4 Experimental Evaluation

We begin this section by describing the specific context of our eye tracking datasets and experiments. We then compare the I-VT and FID filters with the aforementioned four metrics, and discuss our findings.

4.1 Dataset and Equipment

We perform our experiments on eye movement datasets obtained from a total of 28 university students who were assigned to read a text passage shown on a standard desktop computer monitor. Prior to the experiment, each participant completed a brief eye-calibration process lasting less than one minute. We used the Tobii X300 eye tracker [7] to collect participant's eye-movements. The software version is 3.2.3 and the sampling rate was set to 300 Hz.

The 28 recordings were further analyzed using an Intel core i7-6700MQ computer with 3.40 GHz and 16.0 GB RAM running 64-bit Windows 10. Matlab 2016a and Python 2.7 were used for additional data analysis and processing.

4.2 Data Processing

For each eye tracking record, we used the Tobii Studio I-VT filter [8] to generate I-VT fixation identification results. The velocity threshold V was set to 30°/s, which is the recommended threshold in [8]. The minimum fixation duration is set to 100 ms which is the theoretical minimum fixation duration suggested by other eye tracking studies [9, 10].

We further used the results of the I-VT fixation identification as the input data chunks for the mixed integer programming formulation (MIP) for minimizing square area of fixations from [5]. The Gurobi Optimizer 7.5.1 [11] is used as the solver.

Table 1. Comparison of fixation density for I-VT and FID filters.

Fixation density	I-VT		FID (α = 0.1)		*t*-test	
	Mean (pixel)	STD (pixel)	Mean (pixel)	STD (pixel)	*p-value*	Result
D_1	6.769	2.382	5.994	1.961	<0.0001	Reject
D_2	7.690	10.450	5.112	3.920	0.0025	Reject

The FID filter is parametrized by a manually assigned constant α that enables decision-makers to have fine-tuned control over the density. We varied α from 0 to 1 by steps of 0.1 on one randomly selected eye tracking record and examined the fixation identification results manually. When $\alpha = 0.1$, the clustering result appeared the most reasonable, and averaging D_2 values over all fixations yielded the smallest value, suggesting the algorithm finds the (averaged) densest fixations at $\alpha = 0.1$ comparing to other α levels. Therefore, we set $\alpha = 0.1$ when running the FID filter on the other 27 records. In the following evaluations, we discard the record used for selecting α to avoid data snooping.

4.3 Experimental Results

After discarding the single record above, in this section we first report our statistical analyses from the point of view of a single record. Subsequently, we expand it to all 27 of the (remaining) records in our dataset.

4.4 Comparing I-VT and FID Filters for a Single Record

Fixation inner-density and the distribution of gaze points within an individual fixation are micro-patterns in gaze data. Such patterns are relatively difficult to evaluate by averaging over all eye tracking records. To more thoroughly investigate micro-patterns, we first illustrate the comparison results on the eye tracking record of one randomly selected participant. Toward the end of this section, the comparison summary over all recordings is also included.

For this gaze data record, there are 9,788 gaze points and 110 fixations. We calculated fixation inner-density metrics D_1 and D_2 on each individual fixation. The resulting average of both D_1 and D_2 from the I-VT filter is larger than that of FID, which indicates that fixations from the FID filter are denser than those in I-VT filter result. We performed a paired *t*-test with the following hypothesis:

$$H_0 : \overline{D}_{I-VT} = \overline{D}_{FID},$$
$$H_a : \overline{D}_{I-VT} > \overline{D}_{FID}.$$

The *t*-test on both D_1 and D_2 returns a *p-value* smaller than 0.05, so at a 95% confidence level we reject H_0, which implies $\overline{D}_{I-VT}$ is statistically larger than $\overline{D}_{FID}$ (Table 1).

The *SD* metric measures the dispersion of fixation points around their center. Table 2 reveals that the *SD* mean and standard deviation for the I-VT filter are larger than that of the FID filter. We also performed a paired *t*-test when comparing the *SD* metric. The hypotheses are:

$$H_0 : \overline{SD}_{I-VT} = \overline{SD}_{FID},$$
$$H_a : \overline{SD}_{I-VT} > \overline{SD}_{FID}.$$

With the same 95% confidence level as the previous test, the *t*-test result rejects the H_0. It indicates that the FID filter tends to identify fixations having points that are more dispersed around the center. It further demonstrates that identifying fixations by optimizing for fixation inner-density yields fixations with more compact regions.

Table 2. Comparison of *SD* for I-VT and FID filters.

	I-VT		FID (α = 0.1)		*t*-test	
	Mean (pixel)	STD (pixel)	Mean (pixel)	STD (pixel)	*p-value*	Result
D	5.5033	2.189	4.746	1.616	<0.0001	Reject

Finally, we perform a hypothesis test using the *ANN* ratio [6] to see if the gaze points are randomly distributed in a fixation region:

H_0 : gaze points are randomly distributed within fixation region,

H_a : gaze points are not randomly distributed within fixation region.

If the hypothesis test results in a small *p*-value, we would reject the H_0 because of the small probability that the fixation gaze points are randomly distributed in their fixation region.

The *ANN* hypothesis test is rather sensitive with respect to the bounding region used to cover all fixation points in an individual fixation. Therefore, we perform two experimental results using A_{sq} and A_{rt}, respectively, to represent fixation area. Table 3 reports the count of fixations (out of 110) for which H_0 is rejected at 95% confidence level, implying that there is statistical evidence that fixation points are not randomly distributed. Table 3 reveals that, under both fixation regions, more fixations appear to not be randomly distributed when using the I-VT filter. Moreover, the difference between the I-VT and FID filters is greater under the A_{sq} region. This may be due to A_{sq} typically being larger than A_{rt}, as the FID filter specifically minimizes the square area of fixations.

Table 3. Comparison of *ANN* for I-VT and FID filters, reporting the count of fixations (out of 110) for which H_0 is rejected.

	I-VT		FID (α = 0.1)	
	A_{sq}	A_{rt}	A_{sq}	A_{rt}
# of fixations rejecting H_0	95	60	61	50

We now compare fixation duration and fixation center for the I-VT and FID filters. Fixation duration (*FD*) is a commonly used metric in eye tracking research. We compare the average fixation duration on I-VT and FID filters with the hypotheses that

$$H_0 : \overline{FD}_{I-VT} = \overline{FD}_{FID},$$
$$H_a : \overline{FD}_{I-VT} > \overline{FD}_{FID}.$$

The paired *t*-test result shows that $\overline{FD}_{FID}$ is significantly smaller than $\overline{FD}_{I-VT}$ at a 95% confidence level. This outcome may be due to the FID filter eliminating fixation points and refining the fixation region of each of the fixation chunks from the I-VT filter (Table 4).

Table 4. Comparison of fixation duration for I-VT and FID filters.

	I-VT		FID ($\alpha = 0.1$)		*t*-test	
	Mean (second)	STD (second)	Mean (second)	STD (second)	*p-value*	Result
Fixation duration	0.250	0.151	0.204	0.167	<0.0001	Reject

Fixation center is also a basic feature to represent fixation location, used in the depiction the scan path of eye movement. We introduce the center shift, which is the Euclidean distance between the fixation center of the I-VT filter and that of the FID filter. The 110 fixations within the eye tracking record generates mean and standard deviation (STD) of the center shift data as reported in Table 5.

Table 5. Statistics of fixation center shift between I-VT and FID filter.

	Mean (pixel)	STD (pixel)
Center shift	0.881	1.617

When examining the mean and STD of center shift, it may be inferred that the difference of fixation center is negligible. The bivariate distribution of center shift depicted in Fig. 3 displays the long tail distribution in both x and y axis. The 90% quantile of x, y is 0.922 and 1.308 respectively. It shows that while the refined results of the FID filter can skew some I-VT fixation centers, most of the time the center shift remains in a fairly small range.

4.5 Comparing I-VT and FID Filters for all 27 Remaining Records

The results reported above were for a single eye tracking record. The average number of gaze points for all remaining 27 records is 10,959, and the average number of fixations is 127.7. Table 6 reports the results of the corresponding hypothesis tests for

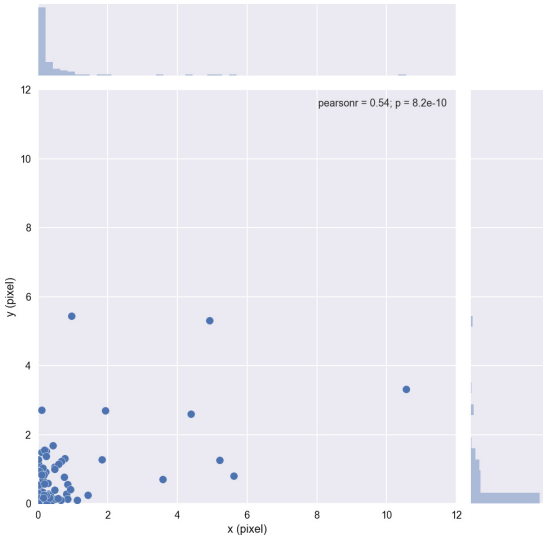

Fig. 3. The bivariate distribution of center shift in x, y coordinates.

D_1, D_2, SD and fixation duration on all the 27 eye tracking records. We find that zero record does not reject the corresponding H_0 in the t-test for D_1, SD and fixation duration, and two for D_2. This analysis shows that the FID filter finds denser and more compact fixations than I-VT filter holds for most of eye tracking records in our dataset in terms of for D_1, D_2 and SD.

Table 6. Summary of hypothesis test results for 27 eye tracking records.

	D_1	D_2	SD	Fixation duration
# of records that do not reject H_0	0	2	0	0

We calculate the center shift between all I-VT and FID filter fixation pairs; the bivariate distribution result is shown in Fig. 4. The distribution on either x or y direction is again a long tail distribution. The 90% quantile value of x, y is 2.095 and 2.411 respectively. Figure 4 shows only a few points that are far away from the origin, indicating that the FID filter identification results can indeed change the fixation center location, though this occured relatively infrequently in our dataset.

We also run the *ANN* hypothesis test on each recording and calculate the count of fixations (FC) for which the *ANN* hypothesis test H_0 ($FC - ANN$) is rejected over all recordings. The average is reported in Table 7. Both the mean and the standard deviation resulting from the FID filter are smaller than that of the I-VT filter.

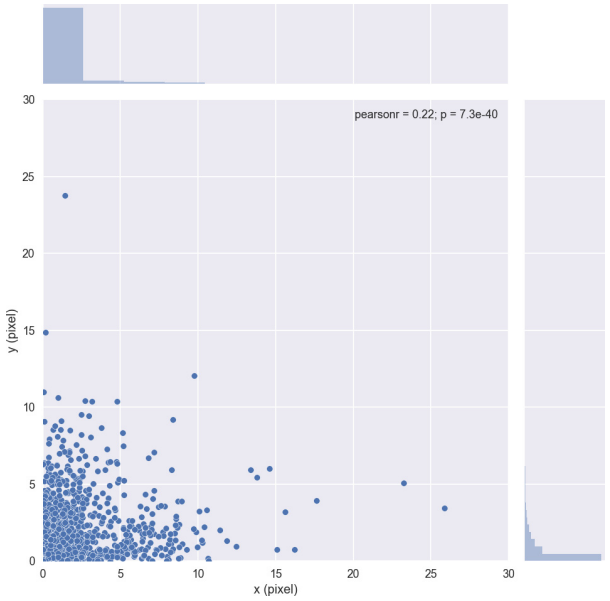

Fig. 4. The bivariate distribution of center shift for all fixations.

We compare the FC results from the I-VT and FID filters by the paired t-test with 95% confidence level and the following hypotheses:

$$H_0 : \overline{FC}_{I-VT} = \overline{FC}_{FID},$$
$$H_a : \overline{FC}_{I-VT} > \overline{FC}_{FID}.$$

The first row in Table 7 shows that when bounding the fixation region by A_{sq}, $\overline{FC}_{FID}$ is significantly smaller than $\overline{FC}_{I-VT}$. It indicates the general trend that the inner gaze points of fixations resulting from the FID filter tend to be randomly distributed. As for A_{rt}, the t-test result also reject H_0, implying that the same conclusion could be drawn on A_{rt}.

Table 7. Comparison of $FC - ANN$ for I-VT and FID filters over all recordings.

Fixation region	I-VT		FID (α = 0.1)		t-test	
	Mean (count)	STD (count)	Mean (count)	STD (count)	*p-value*	Result
A_{sq}	109.1	40.0	70.2	27.4	<0.0001	Reject
A_{rt}	74.5	30.1	64.2	24.6	0.0002	Reject

5 Conclusions

Our results show that the FID filter, as compared to I-VT filter, does indeed identify fixations that are denser and more compact around the center, and more uniformly distributed patterns found in fixation bounding regions. These properties have major implications for two important fixation metrics that are widely used in eye tracking analysis: Fixation duration and location. Our results show that the two filters tend to result in significantly different fixation durations. The results displayed in Figs. 3 and 4 provide evidence that in some cases FID filter can result in quite different fixation centers comparing to I-VT filter. It is important to note that the data used in our study was gathered when users were reading an online text passage, which typically generates more focused fixations. Future investigation using different stimuli are needed to extend the generalizability of these results and to see whether the micro-level differences, including fixation duration and center location, observed in this study between FID and I-VT filters change for different tasks (e.g., reading more challenging text passages, viewing a picture, or browsing a website). For example, in this study we used a reading task which typically results in compact fixations. Using a browsing task may result in much larger differences in fixation center location, because gaze points within fixations in browsing tasks tend to more dispersed [5]. The metrics introduced in this study to compare fixations at a micro level serve to refine the analysis of eye movements to a deeper level. Future studies, however, are needed to validate and extend our findings.

The results of this study contribute in two ways to eye tracking studies that examine user behavior. First, they show that researchers can identify focused attention with the FID filter and thereby improve the sensitivity of their analysis with regard to duration and center location of intense attention. Second, the micro-analysis introduced in this study provides a new way to compare gaze points within a fixation. This is important because it allows researchers to examine relationships between eye movements and behavior at a much smaller unit of analysis, namely fixation micro-patterns.

References

1. Djamasbi, S.: Eye tracking and web experience. AIS Trans. Hum.-Comput. Interact. **6**(2), 37–54 (2014)
2. Nyström, M., Holmqvist, K.: An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behav. Res. Methods **42**(1), 188–204 (2010)
3. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 71–78. ACM, November 2000
4. Shojaeizadeh, M., Djamasbi, S., Trapp, A.C.: Density of gaze points within a fixation and information processing behavior. In: Antona, M., Stephanidis, C. (eds.) UAHCI 2016. LNCS, vol. 9737, pp. 465–471. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40250-5_44
5. Trapp, A.C., Liu, W., Djamasbi, S.: New density-based optimization formulations and algorithms to identify fixations in gaze data. In: Presented in INFORMS Annual Meeting, Houston, TX, INFORMS, Hanover, MD (2017)

6. Mitchell, A.: The ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics. ESRI Guide to GIS analysis (2005)
7. Tobii: Tobii technology (2017). http://www.tobii.com. Accessed 18 Dec 2017
8. Olsen, A.: The Tobii I-VT fixation filter. Tobii Technology (2012)
9. Blignaut, P.: Fixation identification: The optimum threshold for a dispersion algorithm. Atten. Percept. Psychophys. **71**(4), 881–895 (2009)
10. Komogortsev, O.V., Gobert, D.V., Jayarathna, S., Koh, D.H., Gowda, S.M.: Standardization of automated analyses of oculomotor fixation and saccadic behaviors. IEEE Trans. Biomed. Eng. **57**(11), 2635–2645 (2010)
11. Gurobi Optimization, Inc.: Gurobi Optimizer 7.5.1 Reference Manual (2017)

Cognition and Predictors of Password Selection and Usability

Lila A. Loos(✉) and Martha E. Crosby(✉)

University of Hawai'i at Mānoa, Honolulu, HI 96822, USA
{lila7194,crosby}@hawaii.edu

Abstract. Computer passwords represent a secure authentication process used to access electronic information. Inconsequential of data storage location many of us utilize multiple unique computer passwords to access information on a daily basis. Since the design of password requirements are contingent upon the system provider, recalling various passwords is cognitively demanding and results in insecure practices such as writing down passwords visible to passerby. This study examines the task of password selection to improve human computer interaction. Categorizing personality through the locus of control internal and external scale and cognitive factors through memory associations advances understanding of password decision making. These classifications establish associations for predictive password selection informed by the behavioral decision process. This study addresses a design gap in the utility of passwords and describes quantified convergent dispositional factors gathered through valid instruments. Psychological fields of personality, memory cognition and behavioral decision making inform usability in the human computer interaction area of computer science.

Keywords: Authentication · Usability · Cognition · Memory · Password
Locus of control

1 Introduction

The goal of this study is to improve awareness of computer password selection and augment the security mechanism by evaluating locus of control and cognitive memory dynamics for human centered design enhancement. Most users choose short passwords to facilitate memorability and facilitate memorability with short passwords [27]. Studies excluding memorability from password security are able to determine the effect of visual password strength meters as a method to address security concerns with weak passwords. User behavior was positively affected by circumstantial messages from the strength meter resulting in users creating stronger passwords. Their meter was constructed with contextual information appealing to the users as well as a link providing training on password security [35]. Similarly, Jang-Jaccard and Nepal [34] argue for visual or biometric passwords as an option as they don't require memory.

Evaluating individual password decision making supports user centric factors. "While there is no silver bullet solution to the user authentication problem, it is still important to work toward improvements in password usage, security systems, and

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 117–132, 2018.
https://doi.org/10.1007/978-3-319-91467-1_10

understanding threats" [33, p. 78]. A study exposing the differences in awareness and practice of strong password use among college students found most authenticate utilizing seven passwords. As a result, a security awareness strategy established unique passwords for each login, changing passwords on a regular basis and keeping passwords private. In an effort to create passwords that are difficult to guess by hackers, it was suggested to make simple changes such as adding a symbol or upper-case letter to existing passwords [6].

Considering the number of unique passwords used for educational, personal and occupational purposes, "a solution to the problem of password security versus memorability has yet to be found" [25, p. 3761]. Additionally, this study of weak passwords employed persuasive technology to strengthen user security and memorability by inserting random characters into the password string. The experiment results improved password security for users with weak and strong original passwords, however, memorability was not improved for users with strong passwords. Likewise, Gehringer [26] recognizes that multiple password logins necessitate recording them for future retrieval and advises to involve the human component of security and memorability to future inquiry. As this study combines cognitive behavioral activities with technology, Choong [15] suggests a holistic approach to alleviate the memorability burden on users and brings attention to the need of usability research.

2 Human Computer Interaction: Usability

According to Norman [47] usability design combines psychology, computer science, engineering and analytical disciplines. Security is a technical issue imposed on humanity and disrupted by excessively complex technology measures that daunt employee behaviors leading to insecure conduct such as posting passwords in their work spaces in open view. A gap between usability and security is acknowledged and password usability is deserving of examination.

Jang-Jaccard's and Nepal [34] study addresses the relationship between usability and security resulting in higher recall between passphrases and self-selected passwords compared with random passwords. Unlike self-selected passwords, the passphrases withstood simulated dictionary and brute force attacks. Another memorability study indicated difficulty with learnability and recall when using more secure passwords [31]. Likewise, Greene et al. [28] agree that passwords are not memorable and security is threatened by compromised password data banks. Employing security and usability experts, the study measured the loss of security in passwords specific to the multiple keyboards presented on mobile devices. Results define effectiveness measured through password or character login success and failures; efficiency is measured by the length of time it takes to enter a password and satisfaction is measured through subjective user experiences. Similarly, Grassi et al. [27] define usability as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (p. 61). Choong [15] argues that usability is the main concern for users managing multiple passwords.

> Research focusing on human factors and usability of passwords has been challenging the view that users are the primary cause for cyber security issues and pointing out that security policies are often imposing unreasonable requirements and pushing users' cognitive limits. (p. 128)

Although people prefer to use memorable passwords, favoring usability over security presents authentication risks to defending systems as the goals between usability and security are dissimilar [2]. Likewise, Choong [15] suggests collaboration among interdisciplinary influences to discover the intersection between security and usability and acknowledges the need for research to ensure security while reducing the burden on users. Considering the weakest link in security systems are individuals, human computer interaction principals rooted in psychology and cognition impact behavior and warrant further study to improve the authentication processes.

Norman [48] argues "without usable systems, the security and privacy simply disappear as people defeat the processes in order to get their work done" and furthermore, "the more secure you make something, the less secure it becomes" (p. 60). Security professionals attest to challenges between security and usability that trigger insecure behaviors in response to usability difficulties. "The reasonableness of the effort required" (p. 61) to comply with security requirements is a design issue yet to be solved. Renaud's et al. [52] framework considers usability and offers authentication options for decision makers based on their system requirements and preferences. The value given to a resource is aligned with the authentication method. Alternative authentication allows the decision maker to personalize memorability and risk mitigation properties such as strength of password. This framework considers the quality of password authentication for the business and the user as an alternative approach to computer security.

3 Risk Assessment and Authentication

Risk assessments examine computer authentication, human computer interaction and fiscal responsibilities to secure information systems. Grassi et al. [27] defines authentication as "verifying the identity of a user, process, or device, often as a prerequisite to allowing access to a system's resources" (p. 47). Additionally, an authentication secret is a "value that an attacker could use to impersonate the subscriber in an authentication protocol" (p. 47). Passwords are defined as "a type of authenticator comprised of a character string to be memorized or memorable by the subscriber, permitting the subscriber to demonstrate something they know as part of an authentication process" (p. 54). This study considers passwords an authentication secret and investigates the process of password construction and memorability to impede risk.

Security insurance levels examined by protection requirements create digital authentication realities affecting anyone who accesses an information system through a login dialogue. The interface is primarily evaluated by the user who adheres to the login requirements or who may create a supplemental coping mechanism to successfully authenticate. Grassi et al. [27] integrates organizational risk to operational security practice.

> The process of identifying, estimating, and prioritizing risks to organizational operations (including mission, functions, image, or reputation), organizational assets, individuals, and other organizations, resulting from the operation of a system. It is part of risk management, incorporates threat and vulnerability analyses, and considers mitigations provided by security controls planned or in place. Synonymous with risk analysis. (p. 59)

Organizational risk focuses on insider threats to computer security systems. Unlike external exploiters, privileged users hold company knowledge threatening the integrity of information systems. Nurse et al. [49] explores the impact of fraud, theft of intellectual property, and sabotage of infrastructure and provides insight to detecting and preventing malicious behavior utilizing automated analytical tools along with policy awareness to protect organizations. An environment in which access to information is controlled by users compels memorable password authentication, devoid of non-secure actions such as writing down passwords that are difficult to recall. According to Choong et al. [16] "an alarming finding is that employees seem to have a false perception of security around their work-related accounts" (p. 13) and consider prevention of system attacks and security breaches the responsibility of their organization. Furthermore, these Federal employees prescribe to the notion that government work is transparent to the public and without consequences resulting from a system compromise.

Organizational property risk is pivotal to security and requires assessing types of authentication attacks, secrets uncovered, stolen, or tampered, copied intellectual property and replication of user identity [27]. Watkins [64] value assessment of information security is estimated by the impact of the following values: "Risk = Likelihood $\times$ Impact Relationship" (p. 22). The possible vulnerabilities and impact range from very low to very high and apply to all risk involving information assets and security. Calder [11] recommends assessing risk with goals to remove, reduce, tolerate or transfer risk through a contract such as an insurance policy to protect organizations from business harm. Furthermore, a risk management plan identifies action, responsibilities, and priorities of information security while management details changes, corrective and preventative action, and recommends improvement.

An alternative study suggested by Cavusoglu et al. [12] involve game theory as a security investment decision maker.

> The firm's payoff from security investment depends on the extent of hacking it is subjected to. The hacker's payoff from hacking depends on the likelihood he or she will be caught. Thus, the likelihood of the firm getting hacked depends on the likelihood the hacker will be caught, which, in turn, depends on the level of investment the firm makes in IT security. (p. 90)

Selecting the preventative scheme results in maximum savings. Therefore, each recommended security option is weighed against its cost and estimated intrusion parameters. Cost savings are determined by comparing the cost of implementing or not implementing security technology.

Risk assessments indicate password attack methods involve guessing by brute force or dictionary listing of words, guessing, eavesdropping, social engineering, and physical presence. Since text passwords continue to be the dominating login for authentication, Ives et al. [33] exploit reuse of passwords for multiple accounts to obtain higher level system access. Furthermore, authenticating to e-commerce sites

using the same password presents security risks for the user as hackers could obtain access to multiple sites, "there is an obvious and probably sizeable overlap between AOL and Citibank or BankOne and Amazon.com customers" (p. 75). Jang-Jaccard and Nepal [34] discuss the significance of acquiring passwords through deceptive practices: cyber-attacks by malware phishing or injecting computer code to obtain a database of passwords using unsecure technologies such as Wi-Fi or Bluetooth connections. Defense mechanisms include locking an account after failed attempts, establishing secure connectivity during communication, and enforcing password requirements [28]. Although password managers address memorability by collecting user identification and passwords used on various web sites, such tools become susceptible to attack [57]. Aurigemma et al. [4] suggest password managers provide a security mechanism for home-end users however their investigation discovered that insufficient time was the main inhibitor to adoption followed by threat apathy or lack of password security threats.

Growing threats to technology and malicious attack patterns compel security policies to embrace usability and safeguard information systems. Adams and Sasse [1] recognize the shift to user centric design of security systems by determining how their systems are utilized. Although text-based passwords pose a security risk to dictionary attacks [14, 65], Wu's [65] study allows the use of simple passwords defended by encryption against such attacks. Notwithstanding user involvement risk in security technology, passwords and usability are coupled in human computer interaction and form the basis of this study.

4 Authentication and the Human Factor of Text Passwords

According to Norman [47] password security continues to be problematically a human element. In response to security procedures paired with bad password policies, the human factor creatively adapts to solve forgotten password problems. As the majority of individuals compose passwords using the name of a person, place or thing, date, and number, Brown et al. [9] suggest passwords to be easily recalled by the user and not others. Nevertheless, the conditions of creating a password are often rushed without much time for thoughtful composition. In response to security concerns, mechanisms exist to warn users of insecure password fields on web browsers [36].

Creating a password is one part of the authentication process. Users participate in a password lifecycle system to generate, maintain, and authenticate using a process comprised of goals, constraints, memory storage, and authentication experiences that influence the recurrence of password selection [15]. This repetitive method includes individual factors such as attitudes, motivations, and emotions to comply with password requirements and individual needs. Workarounds to password memorability utilize a special character and digit to user generated passwords or created a password from the first letter of a phrase [62]. Having the user login multiple times produced secure and memorable passwords. Varying password requirements include regularly changing passwords circumvent memorability [61]. Additionally, their study suggests short-term recall testing, unlike immediately using the password, improves password retention since retrieval is from long term retention rather than from working memory.

Furthermore, using a mnemonic method to generate a password containing the first letter of a phrase along with a maximum of three attempts to authenticate is recommended to overcome the security and memorability tradeoff.

Once a password is generated, it is combined with random data or salt which is then cryptographically hashed before it is stored in a database. During authentication, the password is decrypted and compared with the stored salted hash [28]. Bonneau et al. [8] recognize the unbalance between security and authentication and focus on uniting their differing roles. Although passwords were initially designed to access mainframe systems, today's graphical environment is dominated by web authentication.

> Failure to recognize the broad range of usability, deployability, and security challenges in Web authentication has produced both a long list of mutually incompatible password requirements for users and countless attempts by researchers to find a magic-bullet solution despite drastically different requirements in different applications. No single technology is likely to 'solve' authentication perfectly for all cases: a synergistic combination is required. (p. 79)

Usability calls for examinations in password selection united with security. Komanduri et al. [37] study of password policies on password strength and user behavior found that crafting a password policy using 16-characters without additional requirements provides greater resistance to brute force attacks with an increase in usability compared with eight-characters containing a number and symbol requirement.

5 Cognition and User Behavior of Text Passwords

According to Michaelian and Sutton [44] "cognitive science is the interdisciplinary study of mind and intelligence, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology" (p. 1). Norman [46] combines cognition with emotion as part of the psychology of design. "Cognition provides understanding: emotion provides value judgements. A human without a working emotional system has difficulty making choices. A human without a cognitive system is dysfunctional" (p. 47). The study emphasizes emotion produced from well-designed devices can lead to pleasure or despair and poses the question, "do we count our technology as an extension of our memory systems" (p. 46). "It is one thing to have to memorize one or two secrets: a combination, or a password, or the secret to opening a door. But when the number of secret codes gets too large, memory fails" (p. 86). Memory overload is addressed by using few passwords for multiple logins. "Even security professionals admit to this, thereby hypocritically violating their own rules" (p. 87). Furthermore, complex passwords stymie memory leading to security violations by employees who use external memory options such as paper to aid in password retrieval.

Users employ risky behaviors when engaging in simple passwords, writing down or sharing passwords and not changing passwords on a regular basis [64]. Simple passwords manifest as "proper names and birthdays are the primary information used in constructing passwords, accounting for about half of all passwords. Almost all respondents reuse passwords" [9, p. 641]. Study results found that participants had a mean of about eight passwords and half are unique.

Individuals with numerous passwords inherit usability problems; a decrease in memorability was associated with an increase in cognitive overhead [1]. Likewise, insecure password behaviors are a result of insufficient awareness of password procedures and security threats like password cracking [1]. Similarly, Florencio's and Herley [23] study of more than 500,000 users each having 25 accounts who use an average of eight passwords a day, resulted in participants remembering groups of passwords through combinations of memory, writing them down, and password resets. Users select weak passwords consisting mostly of lowercase letters, unless required to use uppercase and special characters, and reuse passwords for multiple authentication across websites. Moreover, a case study of Federal employees resulted in an average of nine accounts requiring logins. Twenty five percent of employees managed 11 through 20 passwords. Password requirements are considered complex with frequent changes. Frustrations of mistyping and forgetting often resulted in getting locked out of their account make the password management lifecycle of generating and tracking troublesome. Eighty one percent of respondents prefer passwords that are easy to remember and prefer a single sign on system [16]. Pilar et al. [51] acknowledge the need to improve password-based authentication procedures. Their study showed respondents utilized approximately eight passwords of which at least one password is reused. Memory difficulties in the form of forgetting or mixing up passwords increased with groups using multiple unique passwords. Password lengths increased with the younger and more educated group.

To overcome excessive demands on memory, España [22] examined a technique that combines pieces of information that result in positive memorability for standard passwords and not for multiword or mnemonic passwords. An approach by Tam et al. [59] suggest that users select common passwords based on convenience. "Focusing on the user is important because, although stronger authentication techniques are available, corporations tend to continue to use a password-based system to control system access" (p. 233). Therefore, understanding why users mismanage passwords is essential to enhancing password behavior. Their study showed that users value convenience even though compromising password practices could lead to security breaches of their personal data. Such findings are a result of users placing emphasis on near versus distant future events. Thus, importance of convenience or feasibility of a near future event is favored over the security of a distant future event. However, the study suggests that stronger passwords are chosen for bank accounts than for email accounts resulting in a tradeoff between convenience and security.

Norman [46] explains the encoding of mnemonic phrases help memory retention as it is affected by time and quantity. "Most of us can't (remember all these secret things) even with the use of mnemonics to make some sense of nonsensical material" (p. 88). Users cognitively problem-solve and reason when selecting characters to create a password [15]. Factors affecting the authentication process include the frequency of use, maintenance and interferences from other passwords.

In addition to the great number of complex passwords and memory overload, Greene et al. [29] detail increased task constrictions on mobile devices; it is an overall challenging authentication method requiring smaller keys, multiple character keyboards and task interruptions associated with switching screens. Their study of 158 participants averaging 33.2 years in age suggests constraints of password recall between mobile and desktop platforms.

6 Locus of Control Personality Variables

Applying psychological variables of locus of control to technology is expected to increase understanding of personality influences on the selection and construction of computer passwords and contribute to the design of memorable passwords. This study operationalizes internal and external locus of control as an influencer to decision making in computer security. Individuals respond to perception based on attitudes and behavior. External control "is typically perceived as the result of luck, chance, fate, as under the control of powerful others, or as unpredictable because of the great complexity of the forces surrounding him" while internal control is "contingent upon his own behavior or his own relatively permanent characteristics" [53, p. 1]. Based on social learning theory, the relationship between behavior and consequences is evident in Rotter's [53] hypotheses "when the reinforcement is seen not contingent upon the subject's own behavior that its occurrence will not increase an expectancy as much as when it is seen as contingent" (p. 2). Furthermore, "once a person has established a concept of randomness or chance the effects of reinforcement will vary depending upon what relationship he assigns to the behavior reinforcement sequence" (p. 4). Therefore, a person's internal (skill) control or external (chance) control variable affects reinforcement and subsequently, behavior. The study determined that same situations are considered differently depending if the individual's personality is characteristic of internal or external control factors and predictions of behavior can then be determined. Likewise, reinforcement is perceived as:

> Learning processes such that people with a belief in internal control are more likely to change their behavior following a positive or negative reinforcement than are people with a belief in external control. For behavior change to occur, however, the reinforcement must be of value to the person. [42, p. 251]

Various technology studies apply locus of control to better understand user behavior. Coovert and Goldstein [18] demonstrate the use of locus of control to understand how employees perceive computer related changes in the work environment; internal control personnel resulted with a higher positive attitude compared with external control personnel. Chak and Leung [13] suggest external locus of control or trust on chance contributes to Internet addiction disorder. Li et al. [39] indicate internal locus of control individuals are more likely to regulate mobile phone use to not interfere with their well-being. Fong's et al. [24] research on the re-adoption of mobile phone applications determined that locus of control is an influencer through self-efficacy. Individuals with internal locus of control are driven by success and are likely to overcome operational difficulties with the mobile applications and adopt reuse.

Specialized studies [10, 32, 41, 45, 55, 58, 63] modify locus of control's general internal and external variables instrument to determine control beliefs and behavior in various sectors including consumer strategic shopping, e-learning systems, organizational impression management, meta-analysis of well-being including motivation and behavioral orientation, preventive tobacco, work settings, and health care.

Alternate scale formats deviate from Rotter's [53] forced choice instrument. Studies opting for Likert's multidimensional scale [3, 40] measure social and cultural situations and aspects pertaining to personal well-being. Supported by Rotter [53], the locus of

control internal-external scale "correlates satisfactorily with other methods of assessing the same variable such as questionnaire, Likert scale, interview assessments, and ratings from a story-completion technique" (p. 25).

Although Rotter's [53] locus of control is a unified measurement of personality constructs by design, Lange and Tiggemann [38] suggest multidimensionality in the widely used personality scale consisting of 29 questions on topics such as "social-political events, social recognition, academic recognition and general life philosophy" (p. 398). However, Rotter [54] argues for instrument validity to generalize situational reinforcement of internal or external control variables for potential behavior expectancy where "expectancies in each situation are determined not only by specific experiences in that situation but also, to some varying extent, by experiences in other situations that the individual perceives as similar" (p. 57). Moreover, Ng et al. [45] operationalize locus of control as a continuous variable using Rotter's [53] scale to predict workplace attitudes and behavioral intent to control.

7 Memory Cognition Factor

This study identifies memory aptitude factor using Ekstrom's et al. [21] cognitive tests. Although "there are probably no such things as truly 'pure' factors, a study of individual differences in abilities can profit greatly if it is closely tied to the experimental analysis of particular cognitive tasks" (p. 3). Memory cognition is explored to better understand password recall abilities.

Baddeley [5] defines working memory as "temporary storage and manipulation of the information necessary for such complex tasks as language comprehension, learning, and reasoning" which "evolved from the concept of a unitary short-term memory system" (p. 556). Additional findings associate memory cognition with attention and behavior control. Similarly, the ability to encode active information to long-term memory corresponds with maintaining information in working memory aided by attention control [60]. Moreover, working memory's storage and attention control operations have the ability to sidestep disturbances; increased awareness in a discipline contributes to working memory capacity [17]. Norman [46] suggests information stored in working memory disappears with distraction. Therefore, it is suggested to portray information in various forms to enhance memory recall.

Although immediate memory retrieval is described as effortless recollection, recollection difficulty increases as time passes [46]. Without repetition, working memory's capacity is seven items compared with 10 or 12. Since recalling arbitrary items like passwords is considerably challenging, individuals learn to develop associations by creating organization. Generating meaningful understanding to mixtures of characters, numbers, and symbols is an effective memorability technique. Recalling a password whose length is greater than working memory capacity or numerous passwords with diverse conditions is yet to be solved. Continued development is vital to enhance interfaces and interaction toward usable security [48].

8 Behavioral Decision Theory

Decision making is an action mechanism encountered by individuals during the construction of computer passwords as is organized by gathering information and evaluating alternatives to reach a specified goal.

> The importance of behavioral decision theory lies in the fact that even if one were willing to accept instrumental rationality as the sole criterion for evaluating decisions, knowledge of how tasks are represented is crucial since people's goals form part of their models of the world. [20, p. 60]

Decision behavior is subject to information processing of the task resulting from a cost benefit approach. Evaluating decision promptness against effort are deliberated resulting in lessening alternatives or processing complexity. Cost benefit is characterized as effort error [50].

> Furthermore, the cost of thinking is simply the number of comparisons that are made. The number of comparisons is seen as a function of (a) the desired probability of making a correct choice and (b) the difficulty of making a choice. (p. 396)

Additional studies [20, 50] imply decision strategies depend on past learning experiences of task variables and expected cost. Consequently, while comparisons between high benefit and low-cost decision outcomes create a good decision, Higgins [30] suggests regulatory fit increases the choice value of judging criteria. The favored result produces experiences of motivation and positivity of the decision-making process. Accordingly, decision making is dependent upon a cost benefit framework where the cost of the resource is attributed to the most advantageous result selected [7]. Although personality factors contributing to decision making are ignored because of lack of priori research, the study acknowledges individual characteristics, opinions, perception and knowledge factors leading to a decision. Simplifying alternatives in the decision-making process produces an alternative measurement of the cost of thinking [54].

Norman [47] argues emotion is cognition's necessary partner of judgments that enhance our decision-making process. Influencing behavior are cognitive and emotional factors that interplay in determining how we respond to technical problems with security. The elements specific to this study examine psychology factors and the assessment of memory encoding and decoding capacity that are associated with the decision-making process of creating passwords.

9 Methodology

The purpose of this study is to improve attentiveness of computer password selection and heighten the security mechanism by presenting design conclusions based on results. Outcomes from descriptive quantitative research will suggest associations between the operationalized variables and the represented population. Rooted in psychological variables and memory cognition constructs, assessments in control beliefs are applied to technology to predict security-based behavior. Results will increase

understanding of personality influences and password recall abilities on the selection and construction of passwords to enhance human centered design.

The significance of personality and cognitive factors has serious and practical applications addressing information security and usability. Contributions from user interpretations drive reinforcement of personal traits and its association with behavior. The combination of theoretical perspectives provides objective methods of assessment for predictive technological decision making. "Cognitive style research is based on Carl Jung's 1921 premise that the mental functions related to information gathering and decision making are central to one's personality" [43, p. 811]. Their predictive Internet acceptance study operationalized personality and cognitive dispositional factors that had been ignored in prior research. Although personality resulted as a predictor of Internet adoption, other measures of cognitive style may be influencers.

The research instruments consist of Rotter's Locus of Control Internal External self-evaluation questionnaire, Memory Associative Factor-Referenced Cognitive Tests), Password Selection Survey (Appendix) and Password Recall Survey. These mechanisms operate to produce a descriptive quantitative research study in two phases. The first phase pilots a study of adult university computer science students with objectives to test the validity of the instruments. The second phase applies the research methods to a larger study consisting of an employee population who authenticate to business applications with multiple passwords.

The data collected will not contain personal information. To ensure confidentiality, the research data will not be shared with anyone. To ensure anonymity, no identifying characteristics are recorded on the data and therefore, the researcher will not know who contributes a given piece of data. Pseudonyms may be used to report findings in a way that protects privacy and confidentiality. Participating in either study is optional.

The statistical analysis on the data collected will be logged and represented as patterns of decision making to determine relationships in answering the research questions. ANOVA will determine the main effects and interactions among the locus of control and memory associative factors and password selection. Correlation and linear regression will determine the relationships among the personality and cognitive factors and password selection. ANOVA is designed to contribute to decision making about the differences among the personality groups and selected passwords that contribute to usability. Depending on the sample size, either the z-test or t-test that compare the means of populations will be analyzed along with the f-ratio that finds variance or measure of sample dispersion from the mean.

10 Future Studies

Enhancing authentication security involves incorporating augmented cognition in the usability equation. Considerations of physiological measures enhance psychological factors and further understanding in behavioral decision making of password construction. Future sensory input measurements from eye movements and body heat including perspiration support opportunities to discover cognitive variable associations in the design of system interfaces that aid memorability.

Appendix

Password Selection Survey

Introduction. From the Desk of Thomas F. Duffy, Chair, MS-ISAC
Cybersecurity experts continually identify the use of strong, unique passwords as one of their top recommendations. However, this is also one of the least commonly followed recommendations because unless you know the tricks, it's difficult to remember strong, unique passwords for every login and website.

Why Strong, Unique Passwords Matter
Cybersecurity experts make the recommendation for strong, unique passwords for several reasons – the first being that every day malicious cyber threat actors compromise websites and online accounts, and post lists of usernames, email addresses, and passwords online. This exposes people's passwords, and worse yet, they are exposed with information that uniquely identifies the user, such as an email address. That means that a malicious actor can look for other accounts associated with that same person, such as work related, personal social media, or banking accounts. When the malicious actor finds those accounts, they can try logging in with the exposed password and if the password is reused, they can gain access. This is why unique passwords matter.
Secondly, when malicious cyber threat actors can't easily find or a guess the password, they can use a technique called brute forcing. This is a technique where they try every possible password until the correct password is identified. Computers can try thousands of passwords per second, but for this technique to be worthwhile, the malicious cyber threat actor needs the password to be easy to identify, which is why a strong password matters. The stronger the password the less likely brute forcing will be successful.

When malicious actors use brute forcing techniques they often try every word in the dictionary because it's easier to remember words than random letter combinations. This technique is not limited to English-language dictionaries, so switching languages will not help. And since many passwords require a combination of uppercase and lowercase letters, numbers, and symbols, the malicious actors rely on human instinct to narrow down the possibilities. For instance, most users when faced with choosing a password that fits these requirements, will pick a word, put the uppercase letter first, and end the password with the number and symbol. Alternatively, many people will replace common letters with a number or symbol that represents that letter. This changes a common password, such as "password," into the only slightly more complex password of "p@ssw0rd," which is still an easy to guess pattern.

Another technique to assist in building strong, unique passwords, is to choose a repeatable pattern for your password, such as choosing a sentence that incorporates something unique about the website or account, and then using the first letter of each word as your password. For example, the sentence: "This is my January password for the Center for Internet Security website." would become "TimJp4tCfISw." This password capitalizes 5 letters within the sentence, swaps the word "for" to the number "4," and adds the period to include a symbol. The vulnerability in this technique is that if multiple passwords from the same user are exposed it may reveal the pattern.

Variations on this technique include using the first letters from a line in a favorite song or a poem.

1. Select the most memorable password from the following list of random passwords:
   (a) ksitjgJ8@9
   (b) i5euyrpAT(
   (c) TimMp4ticsPSRp
   (d) 2jU40t#fBa
   (e) tcJotr2atM
2. Modify one of the passwords in question #1 shown above to make it memorable for you. You may also select one of the given random passwords.
3. Enter a strong password of your choice that is memorable for you. A strong password is a unique password that is only used with one account and follows the following format. The password should be at least ten (10) characters in length and include uppercase and lowercase letters, at least one number, and at least one symbol.
4. Describe how you created the strong password and made it memorable for you

References

1. Adams, A., Sasse, M.A.: Users are not the enemy. Commun. ACM **42**(12), 40–46 (1999)
2. Andriotis, P., Tryfonas, T., Oikonomou, G.: Complexity metrics and user strength perceptions of the pattern-lock graphical authentication method. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 115–126. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07620-1_11
3. Ashkanasy, N.M.: Rotter's internal-external scale: confirmatory factor analysis and correlation with social desirability for alternative scale formats. J. Pers. Soc. Psychol. **48** (5), 1328 (1985)
4. Aurigemma, S., Mattson, T., Leonard, L.: So much promise, so little use: what is stopping home end-users from using password manager applications? In: Proceedings of the 50th Hawaii International Conference on System Sciences (2017)
5. Baddeley, A.: Working memory. Science **255**(5044), 556–559 (1992)
6. Bain, L.Z., Hayden, M., Sneesby, S.: An empirical study of user authentication: the perceptions versus practice of strong passwords. Issues Inf. Syst. **XI**(1), 256–265 (2010)
7. Beach, L.R., Mitchell, T.R.: A contingency model for the selection of decision strategies. Acad. Manag. Rev. **3**(3), 439–449 (1978)
8. Bonneau, J., Herley, C., Van Oorschot, P.C., Stajano, F.: Passwords and the evolution of imperfect authentication. Commun. ACM **58**(7), 78–87 (2015)
9. Brown, A.S., Bracken, E., Zoccoli, S., Douglas, K.: Generating and remembering passwords. Appl. Cogn. Psychol. **18**(6), 641–651 (2004)
10. Busseri, M.A., Lefcourt, H.M., Kerton, R.R.: Locus of control for consumer outcomes: Predicting consumer behavior. J. Appl. Soc. Psychol. **28**(12), 1067–1087 (1998)
11. Calder, A.: ISO27001/ISO27002 A Pocket Guide. IT Governance Pub (2008)
12. Cavusoglu, H., Mishra, B., Raghunathan, S.: A model for evaluating IT security investments. Commun. ACM **47**(7), 87–92 (2004)
13. Chak, K., Leung, L.: Shyness and locus of control as predictors of internet addiction and internet use. CyberPsychol. Behav. **7**(5), 559–570 (2004)

14. Chang, T.Y., Tsai, C.J., Lin, J.H.: A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices. J. Syst. Softw. **85**(5), 1157–1165 (2012)
15. Choong, Y.-Y.: A cognitive-behavioral framework of user password management lifecycle. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 127–137. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07620-1_12
16. Choong, Y.Y., Theofanos, M., Liu, H.K.: United States Federal Employees' Password Management Behaviors: A Department of Commerce Case Study. US Department of Commerce, National Institute of Standards and Technology (2014)
17. Conway, A.R., Cowan, N., Bunting, M.F., Therriault, D.J., Minkoff, S.R.: A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. Intelligence **30**(2), 163–183 (2002)
18. Coovert, M.D., Goldstein, M.: Locus of control as a predictor of users' attitude toward computers. Psychol. Rep. **47**(3_suppl), 1167–1173 (1980)
19. Duffy, T.F.: Why Strong, Unique Passwords Matter - Office of Enterprise (n.d.). http://ets.hawaii.gov/why-strong-unique-passwords-matter. Accessed 8 Feb 2018
20. Einhorn, H.J., Hogarth, R.M.: Behavioral decision theory: processes of judgement and choice. Annu. Rev. Psychol. **32**(1), 53–88 (1981)
21. Ekstrom, R.B., Dermen, D., Harman, H.H.: Manual for Kit of Factor-Referenced Cognitive Tests, vol. 102. Educational Testing Service, Princeton (1976)
22. España, L.Y.: Effects of password type and memory techniques on user password memory. Psi Chi J. Psychol. Res. **21**(4) (2016)
23. Florencio, D., Herley, C.: A large-scale study of web password habits. In: Proceedings of the 16th International Conference on World Wide Web, pp. 657–666. ACM (2007)
24. Fong, L.H.N., Lam, L.W., Law, R.: How locus of control shapes intention to reuse mobile apps for making hotel reservations: evidence from Chinese consumers. Tour. Manag. **61**, 331–342 (2017)
25. Forget, A., Biddle, R.: Memorability of persuasive passwords. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 3759–3764 (2008). https://doi.org/10.1145/1358628.1358926
26. Gehringer, E.: Choosing passwords: security and human factors. In: Proceedings of the IEEE 2002 International Symposium on Technology and Society (ISTAS 2002). Social Implications of Information and Communication Technology, (Cat. No.02CH37293) (2002). https://doi.org/10.1109/istas.2002.1013839
27. Grassi, P.A., Garcia, M.E., Fenton, J.L.: Digital identity guidelines (2017). https://doi.org/10.6028/NIST.SP.800-63-3
28. Greene, K.K., Franklin, J.M., Greene, K.K., Kelsey, J.: Measuring the Usability and Security of Permuted Passwords on Mobile Platforms. US Department of Commerce, National Institute of Standards and Technology (2016)
29. Greene, K.K., Gallagher, M.A., Stanton, B.C., Lee, P.Y.: I can't type that! P@$$w0rd entry on mobile devices. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 160–171. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07620-1_15
30. Higgins, E.T.: Making a good decision: value from fit. Am. Psychol. **55**(11), 1217 (2000)
31. Hub, M., Capek, J., Myskova, R.: Relationship between security and usability-authentication case study. Int. J. Comput. Commun. **5**, 1–9 (2011)
32. Hsia, J.W., Chang, C.C., Tseng, A.H.: Effects of individuals' locus of control and computer self-efficacy on their e-learning acceptance in high-tech companies. Behav. Inf. Technol. **33**(1), 51–64 (2014)
33. Ives, B., Walsh, K.R., Schneider, H.: The domino effect of password reuse. Commun. ACM **47**(4), 75–78 (2004)

34. Jang-Jaccard, J., Nepal, S.: A survey of emerging threats in cybersecurity. J. Comput. Syst. Sci. **80**(5), 973–993 (2014)
35. Khern-am-nuai, W., Yang, W., Li, N.: Using context-based password strength meter to nudge users' password generating behavior: a randomized experiment (2016)
36. Kolb, N., Bartsch, S., Volkamer, M., Vogt, J.: Capturing Attention for Warnings about Insecure Password Fields - Systematic Development of a Passive Security Intervention (2014). https://doi.org/10.1007/978-3-319-07620-1_16
37. Komanduri, S., Shay, R., Kelley, P.G., Mazurek, M.L., Bauer, L., Christin, N., Cranor, L.F., Egelman, S.: Of passwords and people: measuring the effect of password-composition policies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2595–2604. ACM (2011)
38. Lange, R.V., Tiggemann, M.: Dimensionality and reliability of the Rotter IE locus of control scale. J. Pers. Assess. **45**(4), 398–406 (1981)
39. Li, J., Lepp, A., Barkley, J.E.: Locus of control and cell phone use: implications for sleep quality, academic performance, and subjective well-being. Comput. Hum. Behav. **52**, 450–457 (2015)
40. Lumpkin, J.R.: Validity of a brief locus of control scale for survey research. Psychol. Rep. **57**(2), 655–659 (1985)
41. Madan, P., Srivastava, S.: Investigating the personality variable (LOC) & impression management relationship: exploring the role of demographic variables & sectoral difference of managers. OPUS **7**(1), 52–71 (2016)
42. Marks, L.I.: Deconstructing locus of control: Implications for practitioners. J. Couns. Dev. **76**(3), 251–260 (1998)
43. McElroy, J.C., Hendrickson, A.R., Townsend, A.M., DeMarie, S.M.: Dispositional factors in internet use: personality versus cognitive style. MIS Q. **31**, 809–820 (2007)
44. Michaelian, K., Sutton, J.: Memory. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2017). https://plato.stanford.edu/archives/sum2017/entries/memory
45. Ng, T.W., Sorensen, K.L., Eby, L.T.: Locus of control at work: a meta-analysis. J. Organ. Behav. **27**(8), 1057–1087 (2006)
46. Norman, D.: The Design of Everyday Things: Revised and Expanded Edition. Basic Books AZ, New York (2013)
47. Norman, D.A.: Emotional Design: Why We Love (or Hate) Everyday Things. Basic Civitas Books, New York (2004)
48. Norman, D.A.: THE WAY I SEE IT when security gets in the way. Interactions **16**(6), 60–63 (2009). https://doi.org/10.1145/1620693.1620708
49. Nurse, J.R.C., Legg, P.A., Buckley, O., Agrafiotis, I., Wright, G., Whitty, M., Upton, D., Goldsmith, M., Creese, S.: A critical reflection on the threat from human insiders – its nature, industry perceptions, and detection approaches. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 270–281. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07620-1_24
50. Payne, J.W.: Contingent decision behavior. Psychol. Bull. **92**(2), 382 (1982)
51. Pilar, D.R., Jaeger, A., Gomes, C.F., Stein, L.M.: Passwords usage and human memory limitations: a survey across age and educational background. PLoS One **7**(12), e51067 (2012)
52. Renaud, K., Volkamer, M., Maguire, J.: ACCESS: describing and contrasting. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 183–194. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07620-1_17
53. Rotter, J.B.: Generalized expectancies for internal versus external control of reinforcement. Psychol. Monogr.: Gen. Appl. **80**(1), 1 (1966)

54. Rotter, J.B.: Some problems and misconceptions related to the construct of internal versus external control of reinforcement. J. Consult. Clin. Psychol. **43**(1), 56 (1975)
55. Sheffer, C., MacKillop, J., McGeary, J., Landes, R., Carter, L., Yi, R., Jones, B., Christensen, D., Stitzer, M., Jackson, L., Bickel, W.: Delay discounting, locus of control, and cognitive impulsiveness independently predict tobacco dependence treatment outcomes in a highly dependent, lower socioeconomic group of smokers. Am. J. Addict. **21**(3), 221–232 (2012)
56. Shugan, S.M.: The cost of thinking. J. Consum. Res. **7**(2), 99–111 (1980)
57. Silver, D., Jana, S., Boneh, D., Chen, E.Y., Jackson, C.: Password managers: attacks and defenses. In: USENIX Security Symposium, pp. 449–464 (2014)
58. Spector, P.E.: Development of the work locus of control scale. J. Occup. Organ. Psychol. **61** (4), 335–340 (1988)
59. Tam, L., Glassman, M., Vandenwauver, M.: The psychology of password management: a tradeoff between security and convenience. Behav. Inf. Technol. **29**(3), 233–244 (2010)
60. Unsworth, N., Fukuda, K., Awh, E., Vogel, E.K.: Working memory and fluid intelligence: capacity, attention control, and secondary memory retrieval. Cogn. Psychol. **71**, 1–26 (2014)
61. Vu, K.L., Cook, J., Bhargav-Spantzel, A., Proctor, R.W.: Short- and long-term retention of passwords generated by first-letter and entire- word mnemonic methods. In: Proceedings of the 5th Annual Security Conference, Las Vegas, NV, pp. 1–13 (2006). https://doi.org/10.15417/1881
62. Vu, K.L., Proctor, R.W., Bhargav-Spantzel, A., Tai, B., Cook, J., Schultz, E.E.: Improving password security and memorability to protect personal and organizational information. Int. J. Hum Comput Stud. **65**(8), 744–757 (2007). https://doi.org/10.1016/j.ijhcs.2007.03.007
63. Wallston, B.S., Wallston, K.A., Kaplan, G.D., Maides, S.A.: Development and validation of the health locus of control (HLC) scale. J. Consult. Clin. Psychol. **44**(4), 580 (1976)
64. Watkins, S.G.: An introduction to information security and ISO27001. IT Governance Pub (2008)
65. Wu, T.D.: A real-world analysis of Kerberos password security. In: NDSS (1999)

Forget the Password: Password Memory and Security Applications of Augmented Cognition

Nancy Mogire[1(✉)], Michael-Brian Ogawa[1], Randall K. Minas[2], Brent Auernheimer[3], and Martha E. Crosby[1]

[1] Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822, USA
{nmogire, ogawam, crosby}@hawaii.edu

[2] Shidler College of Business, University of Hawaii at Manoa, 2404 Maile Way Suite E601f, Honolulu, HI 96822, USA
rminas@hawaii.edu

[3] Computer Science Department, California State University, Fresno, CA 93740, USA
brent@csufresno.edu

Abstract. Individual security behavior plays a central role in achieving secure computing. However, secure usage is difficult to guarantee in an open-ended context where different users have different perceptions of security as well as different cognitive loads when using security tools. In designing secure systems, it is not only necessary to define secure behavior but also to provide built-in support for such behavior in order to enable users to be complaint. In this work, we explore the viability of augmented cognition as a modality that can be used to support security-oriented behavior in authentication systems. Specifically, we explore how transformations of password character properties such as font and weight can improve password recall and recognition and reduce insecure habits, such as writing down passwords. In a previous study, we tested the accuracy of recall and recognition in an augmented password system. The system was designed to make use of character property transformations to minimize the need for complex passwords while not compromising security. Here we repeat the study, incorporating the use of neurophysiological measures to study human physiological responses during recognition and recall of character sets with different types of transformation. The results suggest that cognitive effort in recall of complex passwords can be alleviated with the performance of the augmented password task. This finding has important implications for future research.

Keywords: Augmented cognition · Recall · Recognition · Password memory Physiological measures · Authentication · Cybersecurity

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 133–142, 2018.
https://doi.org/10.1007/978-3-319-91467-1_11

1 Introduction

The pervasive presence of information technology has increased the information processing demands on the human cognitive system, elevating information intake, requiring faster assessment and more accurate decision-making [1]. On the other hand, technology is yet to sufficiently augment human cognition to meet the new expectations. Password authentication is a perfect case for this imbalanced scenario between cognitive processing expectations and human ability. Password authentication falls under the "what you know" paradigm of authentication which is commonly used due to lower cost and ease implementation compared to other methods. Password authentication is dependent on two cognitive processes namely recall and recognition [2]. Cabeza et al. [3] conducted an experiment to determine which parts of the brain are activated by recognition and recall processes. Their experiment involved the use of positron emission tomography (PET) to take measurements while participants recognized or recalled previously studied word pairs mixed with words not previously studied. They found that both processes caused observable activations in regions of the brain, with activation levels in the frontal cortex being indistinguishable between the two processes.

Recall and recognition are demanding on the human cognitive system and users typically have trouble remembering good passwords [4]. The result of this problem is that users typically choose weak passwords for easier memorization. For example, Florencio and Herley [5] conducted a study covering half a million accounts on the Internet over a period of 3 months. They found that when the participants were unconstrained, the majority created passwords that contained only lower case letters leaving out uppercase letters, digits, and special characters (modifications that make passwords more difficult to crack). The passwords they created represented an average bit strength of 40.54 bits, making them easy to crack in dictionary attacks. Additionally, the average password is reused in at least six different websites. In an attempt to improve password memory, different memory training techniques have been used including password rehearsal games [4].

Melby-Lervåg and Hulme [6], in their meta-analysis of memory training studies, concluded that memory training programs can yield reliable improvements on both verbal and nonverbal working memory tasks. However, these effects were likely to be short-term with minimal near-transfer effects. Near-transfer effects are those reflected in tasks similar to the ones in the training program [6]. Given the difficulties involved in improving cognitive ability, methods that aid the user's cognitive processes during the use of a system might be more reliable for promoting secure password behavior.

In this paper, we explore the use of augmented cognition strategies for the purpose of aiding recognition and recall of passwords. Specifically, we study the use of character transformations to make passwords unique and memorable. We examine the extent to which character transformations make the passwords easier to recognize and recall. This strategy would also expand the password space within shorter passwords and hence reduce the need for longer and more complex ones that are difficult to use. In addition, we examine the neurophysiological correlates of password recognition and

recall using electroencephalography (EEG). We conclude with future avenues for augmented cognition work in password recognition and recall.

2 Related Work

2.1 Password Security Habits

Several studies have shown that typical computer users' password habits do not promote security [5, 8, 9]. Florencio and Herley [5] conducted a study covering half a million accounts online over a period of 3 months and found that majority of users when not restricted, chose passwords that contained only lower case letters leaving out uppercase letters, digits, and special characters. In another study conducted within a university setting, results showed that majority of the surveyed users used uppercase letter and numbers in their password formulation but typically used the numbers at the end of the string [9]. As the authors discuss, numbers placed at the end of the password string make it easier for attackers to build password cracking dictionaries, and 60% of the passwords they studied were breakable within days. Interestingly, this study found that there was a correlation between user self-rated security awareness and password strength. What they found was that users tend to overrate their security awareness [9]. The authors interpreted this as potentially resulting from misunderstanding of security concepts [9].

As Alohali et al. [10] found in their study on human security behavior, several factors influence the likelihood that a user will practise good security habits. These factors include: cultural perception of security, technical savviness, awareness of security risk, security tool usability, past experiences with security incidents, individual cognition models and the user ability to understand risk, as well as personality traits and attitudes towards risk messages [10].

2.2 Information Overload and Cognition

Human actions can be understood in terms of dual process cognition with most of our actions being completed with little or no cognition (i.e., Kahneman's System 1 cognition) [12]. The study of cognitive sciences helps define what the human brain can do as well as where humans excel and the extent to which cognition can be influenced [11]. Password recall and recognition could be described in the context of problem solving which has been found to impose cognitive load on the memory [14].

Woods and Siponen [7] in their contextualized metamemory theory, suggest that the password recall problem may not be solely an indication of poor memory performance, but may also be caused by users' inaccurate perception of their own memory. In their study on password recall, memory performance, and metamemory, they found no significant correlation between memory performance and correct password recall. They also did not find a correlation between digit span and correct, immediate, or long-term recall. They argue that users who believe they have more memory capacity or that they have more control over remembering their passwords have a better password correct recall rate. These findings may mean a solution to password recall problems lies in

designing password systems that give more control to a user, and increase their motivation to memorize passwords and influence their expectations they will be able to remember passwords.

Either way, humans behavior eventually suffers from cognitive limitation as suggested by cognitive load theory [13]. The cognitive limitations of users, in turn, influences security behavior [15]. Specifically, awareness of cognitive limitations (or perhaps overestimation of them) influences security behavior [7]. In order to make passwords easier to remember, password users may select characters and words that are familiar to them such as their own names or birth dates. In cases where a password system enforces rules that lead to passwords that are not easily memorable, users may resort to writing them down on paper so as to be able to refer later. In addition, once a password is learned, the password owner may tend to reuse it on as many systems as possible. As it turns out, poor password choices are not by themselves an indication of security ignorance but often a result of the pursuit for simplicity and ease of use of computing systems [16]. People tend to create easier passwords and reuse them so they do not fall victim to their cognitive limitations.

Due to these realities, cognitive load is a critical factor to consider in information security design. For information system users, a high information security workload comes into direct conflict with system usability and results in less security-oriented behavior [17]. Albretchsen [17] found that while awareness campaigns on their own did not cause a change in behavior, users were more motivated towards security conscious behavior when using a system that actively engaged them in the security process. Dorneich et al. [1] defined a framework for mitigating cognitive bottlenecks in systems by employing augmented cognition design. They identified cognitive bottlenecks as those factors that made it difficult for the user to process information in a correct and timely manner. These bottlenecks include excess information, lower information processing ability than the computer they are interacting with and function misallocation where the user is left with a task they are not cognitively able to handle [1]. Augmented cognition models allow for redesign of the information processing interaction in a way that lessens cognitive load and allows the user's preparation to be better aligned to the cognitive task requirement.

Augmented cognition has been defined as "a form of human-systems interaction in which a tight coupling between the user and computer is achieved via physiological and neurophysiological sensing of a user's cognitive state. "This paradigm leverages knowledge of cognitive state to precisely adapt user-system interaction in real time [18].

2.3 Improving Password Habits Using Augmented Cognition

In the context of password creation, memory and usage, one of the main cognitive bottlenecks is function misallocation [1]. Restrictively directing users into creating passwords that combine characters in complex ways leads to misallocation of the information processing tasks between the computer and the user. The user is then left with an enormous task of recalling complex passwords in order to keep the system secure. The adjustable autonomy approach suggested in Dorneich et al. [1], indicates a situation where the user would have more leeway to define their own cognitive task.

As it relates to password use, this would mean the option to use a simpler password as needed. However, simpler passwords present a security risk as they can be easily cracked using rainbow tables and dictionaries [9]. A possible middleground is an augmented password system that utilizes character transformations to expand the password space and hence reduce the complexity requirement on the user. Such a system is discussed in Mogire et al. [2]. The design is aimed at making the password usage easier both by reducing the need for character complexity and by increasing memorability. Albretchsen [17], in a qualitative study conducted to interpret participants' experiences of information security, found that while awareness campaigns on their own did not change behavior, users were more motivated towards security conscious behavior when using a system that actively engaged the user in the security design. The process of character transformation engages the user more actively in the creation of the password and could potentially increase their ability to recall it later.

2.4 Augmented Passwords: A Study

In 2017, a study was conducted to test the cognitive experience in an augmented password system [2]. Specifically, the study was designed to determine accuracy in recognition and recall of passwords using different font styles. 150 participants were recruited from a large-enrollment introductory computer science course at a research extensive university.

For the study, a six-character string password was used. The second character of the password string was modified for the different groups. There were five formats of the second character: the non-modified character, bold, italicize, underline, and strikethrough modification. The password was displayed on the projector twice on the first day, then erased. Students were then asked to select the password from a group of similar password suggestions to assess recognition.

Two days later, students were asked to enter the password from memory to assess recall. After entering the password, a survey was administered to determine the methods used to recall the password.

Results showed that students had a recognition rate of 70% for no font style (plain text), 76% for bold text, 74% for italicize text, 75% for underline text, and 86% for strikethrough text. When asked to recall, students were accurate: 64% for no font style, 93% for bold text, 91% for italicize text, 94% for underline text, 76% for strikethrough text. Performance improved between recognition and recall for bolded, italicized, and underlined text by 17–19%. Conversely, performance decreased between recognition and recall: for no font style (−6%), strikethrough (−9%). Augmented password recall accuracy was higher than the non-augmented passwords.

Due to the higher recognition rate of augmented passwords, the authors believe that font style augmentation helped the participant to focus on the string and accurately replicate it with ease. Similarly, the researchers inferred that augmented characters were more memorable due to distinctiveness and hence had a higher recall rate [2]. Based on the greater recall rate, the authors believe that augmented passwords could be used in practice. An advantage of using the augmented characters is that it expands the password option space of traditional passwords.

However, further investigation is necessary, particularly to observe the neurophysiology of engaging with the augmented passwords. Neurophysiology would prove useful in defining the extent to which character transformations aid the cognition processes of recall and recognition compared to transformations where they are absent. The study below investigates this question by examining the neurophysiological correlates of password recognition and recall.

3 Methodology

3.1 Participant Demographic

This study drew participants from two senior computer science courses at a research intensive university. Participation was voluntary and extra credit was given as compensation for those choosing to participate. Each of the courses had alternative extra credit tasks for participants that did not wish to take part in a study. Nineteen people participated, there were 7 female and 12 male participants.

3.2 Study Task

To determine recognition accuracy, we used a Sakai course management system to create an experiment website that allowed participants to input a six-character string that included either no augmentation or one augmented character. The character string was six characters long (RNGKBV) and included bold, italicize, underline, strikethrough, and no modification. The set included 25 variations that were delivered to participants in a random order with the exception of the first and last questions. These questions were the same for all participants to test for primacy and recency effects in recall. Another design flipped the order of the first and last question to reverse the test for primacy and recency. This created two separate groups for the test. Participants were randomly assigned to each group. At the end the participants were asked to enter the first and last strings in the set of 25.

3.3 Protocol

The participants arrived to the lab and informed consent was obtained. They then sat at the experiment desk. They logged into the university e-learning system and navigated to the study that was set up as a course with participants added to it.

After these initial steps, two standard-size disposable sensors were applied to the forearms to measure heart rate. Two standard-sized disposable sensors were attached to the sole of the participant's foot to measure skin conductance. Two small sensors were applied just above the participant's left eyebrow to measure frown muscle activity. An Emotiv EEG data collection headset was placed over the head to measure brainwave activity. The task began once the psychophysiological data collection sensors were applied.

3.4 Data Cleaning and Preparation

EEG data will be cleaned and analyzed using EEGLab. One limitation of EEG is that cortical bioelectrical activity is extremely small in magnitude when compared to muscle movements across the head. Therefore, participant movement introduces artifacts of high-frequency and magnitude into the EEG data. These will be removed using two methods: EEGlab probability calculations and visual inspection. The EEGLab artifact rejection algorithm uses deviations in microvolts greater than three standard deviations from the mean to reject specific trials. However, additional artifacts are also apparent to the trained eye, so visual inspection of trials is essential in artifact removal.

Electrodermal and facial EMG data were aggregated to mean values per second using BIOPAC's AcqKnowledge software. Change scores will be calculated by subtracting the physiological level at the onset of each target statement during the online discussion from each subsequent second across a 6-s window.

3.5 ICA Analysis of EEG Data

Initially, an EEGLab ICA performs a Principal Components Analysis (PCA). At each electrode site the program assesses which of the other electrode sites account for the most variance in the signal. Taking these weighted values it then relaxes the orthogonality constraint of PCA to isolate individual components of activation [19, 20]. Each ICA component then represents a pattern of activation over the entire brain, not solely the activity present at a specific electrode. The number of independent components (ICs) depends on the number of electrodes in the dataset, as the algorithm is working in an *N*-dimensional space (where N is the number of electrodes). Most participants in the current study are expected to generate 14 distinct ICs, since our recording device has 14 electrodes [20].

Finally, using the *K*-means component of EEGlab the independent components at the individual level will be grouped into clusters containing similar components using procedures recommended by Delorme and Makeig [20]. This procedure clusters similar ICs based upon their latency, frequency, amplitude, and scalp distribution. Relevant clusters will be identified and a time-frequency decomposition will be performed to examine changes in event-related desynchronization of the alpha rhythm.

4 Results

The results show increased alpha attenuation during the password augmentation task, which in turn led to better recall of the password. 11 participants correctly recalled the password asked at the end of the task, indicating the augmented password worked for most participants in invoking memory of the password. Significant alpha attenuation was observed during the recognition task (Fig. 1) with participants showing increased cognition while performing the augmentation task as compared to the recall task. These findings suggest that password recall can be aided by augmenting the password characters.

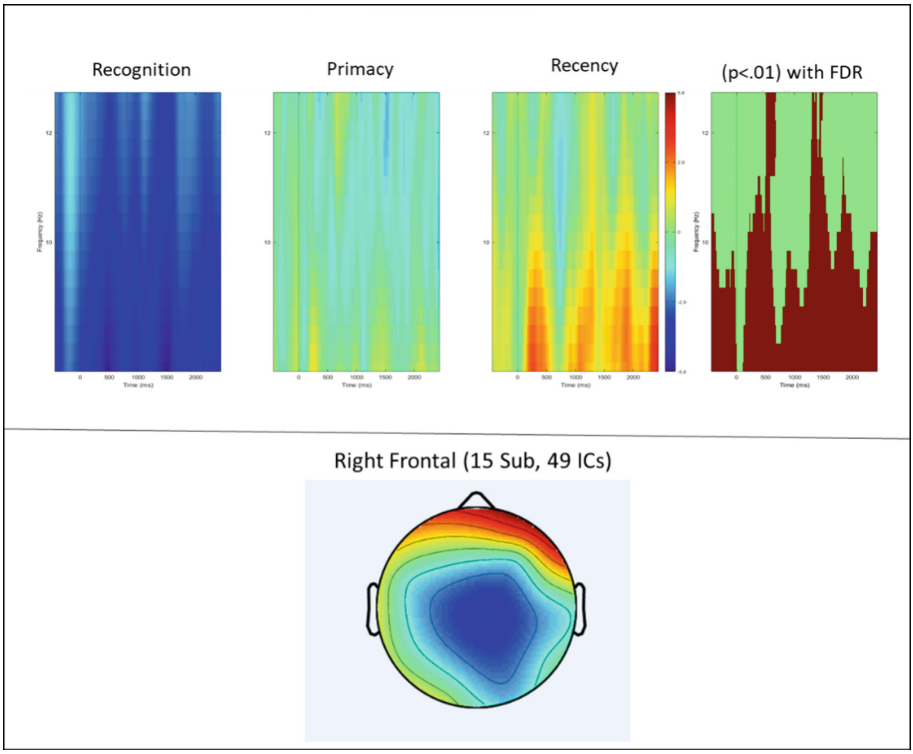

Fig. 1. Significant differences in the alpha band for recognition and recall (primacy and recency effects). Results indicate the password augmentation task increased cognition compared to the recall task.

We observed more alpha attenuation for the primacy than the recency recall condition across the participants measured (Fig. 1). This shows the participants required more cognition during recall of the first password augmentation than the last. Both recall conditions required less cognitive effort than the recognition task, indicating an alleviation of cognitive load during recall of a complex password. This finding indicates that this password augmentation task can aid in alleviating cognitive effort required for recall of complex passwords and could be useful in future applications.

5 Conclusion

Passwords still play an important role in security and poor user habits such as choosing easy-to-guess passwords, are driven by their desire to keep computer usage simple. Augmenting passwords strings using character font transformations can help improve users' cognition and memory of password. This can in turn motivate and promote security-conscious behavior in the password authentication process.

References

1. Dorneich, M., Whitlow, S., Ververs, P., Rogers, W.: Mitigating cognitive bottlenecks via an augmented cognition adaptive system. In: SMC 2003 Conference Proceedings, 2003 IEEE International Conference on Systems, Man and Cybernetics, Conference Theme - System Security and Assurance (Cat. No. 03CH37483) (2003)
2. Mogire, N., Ogawa, M.-B., Auernheimer, B., Crosby, M.E.: Augmented cognition for continuous authentication. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10284, pp. 342–356. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_27
3. Cabeza, R., Kapur, S., Craik, F.I.M., McIntosh, A.R., Houle, S., Tulving, E.: Functional neuroanatomy of recall and recognition: a PET study of episodic memory. J. Cogn. Neurosci. **9**(2), 254–265 (1997). https://doi.org/10.1162/jocn.1997.9.2.254
4. Forget, A., Chiasson, S., Biddle, R.: Lessons from brain age on password memorability. In: ACM, New York (2008). https://doi.org/10.1145/1496984.1497044
5. Florencio, D., Herley, C.: A large-scale study of web password habits. In: ACM, New York (2007). https://doi.org/10.1145/1242572.1242661
6. Melby-Lervåg, M., Hulme, C.: Is working memory training effective? A meta-analytic review. Dev. Psychol. **49**(2), 270–291 (2013). https://doi.org/10.1037/a0028228
7. Woods, N., Siponen, M.: Too many passwords? How understanding our memory can increase password memorability. Int. J. Hum.-Comput. Stud. **111**, 36–48 (2018)
8. Klein, D.V.: Foiling the cracker: a survey of, and improvements to, password security. In: Proceedings of the 2nd USENIX Security Workshop (1990)
9. Awad, M., Al-Qudah, Z., Idwan, S., Jallad, A.: Password security: password behavior analysis at a small university. In: 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) (2016)
10. Alohali, M., Clarke, N., Furnell, S., Albakri, S.: Information security behavior: recognizing the influencers. In: 2017 Computing Conference, London, pp. 844–853 (2017)
11. Watanabe, K.: Explicit and implicit aspects of human cognition and behavior. In: 2017 9th International Conference on Knowledge and Smart Technology (KST) (2017)
12. Posner, M.I., Snyder, C.R.R.: Attention and cognitive control. In: Solso, R.L. (ed) Information Processing and Cognition: The Loyola Symposium, pp. 55–85. Lawrence Erlbaum Associates, Hillsdale (1975)
13. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81–97 (1956). https://doi.org/10.1037/h0043158, PMID 13310704
14. Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. **12**, 257–285 (1988)
15. Conklin, A., Dietrich, G., Walz, D.: Password-based authentication: a system perspective. In: Hawaii International Conference on System Sciences, vol. 7, pp. 1530–1605 (2004)
16. Horcher, A., Tejay, G.: Building a better password: the role of cognitive load in information security training. In: 2009 IEEE International Conference on Intelligence and Security Informatics (2009)
17. Albrechtsen, E.: A qualitative study of users' view on information security. Comput. Secur. **26**, 276–289 (2007)
18. Stanney, K., Schmorrow, D., Johnston, M., Fuchs, S., Jones, D., Hale, K., et al.: Augmented cognition: an overview. Rev. Hum. Factors Ergon. **5**, 195–224 (2009)

19. Onton, J., Makeig, S.: Information-based modeling of event-related brain dynamics. In: Christa, N., Wolfgang, K. (eds.) Progress in Brain Research, vol. 159, pp. 99–120. Elsevier (2006)
20. Delorme, A., Makeig, S.: Eeglab Wikitutorial, May 2012. http://sccn.ucsd.edu/wiki/PDF:EEGLAB_Wiki_Tutorial

Designing and Evaluating Reporting Systems in the Context of New Assessments

Diego Zapata-Rivera(✉), Priya Kannan, Carol Forsyth, Stephanie Peters, Andrew D. Bryant, Enruo Guo, and Rodolfo Long

Educational Testing Service, Princeton, NJ 08541, USA
dzapata@ets.org

Abstract. The effective communication of assessment results to the intended audience is an important issue that has implications for accomplishing the goals of an assessment. New assessments can provide score report users with a variety of additional evidence about the test taker's knowledge, skills, and abilities, than has been possible with traditional assessments. Two audience-specific score reporting systems for highly interactive assessments are currently being developed to provide formative feedback for teachers. The first system provides formative feedback to preservice teachers based on their performance teaching a group of virtual student avatars in a simulated classroom. The second system provides teachers with information relevant to how students interact with a conversation-based assessment. These two score reporting systems provide us with good examples of the types of communication and interaction issues that are present in the development of new types of assessments. In this paper, we describe these two reporting systems, discuss commonalities between the two systems particularly focusing on the design and evaluation processes, and elaborate on the implications for future work in this area.

Keywords: Reporting systems · New assessments · Teachers

1 Introduction

Clear communication of assessment results to the intended audience contributes to the appropriate use of assessment information. The Standards for Educational and Psychological Testing [1] contain several guidelines on score reporting issues including the need to provide clear explanations of assessment results, evidence to support interpretations for intended purposes, information about recommended uses, and warnings about possible misuses.

The literature in this area includes guidelines and iterative development frameworks for designing score reports [2–5]. These iterative frameworks usually include activities such as gathering assessment needs from and evaluating score reports with the intended audience. Zapata-Rivera and Katz [6] apply audience analyses to design score reports based on the needs, knowledge and attitudes of the audience.

Score reports for traditional assessment types typically include assessment results at the individual level (e.g., total scores, subscores, performance levels, task-level results, and recommendations for follow-up activities), class level (e.g., roster of individual

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 143–153, 2018.
https://doi.org/10.1007/978-3-319-91467-1_12

results, distribution of scores and performance levels), school level (e.g., distribution of scores per grade and subject) and district level (e.g., aggregate data across schools, and subgroups of students). These results are usually accompanied by introductory and ancillary materials aimed at helping the intended audience make sense of the assessment results embedded in the score report.

New assessments, such as video-based, simulation-based, and conversation-based assessments, can provide users with a variety of novel assessment information. New assessments can gather evidence of students' knowledge, skills and abilities (KSAs) derived from several sources including student responses to predefined questions and process data [7]. In general, the effective communication of assessment results to the intended audience is critical to the validity of the assessment [6]. Therefore, it is important to carefully consider the types of feedback information that would be most appropriate and useful for these new assessment types. Moreover, it is important to carefully consider the feedback needs for formative assessment tasks, which tend to be woven into instruction, and are therefore intended to provide teachers with ongoing feedback about their students' current level of understanding [8, 9].

In this paper, we describe two audience-specific score reporting systems for highly interactive assessments that are being developed to provide formative feedback for teachers. These two score reporting systems provide us with good examples of the types of communication and interaction issues that are present in the development of new types of assessments. We discuss commonalities between these systems focusing on mainly design and evaluation processes mainly, and elaborate on the implications for future work in this area.

2 Formative Feedback for Preservice and Teacher Educators Using a Simulated Classroom

The first system [10] provides formative feedback to preservice (i.e., student) teachers (PSTs) based on their performance teaching a group of virtual student avatars in a simulated performance-based task [11]. A user-based needs assessment constitutes the first step in an iterative multistep process typically recommended for score report development [2–5]. Following this recommendation, the first system was designed specifically to cater to the feedback needs of teacher educators and PSTs when formatively embedding simulated performance-based tasks into science and mathematics elementary methods courses. In these simulated tasks, administered multiple times during a methods course, PSTs are provided an opportunity to learn to facilitate high quality discussions among virtual student avatars within a simulated classroom environment. For feedback to be effective in a formative context, teacher educators (the teachers, in this context) should be able to diagnose gaps in PSTs' (the students, in this context) current learning to modify instruction and PSTs should be able to understand their strengths and weaknesses to improve their performance [8].

Therefore, with the specific goals of developing a system that is primarily intended to inform instructional practice and guide ongoing learning, we designed a preliminary prospective score reporting (PSR) system [3] and used this PSR system to identify audience-specific score reporting needs from teacher educators and PSTs. The feedback

elements incorporated within this PSR system were informed by previous research [5, 10] and through pilot testing. Among other features, the PSR system developed for the teacher educators included the ability to interactively score videos of PSTs' performances, provide written and annotated video-based feedback to PSTs, and view summary level reports of the whole-class and individual student performance. The PSR system developed for PSTs included the ability to view teacher feedback for each dimension, and the ability to view and annotate one's own video to respond and discuss with the teacher educator. Example screenshots of the scoring and reporting functionalities within this PSR system are presented in Figs. 1, 2, 3 and 4. Focus groups with the relevant stakeholder groups (i.e., teacher educators $N = 8$ and PSTs $N = 5$) were carried out to identify additional stakeholder-specific needs. With this goal in mind, participants in the focus group studies responded to usability questions that included comprehension and preference questions, and identified a prioritized list of user-specific needs.

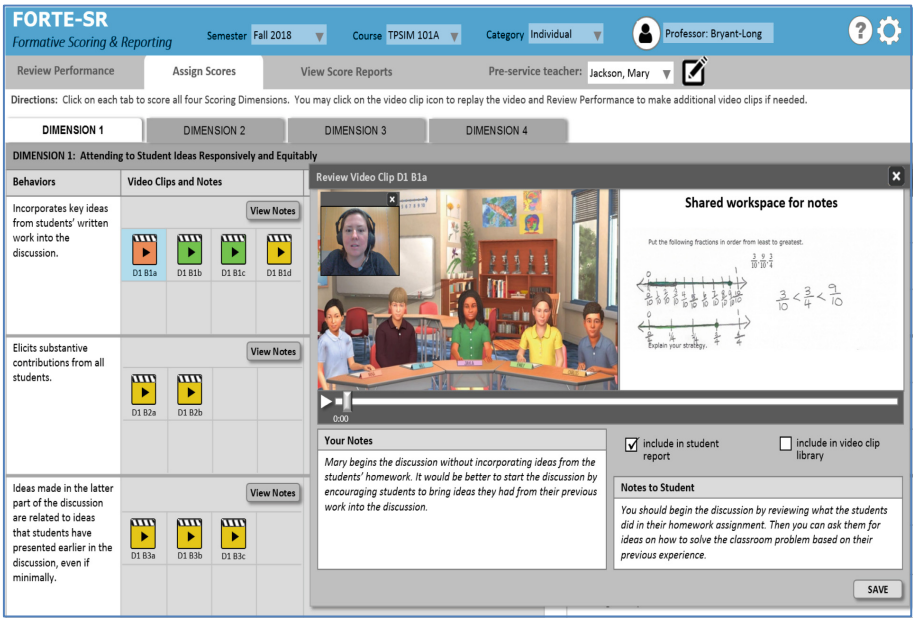

Fig. 1. Example screenshot of scoring functionality with video snapshots of preservice teacher performance that serves as evidence for the score assigned for each behavior.

Results [8] from the focus groups indicated that, in general, all participants reacted positively to the preliminary mockups and had some insightful suggestions for revisions (e.g., include benchmark performances; annotate visual representations; include the ability for PSTs to self-evaluate and respond to feedback). As suggested in the literature (e.g., [12, 13]), teacher educators reiterated that they favored qualitative annotated feedback directed for focused improvement during the first two

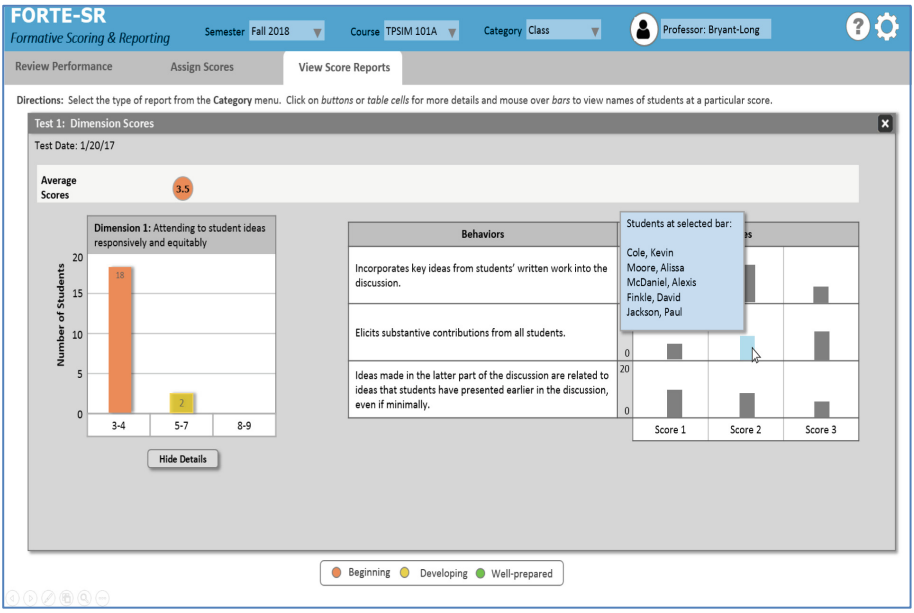

Fig. 2. Example screenshot of detailed class-level feedback on one dimension for one administration that shows the distribution of students across the score points.

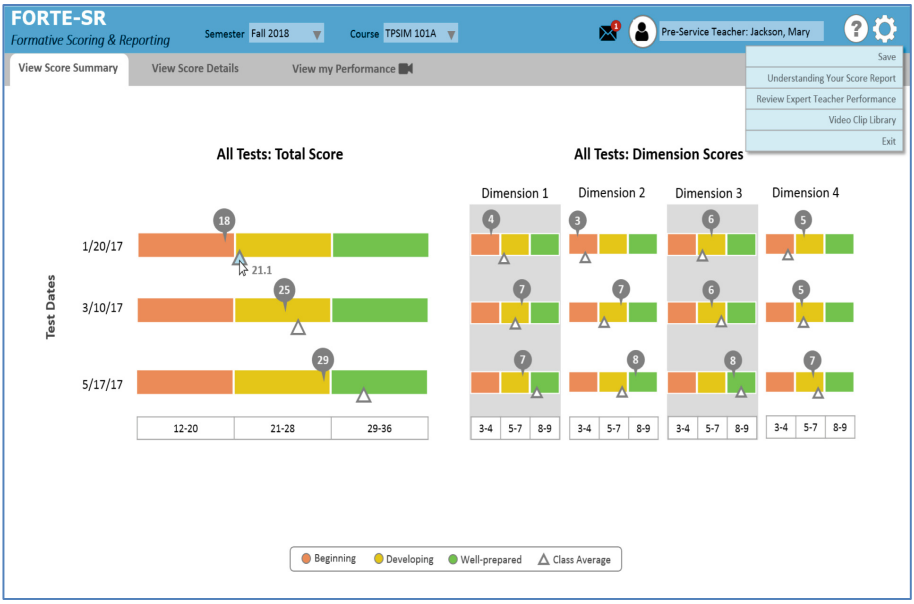

Fig. 3. Example screenshot of preservice teacher report that shows total score and dimension scores across three test administrations.

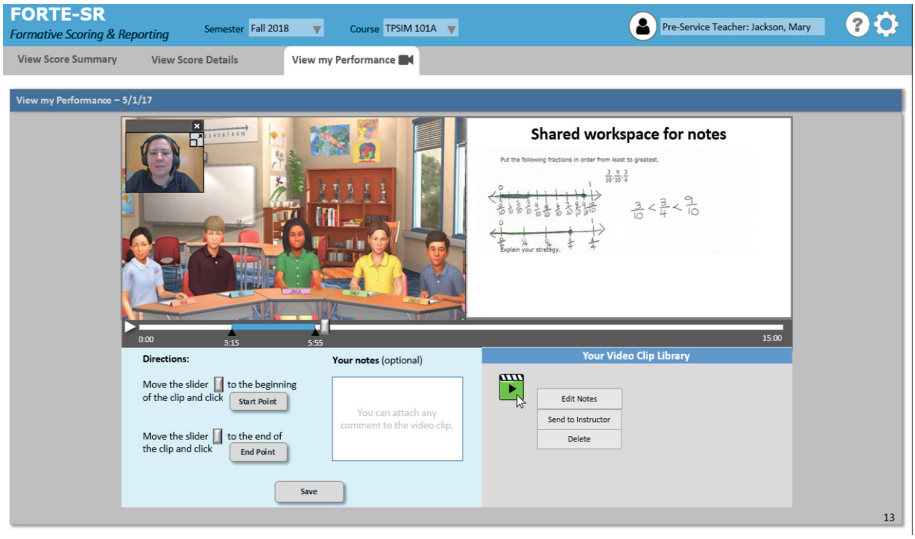

Fig. 4. Example screenshot of preservice teacher report that allows the PST the ability to annotate their own video and use as a discussion tool.

administrations rather than providing quantitative scores. One interesting suggestion from these focus groups was the idea of providing support for self-reflection – both teacher educators and PSTs thought it would be a good idea for the PSTs to first evaluate and annotate their own performance using a teacher-provided check-list before the teacher educator scores and provides feedback on these performances.

Following the focus groups, needs generated from both groups (teacher educators and PSTs) were reviewed and duplicate needs were consolidated. Ultimately, 32 needs for the teacher educators and 19 needs for the PSTs were identified; of these, 16 needs were common across the two stakeholder groups, and therefore, included on both lists. The relative importances of these identified needs were then confirmed through a post-meeting survey. The top 10 needs that emerged for both teacher educators and the PSTs will be considered first in our prototype revisions (if they were not already incorporated in the preliminary mockups) during subsequent phases of the iterative report development cycle.

3 Feedback for Teachers on Students' Interaction with Conversation-Based Assessments

The second system prototype provides teachers with information relevant to how students interact with conversation-based assessments [14]. This system also includes a dialogue-based tutorial aimed at teaching teachers about measurement error and how to make informed decisions based on this concept [15].

The PSR system was created following an iterative design process (See Fig. 4). Specifically, we created mock score reports for teachers to showcase unique features of

conversation-based assessment (CBA) such as ability of students to discriminate between correct and incorrect answers, number of words generated, and amount of scaffolding received. These features were chosen because they are positively correlated with learning [16–18]. We created score reports for various domains including English language learning and assessment (ELLA) and science.

After developing the initial prototype, we refined the score reports across three iterations of teacher focus groups through the process of creating the mockup prototype-> gaining teacher feedback-> reviewing the feedback-> iteratively refining mockup prototypes. In total we conducted 2 focus groups for each domain. Teachers participating in the focus groups included 7 ELLA-mathematics teachers, and 5 science teachers. Three versions of the prototypes were created based on the teachers' feedback. Questions for the focus group addressed comprehension and preference issues.

We discovered from the focus groups various representations that could be more helpful to teachers such as using color bars rather than levels in some instances and providing links to items and incorporating a conversational tutor. Overall, we discovered that teacher feedback was extremely important in creating a score report that meets the needs of teachers and thus may be used to help students.

Figures 5, 6 and 7 show screenshots of the mockup prototypes at a late stage of the process. Assessment results based on both student responses and process data were included in the reports. Figure 5 shows a description of one of the features (number of words). These descriptions are aimed at facilitating teachers' understanding of and appropriate use of the information in the report. Figures 5 and 6 correspond to individual student reports for teachers in the ELLA domain and Fig. 7 in the science domain.

We then developed a conversation-based tutor to better help teachers understand measurement error because teachers who participated above-mentioned focus groups agreed that this type of support would be helpful in interpreting overall score information. Initial results from a pilot study ($N = 6$) suggest that this tutor was well-received by teachers but may need alterations in the dialogic framework to account for teacher answers which are different from common student responses [15]. Specifically, in a pilot study ($N = 8$), teachers answered 62.8% of questions about the system in a positive fashion. In regards to the dialogic framework, teachers gave elaborate responses that were extremely close to the correct answer. However, these responses were judged to be correct by the tutor but only partially correct by two human raters (with interrater reliability of 90.8% agreement). This phenomenon of teachers providing extremely close but not completely correct answers may be more unique to teachers as the framework has been successful in categorizing student responses in other systems.

4 Discussion

In this section we discuss several design and evaluation aspects of score report systems in the context of new assessments.

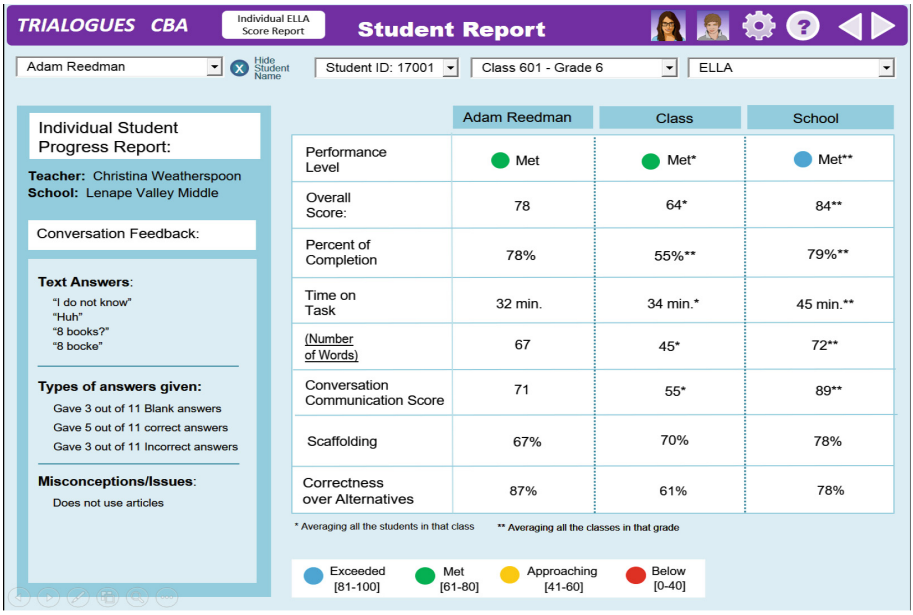

Fig. 5. Example screenshot of an individual student report for teachers in the context of an English language CBA. The report includes performance level information, overall scores and features extracted from process data at different levels (student-, class-, and school-level). (Color figure online)

- *An iterative, audience-centered approach.* These systems followed an audience-centered development and evaluation approach. Various mockups of the system were used to gather information about needs for assessment results in reports. These assessment needs are usually captured in a prospective score report (PSR) [3] that is used throughout the assessment development process by an interdisciplinary group of experts. The PSR can take the form of a paper-based mockup report or a reporting system mockup (or PSR system). As these PSR systems get iteratively refined based on the results of studies with experts and the intended audience, they can be used as communication tools to show different stakeholders the changes made to the original report design, and iteratively refine the information that would be included in the operational reports.
- *Communicating assessment results based on student response and process data.* A clear alignment of the purpose of the reporting system and the types of claims and assessment information in the reporting system is essential for the creation of reporting systems that (a) provide the right type of information needed by the intended audience, and (b) support appropriate use of assessment information. The examples presented in this paper show how the purpose of the assessment (i.e., provide formative feedback for teachers) guided design and evaluation decisions. Results of the studies carried out as part of each project (e.g., focus groups, cognitive labs and usability studies) showed that teachers appreciated the type of

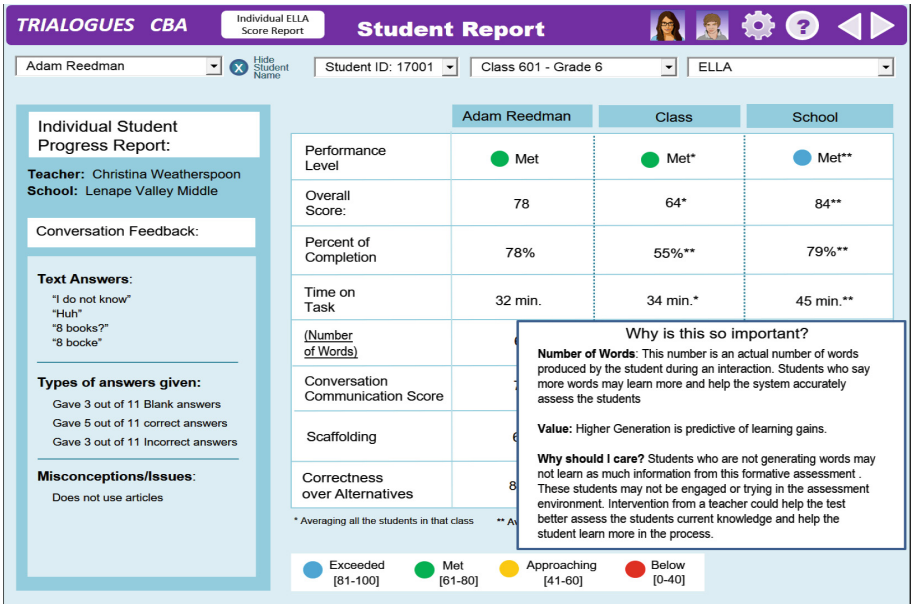

Fig. 6. Example screenshot of an individual student report for teachers showing an explanation for number of words. (Color figure online)

information provided by the system. This information was not limited to total or subscores but also included feedback for PSTs from teacher educators based on a predefined rubric, in the case of the simulated classroom, and general linguistic features based on student performance across tasks, in the case of the CBA system.

- *Evaluating comprehension and preference aspects*. Both comprehension and preference aspects of the reports should be part of the evaluation plan. Preferred representations are not necessarily better at supporting comprehension of assessment information [19–21]. Both, comprehension and preference questions were included in the questionnaires used to evaluate the reporting systems described above.
- *Supporting mechanisms.* When evaluating reporting systems, needs for additional support can be identified. In some cases, providing additional information of the meaning of particular features and how they can be used is enough to help the audience understand and appropriately use assessment results. However, in some cases, additional supporting mechanisms such as video tutorials or dialogue-based tutors are necessary to teach challenging concepts [15, 21].

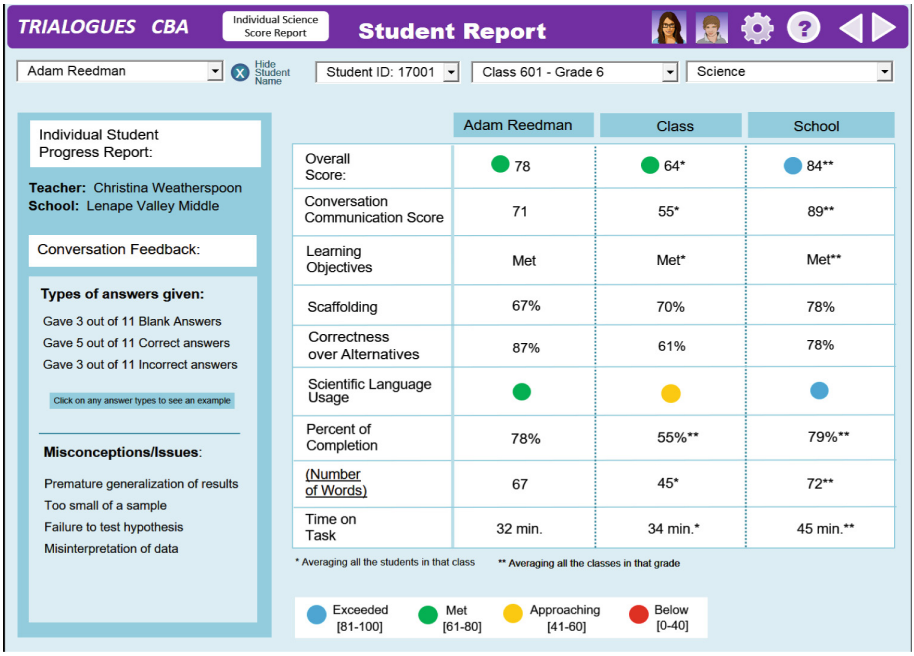

Fig. 7. Example screenshot of an individual student report for teachers in the context of a science CBA. (Color figure online)

5 Future Work

Future work in this area includes designing and evaluating new feedback features. These features may involve the implementation of a continuous feedback loop in which teacher educators can support PSTs while improving their teaching skills using the simulated classroom. Also, providing teachers with video segments of students interacting with characters in the CBA.

In addition, we would like to make improvements to the conversation-based tutor based on the data collected and using this tutor to help teachers understand the concept of measurement error.

References

1. American Educational Research Association (AERA): American Psychological Association (APA), & National Council on Measurement in Education (NCME). Standards for educational and psychological testing, AERA, Washington, DC (2014)
2. Wainer, H., Hambleton, R.K., Meara, K.: Alternative displays for communicating NAEP results: a redesign and validity study. J. Educ. Meas. **36**(4), 301–335 (1999)

3. Zapata-Rivera, D., VanWinkle, W., Zwick, R.: Applying score design principles in the design of score reports for CBAL™ teachers. ETS Research Memorandum RM-12-20. ETS, Princeton (2012)
4. Zenisky, A.L., Hambleton, R.K.: A model and good practices for score reporting. In: Lane, S., Raymond, M.R., Haladyna, T.M. (eds.) Handbook of test development, 2nd edn, pp. 585–602. Routledge, New York (2016)
5. Tannenbaum, R.J., Kannan, P., Leibowitz, E.A., Choi, I., Papageorgiou, S.: Interactive score reports: a strategic and systematic approach to development. In: Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC (2016)
6. Zapata-Rivera, D., Katz, I.: Keeping your audience in mind: applying audience analysis to the design of score reports. Assess. Educ. Princ. Policy Pract. **21**(4), 442–463 (2014)
7. Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., von Davier, A.A.: Assessing science inquiry skills in an immersive, conversation-based scenario. In: Kei Daniel, B. (ed.) Big Data and Learning Analytics in Higher Education, pp. 237–252. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-06520-5_14
8. Black, P., Wiliam, D.: Assessment and classroom learning. Assess. Educ. Princ. Policy Pract. **5**(1), 7–75 (1998)
9. Shepard, L.A.: Formative assessment: Caveat emptor. In: Paper Presented at the 2005 ETS Invitational Conference on The Future of Assessment: Shaping Teaching and Learning, New York (2005)
10. Kannan, P., Zapata-Rivera, D., Bryant, A.D., Long, R.: Providing formative feedback to pre-service teachers as they practice facilitation of high-quality discussions in simulated mathematics methods classrooms. Final Report. ETS, Princeton (2018)
11. Mikeska, J.N., Howell, H., Straub, C.: Developing elementary teachers' ability to facilitate discussions in science and mathematics via simulated classroom environments. In: Paper Presented at the Annual TeachLive Conference, Orlando, FL (2017)
12. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction (KLI) framework: toward bridging the science-practice chasm to enhance robust student learning. Cogn. Sci. **36**, 757–798 (2010)
13. Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008)
14. Peters, S., Forsyth, C.M., Lentini, J., Zapata-Rivera, D.: Score reports for conversation-based assessments: identifying and interpreting evidence. In: Paper Presented at the Annual Meeting of the American Educational Research Association (AERA), San Antonio, TX (2017)
15. Forsyth, C.M., Peters, S., Zapata-Rivera, D., Lentini, J., Graesser, A., Cai, Z.: Interactive score reporting: an autotutor-based system for teachers. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 506–509. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_51
16. VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., Rose, C.P.: When are tutorial dialogues more effective than reading? Cogn. Sci. **31**(1), 3–62 (2007)
17. Forsyth, C.M., Graesser, A.C., Pavlik, P., Millis, K., Samei, B.: Discovering theoretically-grounded predictors of shallow vs. deep- level learning. In: Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (eds.), Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), pp. 229–232 (2014)
18. Swanson, H.L.: Generality and modifiability of working memory among skilled and less skilled readers. J. Educ. Psychol. **84**, 473–488 (1992)
19. Kannan, P., Zapata-Rivera, D., Leibowitz, E.A.: The interpretation of score reports by diverse subgroups of parents. Educational Assessment (in press)

20. Zwick, R., Zapata-Rivera, D., Hegarty, M.: Comparing graphical and verbal representations of measurement error in test score reports. Educ. Assess. **19**(2), 116–138 (2014)
21. Zapata-Rivera, D., Zwick, R., Vezzu, M.: Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. Educ. Assess. **21**(3), 215–229 (2016)

Human Augmentation of UAV Cyber-Attack Detection

Haibei Zhu[1(✉)], Mahmoud Elfar[1], Miroslav Pajic[1], Ziyao Wang[2], and Mary L. Cummings[2]

[1] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
{haibei.zhu, mahmoud.elfar, miroslav.pajic}@duke.edu
[2] Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA
{ziyao.wang, mary.cummings}@duke.edu

Abstract. Unmanned aerial vehicles (UAVs) have extensive applications in both civilian and military applications. Nevertheless, the continued development of UAVs has been accompanied by security concerns. UAV navigation systems are potentially vulnerable to malicious attacks that target their Global Positioning System (GPS). Thus, efficient GPS hacking detection with high success rate is paramount. Significant effort has been put into developing autonomous hacking detection techniques. However, little research has considered how a human operator can contribute to the security of such systems. In this paper, we propose a human-autonomy collaborative approach for a single operator of multiple-UAV supervisory control systems, where human geo-location is used to help detect possible UAV cyber-attacks. An experiment was designed and conducted using the RESCHU-SA experiment platform to evaluate this approach. The primary results show that 65% of all experiment sessions reached over 80% success rate in UAV hacking detection, while only 17% of participants lost one or more UAVs because of incorrect hacking detections. These results suggest that such an approach could help achieve better security guarantees for human-in-the-loop autonomous UAV systems that are prone to cyber-attacks.

Keywords: Unmanned aerial vehicles · Cyber-attack detection
Human geo-location

1 Introduction

Unmanned aerial vehicles (UAVs) have significantly increasing commercial market and extensive applications in both civilian and military realms [1]. Many of these UAVs rely on the Global Positioning System (GPS) for navigation, however, this reliance leaves UAVs vulnerable to malicious attacks targeting GPS signals. One common attack is GPS spoofing, in which attackers deceive GPS receivers to override the navigation systems and redirect UAVs to unexpected destinations [2, 3]. A well-known such incident garnered public attention in 2011 when, a US RQ-170 Sentinel UAV was captured by Iranian forces using GPS spoofing attacks [4]. Thus,

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 154–167, 2018.
https://doi.org/10.1007/978-3-319-91467-1_13

detecting GPS spoofing attacks with a high success rate is important for UAV control systems.

We propose a human-autonomy collaborative approach of human geo-location in that humans can aid in the detection of possible GPS spoofing attacks on UAVs. This approach was evaluated via an experiment, which was designed and conducted using the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) platform. Experiment sessions simulated human supervisory multi-UAV control scenarios with potential UAV GPS spoofing attacks. In this paper, we focus on answering the following questions based on the experiment results: (1) Can human operators successfully identify UAV GPS spoofing attacks? (2) What factors affect human operator general operation? (3) Would hacking detections affect the performance of operators' primary tasks? (4) What types of landmarks used in human geo-location affect operator decisions to hacking detections?

2 Background

A common UAV control scheme is human supervisory control, in which a human operator monitors the multi-UAV system, intermittently navigating UAVs, and conducting other higher-level tasks [5]. The architecture of human supervisory UAV control is shown in Fig. 1. Human supervisory UAV control can be introduced with various level of automation. In this study, we assume that human operators are responsible for higher-level decision, and autonomous systems are in charge of lower-level UAV control and navigation operations [6].

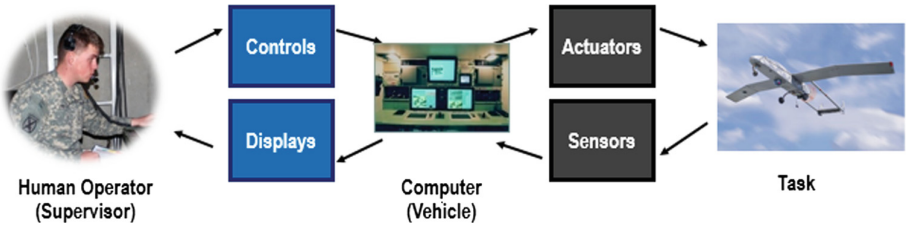

Fig. 1. Human supervisory UAV control architecture.

2.1 GPS Spoofing Detection

UAVs typically rely on an embedded navigation system known as GPS, which provides accurate position, velocity and time information for GPS receivers in most areas on Earth. GPS receivers calculate precise latitude, longitude and height with speed information based on the received satellite signals. Furthermore, GPS receivers can report their locations to UAV control interface to provide location views for operators. However, GPS receivers are vulnerable to GPS spoofing attacks, in which GPS spoofers generate counterfeit signals to attack GPS receivers by manipulating the target position, velocity and time information [2, 3].

Many researchers have presented autonomous GPS spoofing detection methods [7–12], however, false alarms and detection mistakes still exist while applying autonomous detection techniques [13, 14]. Thus, supplementary detection methods are needed.

In the common design of military UAVs, a UAV is usually equipped with both a GPS navigation system and a payload camera, whose signal is independent of the UAV GPS signal [15]. Thus, the UAV payload camera view could be used as an independent reference for detection of GPS spoofing (i.e., navigation based) attacks, which is further explored in the remainder of this paper.

2.2 Human Visual Task

In order to utilize a UAV payload camera to detect UAV GPS attacks, interpreting the UAV real-time location through the camera view and comparing this to a certain landmark or position estimate from a map could be the central mechanism for making such an assessment.

While autonomous localization techniques may have limited performance [16, 17], human vision has advantages in such complex search and surveillance tasks. The process of human vision obtaining information from objects can be divided into two stages. The first stage is the preattentive stage, in which human observers can gather basic information about the target even before the observer become conscious of it [18]. Thus, human vision can process target information relatively fast in complex environment. Human observers also tend to choose areas that maximize information of the target in salience-driven visual search strategy [19], which means human vision has effective strategies to obtain target information. In addition, the direction discrimination threshold of human vision has a low average of 1.8° [20], which means human vision can detect relatively small changes in orientation. Based on these visual advantages, a human operator can potentially aid in UAV localization and thus detect potential UAV GPS spoofing attacks.

Based on the assumption that UAV cameras can show the true surrounding scene of UAVs, we propose that human operators can act as supplementary sensors and assist autonomous system to detect UAV hacking attacks through comparative geo-location between the camera and map position estimates. In human geo-location, the operator can compare the non-tempered video feed coming from the UAV to the potentially falsified GPS location; this allows the user to detect inconsistencies between these two sensing feeds (i.e., whether the feed and the reported locations match). If the operator thinks the location interpreted from camera view does not match the location shown on the map, then the UAV is most likely hacked via GPS spoofing.

An example of applying human geo-location in UAV hacking detection is shown in Fig. 2. If the UAV is hacked, the operator will observe a location other than the GPS reported UAV location through the camera, like shown in the upper left camera view in Fig. 2. When a GPS spoofing attack is discovered, the operator can prevent a hacking event by overriding its physical control.

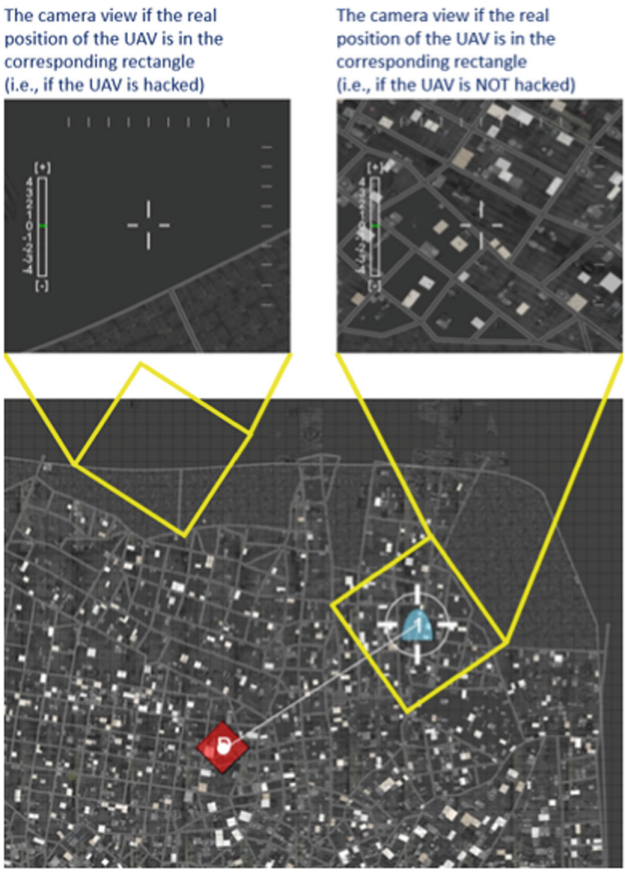

Fig. 2. An example of GPS reported location on the map.

3 Experiment

An experiment was designed utilizing a modified version of the RESCHU experiment platform [21], known as Security-Aware RESCHU (RESCHU-SA) [22]. RESCHU-SA is a Java-based single operator with multi-UAV supervisory control simulation platform, which provides the capability to design multi-tasking scenarios that include both navigational and imagery search tasks. Moreover, the platform allows for simulating GPS spoofing attacks, in which hacked UAVs deviate from their originally assigned path and target to other unexpected destinations, along with warning notifications that simulate autonomous GPS spoofing detection systems.

3.1 Experiment Platform Interface

The interface of the RESCHU-SA platform is shown in Fig. 3. The interface features five main components: the payload camera view, message box, control panel, mission timeline and map area.

- The camera view displays the video stream from the payload camera of the selected UAV. The primary purpose of this view is to conduct real-time image analysis tasks. In this study, it can also be used to determine the actual location of UAVs by locating landmarks.
- The message box displays events that occur during the simulation such as UAV arrival at a target. It also allows operators to communicate the results for the imagery analysis tasks to a "supervisor" that is, in actuality, a bot.
- The control panel provides the UAV damage level, which is caused by UAVs intersecting with hazard areas, as well as instructions for imagery analysis tasks and vehicle updates.
- The timeline shows the estimated remaining time of all UAV arrivals at waypoints and assigned targets.
- The map displays the area of surveillance with real-time locations of all UAVs, hazard areas and targets. For this experiment, the map was created using CityEngine from ArcGIS, a modeling software package that is used for urban planning and architecture design.

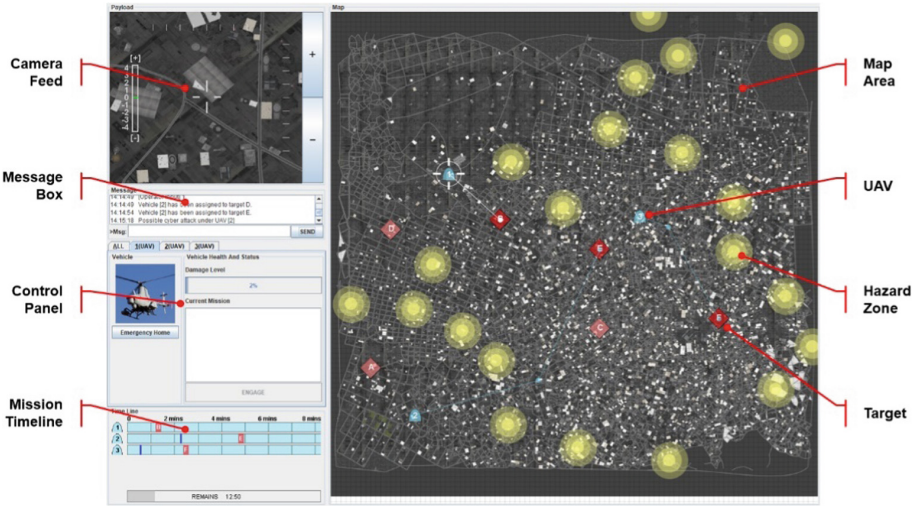

Fig. 3. RESCHU-SA operator interface.

3.2 Experiment Scenarios

The primary objectives for operators in RESCHU-SA are to control multiple UAVs to: (1) perform reconnaissance imagery tasks of counting road intersections when UAVs reach assigned targets, (2) ensure UAVs do not encounter hazard areas, and (3) determine whether UAVs are under GPS spoofing attacks.

For this experiment, GPS spoofing attack events followed a pre-defined schedule, unknown to the participants. When triggered internally, the hacked UAV changed its heading by a random angle within 30 to 60, or 300 to 330, degrees, which was larger than the human direction discrimination threshold. A hacking notification appeared 10 to 20 s after the attack event, simulating an external agency detection of a possible GPS spoofing attack. However, as in real systems, the notification could be a false alarm. In fact, about half of all notifications in this experiment were false alarms in the pre-defined schedule of each test session. Although change blindness and vigilance declination [23, 24] are known problems for humans in such settings, they were not examined in this study, given there was a clear signaling of a potential hacking event. Thus, a detection failure on the part of the human is not considered in this experiment.

Once the operator received notification from the system that a certain UAV was under possible cyber-attack, the operator could then investigate the UAV by checking the UAV's camera view and matching it against the position of the UAV on the map. The operator was expected to make a decision before the hacked UAV either exceeded the map boundary or the experiment ended. If the operator decided the UAV was hacked, the operator could override the hacked UAV and send it home.

When UAVs that were not hacked reached a target, the operator engaged in an imagery task of counting the road intersections from the UAV's camera view at a pre-specified zoom level. This side task represents the primary purpose of such a mission, which is typically information gathering. While engaging in a counting task, the operator was required to enter an answer before the counting task was finished. The counting task allowed us to assess participants' performance based on the number of attempted tasks and the task correctness percentage.

The path planner for the UAVs was intentionally suboptimal, i.e. the planner did not necessarily pick the most efficient assignment of UAVs to targets. In addition, UAVs would possibly encounter hazard areas that appeared and disappeared randomly. The suboptimal planner and the dynamic nature of hazard areas allowed experimenters to assess how much spare attention participants could devote to optimize the navigation and target assignment.

3.3 Experiment Participants

Thirty-six participants took part in this experiment, including 22 males and 14 females. Age ranged from 19 to 34 years with an average of 25.2 and a standard deviation of 3.8 years. Among the participants, 18 had little video game experience, 6 participants had monthly gaming experience, 5 participants played video game several times a week, another 5 participants had weekly gaming experience, and only 2 participants had daily gaming experience.

3.4 Experiment Procedure

The experiment procedure consisted of four main sections. The first section was a self-paced tutorial session, during which participants went over the tutorial slides, and the experimenter answered questions that the participants might have had. The second section was a practice session to allow participants to get more familiar with the user interface. In the first half of the practice session, participants were shown how to operate UAVs and complete all major tasks.

In the second half, participants controlled all UAVs and accomplished different tasks by themselves. The practice session lasted 18 min, which was the same as a single experiment session. The third section included the test sessions with two scenarios of different task loads, which were counterbalanced in terms of order of presentation. The fourth section was the debriefing session, in which the experimenter asked the participant several questions related to participants' performance and strategies for navigating UAVs and detecting hacking events.

Given that many related studies on the RESCHU platform [21, 25, 26] showed a significant impact of task load on system performance, task load was a primary factor in this experiment looking at hacking detection. It should be noted that high task load does not necessarily represent high operator mental workload, since operator mental workload is an individually subjective interpretation of an objective task load.

Thus, for a high task load scenario, operators controlled 6 UAVs with 9 different targets and 9 hacking events, and in each low task load scenario, operators controlled 3 UAVs with 6 different targets and 6 hacking events. In both scenarios, the number of hazard areas, which generated and disappeared randomly, was constantly twenty-one. Each test scenario lasted 18 min, and each participant completed both high and low sessions. Each participant's performance scores were calculated based on the total vehicle damage, the correct percentage of imagery counting tasks, and the correct percentage of hacking identifications.

4 Results

4.1 Performance Statistical Results

We used a multivariate repeated-measures ANOVA model and Pearson correlation with a significance level of 0.05 to analyze data. In data analysis, independent variables included task load, which task load was experienced first, gender and video game experience as a covariate. Task load (low and high) was a within factor variable. Dependent variables included percentage of correct hacking detections, the aggregated damage sustained by vehicles over a test session, and the overall correct percentage intersection counts per test session. These variables represent the primary objectives of performing the counting tasks, keeping vehicles out of the damage areas, and successfully detecting hacking events.

An important question was whether human operators could successfully detect the UAV hacking events. A successful detection was indicated by a correct decision for a specific hacking event, including overriding the UAV and sending it home if the UAV was hacked, or recognizing the notification was a false alarm.

Each high task load experiment session included 9 hacking events, and each low task load session included 6 hacking events. Among all hacking events in both test sessions for each participant, 7 (4 in high task load and 3 in low task load) were predefined as false alarms, which meant the threshold for incorrect hacking notifications was 47%. As shown in Table 1, out of all real hacking notifications across all participants, the overall success rate was 78%, and for the false alarms, the success rate was 84%. In other word, the type one error (false positive, operators considered UAV not hacked with real hacking notification) was 22%, which was slightly higher than the type two error (false negative, operators considered UAV hacked with false alarm notification) of 16%. Thus, operators were slightly better at detecting false alarms than identifying real hacking notifications.

Table 1. The confusion matrix of hacking detection decisions in different notifications.

	Real hacking notification	False alarm notification
Consider UAV hacked	224	40
Consider UAV not hacked	63	207

When looking at each individual's performance per test session, even though they had to multitask in RESCHU-SA in managing multiple vehicles and detecting potential hacking events, results showed that 23 out of total 72 experiment sessions (32%) resulted in 100% of successful hack identifications in a single test session, with another 24 (33%) above 80% successful attack identification. Thus, 65% of total experiment sessions exhibited 80% correct hacking detection or better without having any prior formal training. In terms of incorrect hacking identifications, 12 (17%) participants lost one or more UAVs, meaning that these UAVs were successfully hacked and could not be further controlled.

Additionally, those factors that affected human operators' performance were studied. For the three performance scores of vehicle damage, the correct percentage intersection counts, and correct percentage of hacking events, the only variable affected by task load was vehicle damage ($(F(1, 31) = 32.93)$, $p < 0.001$). In the high task load scenario, the average UAV damage was 31.4, which was much higher than 9.6 in the low task load scenario. Participants with less workload suffered less damage as they had more time to optimize their paths and avoid hostile areas.

One result showed an interesting significant negative correlation between the time expended in hacking detections and correct hacking detections (Pearson = −0.375, $p = 0.001$), which meant that participants who took longer to detect the hacking events had a lower percentage of correct hacking identifications. This suggests that early detection was better from the operator standpoint, which is at odds with those who would argue that longer detection times should yield more correct identifications.

Gender was examined because of the potential difference in self-assessment in cognitive tasks between different genders [27]. However, gender did not affect the participants' general performance. Another covariate, the video game experience, did have a significant effect on participants' correct hacking detections ($F(1, 31) = 4.652$ $p = 0.039$). This means that the more video game experience, the higher the chance of a

correct hacking detection. Not surprisingly, seven participants who lost UAVs had no video game experience, and the other 5 who lost UAVs ranged from some to moderate gaming experience. Participants with daily gaming experience did not lose any UAVs and were 100% correct in hacking identification.

Another result showed that participants' task inputs were effective in that the more time they navigated the UAVs, the less time UAVs intersected with hostile areas (Pearson = −0.345, p = 0.003). This result suggests that improved path planning could reduce operators' workload and free their cognitive resources to attend to other tasks.

We also investigated whether hacking detection affected the performance of operators' primary tasks of counting road intersections. The results showed that the correctness of imagery counting tasks was not affected by either the correctness of hacking detections (Pearson = −0.022, p = 0.854) or the time expended in hacking detections (Pearson = 0.024, p = 0.841). However, time expended in the imagery task was negatively correlated with the percentage of correct hacking detection (Pearson = −0.275, p = 0.019). This result was expected as participants who spent more time in counting tasks were less likely to detect hacking events.

In addition, an interesting observation is that the first experiment scenario affected participants' abilities to correctly finish their primary task of counting the intersections at each target ($F(1, 31) = 5.324$, $p = 0.028$). The participants who had the high task load scenario as the first experiment session tended to have higher correct intersection count percentages. This suggests a fatigue effect since these participants did worse on their second scenarios with low task load, which should have been easier.

4.2 Map Analysis for Hacking Detection

While using human geo-location in UAV hacking detections, operators will compare the non-tempered UAV camera video feed to the potentially falsified GPS location to detect inconsistencies between the UAV video feed and UAV GPS location. After receiving a hacking notification, operators can purposely navigate the notified UAV to some specific areas that can potentially provide more inconsistencies to increase the confidence of making a correct decision to a hacking event. Thus, analyzing the map usage in hacking detections will benefit the future design of autonomous decision supporting tool for hacking identification.

The resulting heat map, which represents the frequency distribution of areas of participant interest during hacking detections, is shown in Fig. 4. Different colors represent varying frequency of operations, including adding waypoints and switching targets for UAVs, on a specific point. The warmer the color, the more participants interacted with a specific point, for example, red represents 5 or 6 operations. The heat map shows that the lower left quadrant is the most popular region, however, some regions, like the middle right of the map, have few operations. Understanding that the density of targets on the lower left quadrant of the map is slightly higher than other regions, this quadrant is more attractive to operators since operators can navigate UAVs between targets to get engaged to more imagery tasks in a shorter time range. Thus, more operations occurred on the lower left quadrant. In addition, red areas on this quadrant indicate the existence of some interesting landmarks that operators tend to investigate during hacking detections.

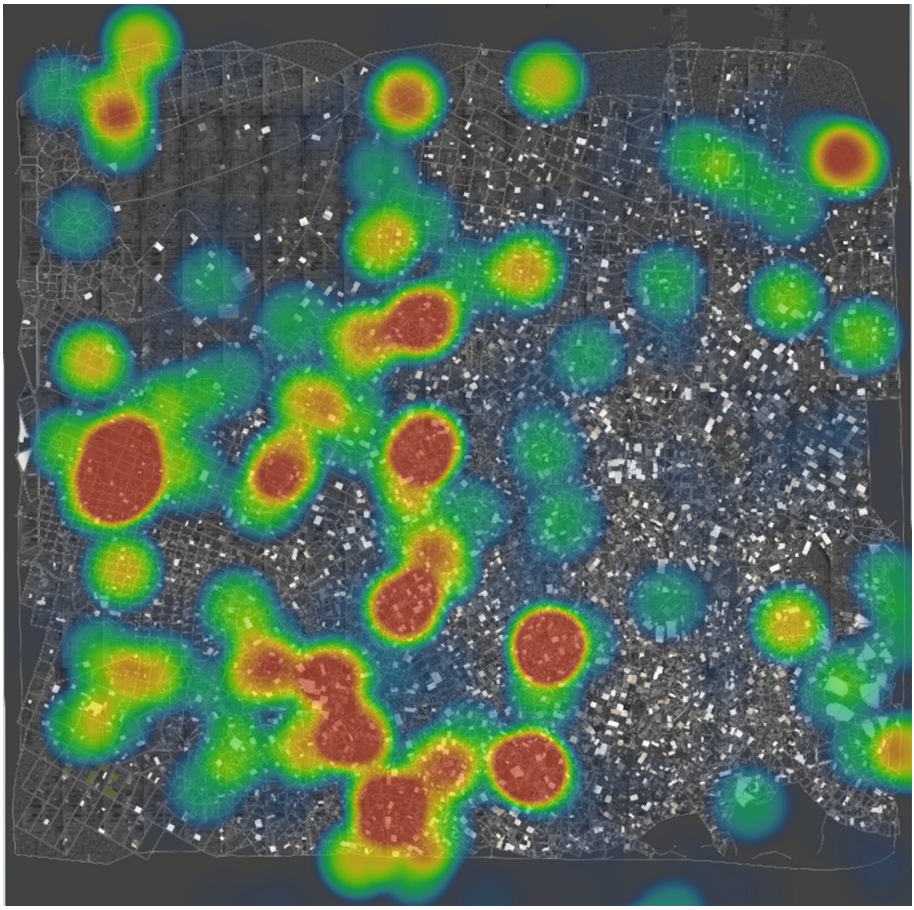

Fig. 4. The heat map of reference points in UAV hacking detection. (Color figure online)

Landmarks used in hacking detections are classified into three categories, including special road patterns, geographic feature transition, and special buildings, like shown in Table 2. Using these different landmarks in hacking detections, operators can investigate the moving orientation of a certain UAV or the relative motion between a UAV and a specific landmark to investigate whether a UAV is potentially hacked. Shown in Table 2, special road patterns were the most frequently used landmarks in hacking detection with an occurrence percentage of 59.3%.

Geographic feature transitions are defined as the transition between land and sea areas, on which operators can clearly observe the sudden change of geographic patterns. Special buildings are defined as distinctive shapes with contrastive colors that are used to represent a single building or a group of buildings on the map. As the percentage of total special road patterns and special buildings are approximate the same, special road patterns are more attractive to operators. Future work will determine why

Table 2. The frequency of different types of landmarks used in hacking detections.

	Special road patterns	Geographic feature transition	Special buildings
Occurrence frequency	380	152	109
Occurrence percentage	59.3%	23.7%	17.0%

Table 3. The frequency of different landmarks used in different detection decisions.

	Real hacking notification		False alarm notification	
	Consider UAV hacked	Consider UAV not hacked	Consider UAV hacked	Consider UAV not hacked
Special road patterns	168	25	28	159
Geographic feature transitions	69	17	14	52
Special buildings	42	13	8	46

exactly people prefer these over other options, but one hypothesis is that these are easier to see than the buildings, and do not take as long to investigate as the sea/land transitions (Table 3).

The frequency of different landmarks used in different detection decisions were examined. In correct hacking detections with both real hacking and false alarm notifications, the percentage of operations based on special road patterns is slightly over 60%, which is higher than the percentage in incorrect hacking detection with real hacking notification (45.5%) and false alarm notification (56.0%). Another interesting fact is that special road patterns lead to the highest success rate of 86.1% in hacking detections, while geographic feature transition lead to 79.6% and special buildings lead to 80.7%. These results provide insight for how a future advanced map-based hacking detection support tool for human operators could be designed.

5 Discussion

The experiment results provide insight into our initial questions with implications for future studies. In this study, we analyzed if a human operator could serve as a supplementary sensor in supervisory UAV control systems by successfully detecting spoofing attacks. Experiment results supported this hypothesis in that 65% of total experiment sessions reached over 80% hacking detection correctness. This result was achieved with no dedicated training and so greater emphasis on optimal search strategies would likely yield even better results.

The experiment results also clearly indicate that some factors affected operators' performance and operations. For example, higher task load tended to cause more UAV damage. This result was supported by a previous study that higher mental workload increased operator attention switching delays [21]. In high task load scenarios, operators tended to experience higher mental workload, which slowed down their attention switches and causing more damage. This could be mitigated in future studies with more optimal path planning as well as better target allocation.

Understanding that the operator's video game experience significantly affected the success rate in hacking detections, future personnel selection strategies for supervisory control systems with human visual tasks could focus more on the experience in similar applications or more training. This fact also raises interesting future research questions, including how video game experience may affect human search strategies and how different types of video games may affect human operators' performance in hacking detection? Also, the result of the negative correlation between the time expended and the success rate in hacking detection provides implications for future studies of increasing hacking detection correctness by guiding better search strategies and earlier detections. However, a fatigue effect was potentially exhibited after just one 18-min scenario, which raises the question of how sustainable such task load levels are over time?

The map analysis shows the heat map of participants preferences for hacking detection. Although the usage percentages of different landmarks in different hacking detection decisions are similar, there was a clear preference for unusual road intersections. These results provide some insights on a more efficient way to utilize different landmarks and raise future research topics of investigating potential different operator hacking detection strategies.

Lastly, all these results establish a baseline of performance of applying human geo-location in UAV hacking detection. Future studies, enabled by an empirical model of security-aware human-autonomy interaction will focus on how higher-level automation or advanced decision support tools could be utilized to assist human operators to improve the success rate of hacking identifications.

6 Conclusion

Navigational GPS systems used in UAVs can be prone to malicious cyber-attacks, especially GPS spoofing attacks. In this study, we have shown that a human operator can assist autonomous systems in hacking detection using human geo-location comparison between maps and downward-facing camera views, even without extensive training. Moreover, we found that an individual factor, video game experience, and the time expended in hacking detection and UAV navigation, affected operators' hacking detection performance. The results from this study indicate that human geo-location is a potentially promising approach for hacking detection, which could be improved by future efforts in improving operator decision support.

Acknowledgements. This work was supported in part by the CNS-1505701 grant. This material is also based on research sponsored by the ONR under agreements number N00014-17-1-2012 and N00014-17-1-2504. We gratefully acknowledge the efforts of those who have assisted in developing the RESCHU-SA experiment platform including Duke undergraduate students, Jeffrey Wubbenhorst, Adithya Raghunathan and Yi Yan Tay.

References

1. Pajares, G.: Overview and current status of remote sensing applications based on unmanned aerial vehicles (UAVs). Photogram. Eng. Remote Sens. **81**(4), 281–329 (2015)
2. Humphreys, T.E., Ledvina, B.M., Psiaki, M.L., O'Hanlon, B.W., Kintner Jr., P.M.: Assessing the spoofing threat: development of a portable GPS civilian spoofer. In: Proceedings of the ION GNSS International Technical Meeting of the Satellite Division, vol. 55, p. 56, September 2008
3. Shepard, D.P., Humphreys, T.E., Fansler, A.A.: Evaluation of the vulnerability of phasor measurement units to GPS spoofing attacks. Int. J. Crit. Infrastruct. Prot. **5**(3), 146–153 (2012)
4. Shane, S., Sanger, D.E.: Drone Crash in Iran reveals secret US surveillance effort. The N. Y. Times **7** (2011). https://www.nytimes.com/2011/12/08/world/middleeast/drone-crash-in-iran-reveals-secret-us-surveillance-bid.html
5. Sheridan, T.B.: Telerobotics, Automation, and Human Supervisory Control. MIT press, Cambridge (1992)
6. Cummings, M.L., Bruni, S., Mercier, S., Mitchell, P.J.: Automation Architecture for Single Operator. Multiple UAV Command and Control. Massachusetts Institute of Technology, Cambridge (2007)
7. Kerns, A.J., Shepard, D.P., Bhatti, J.A., Humphreys, T.E.: Unmanned aircraft capture and control via GPS spoofing. J. Field Robot. **31**(4), 617–636 (2014)
8. Wesson, K.D., Shepard, D.P., Bhatti, J.A., Humphreys, T.E.: An evaluation of the vestigial signal defense for civil GPS anti-spoofing. In: Proceedings of the ION GNSS Meeting, September 2011
9. Broumandan, A., Jafarnia-Jahromi, A., Dehghanian, V., Nielsen, J., Lachapelle, G.: GNSS spoofing detection in handheld receivers based on signal spatial correlation. In: 2012 IEEE/ION Position Location and Navigation Symposium (PLANS), pp. 479–487. IEEE, April 2012
10. Wesson, K., Rothlisberger, M., Humphreys, T.: Practical cryptographic civil GPS signal authentication. Navigation **59**(3), 177–193 (2012)
11. Psiaki, M.L., O'Hanlon, B.W., Bhatti, J.A., Shepard, D.P., Humphreys, T.E.: GPS spoofing detection via dual-receiver correlation of military signals. IEEE Trans. Aerosp. Electron. Syst. **49**(4), 2250–2267 (2013)
12. Pajic, M., Weimer, J., Bezzo, N., Sokolsky, O., Pappas, G.J., Lee, I.: Design and implementation of attack-resilient cyberphysical systems: with a focus on attack-resilient state estimators. IEEE Control Syst. **37**(2), 66–81 (2017)
13. Wesson, K.D., Evans, B.L., Humphreys, T.E.: A combined symmetric difference and power monitoring GNSS anti-spoofing technique. In: 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 217–220. IEEE, December 2013
14. Humphreys, T.E.: Detection strategy for cryptographic GNSS anti-spoofing. IEEE Trans. Aerosp. Electron. Syst. **49**(2), 1073–1090 (2013)
15. Sun, J., Li, B., Jiang, Y., Wen, C.Y.: A Camera-based target detection and positioning UAV system for search and rescue (SAR) purposes. Sensors **16**(11), 1778 (2016)

16. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Trans. Image Process. **14**(3), 294–307 (2005)
17. Blacknell, D., Griffiths, H.: Radar automatic target recognition (ATR) and non-cooperative target recognition (NCTR). The Institution of Engineering and Technology (2013)
18. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980)
19. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
20. De Bruyn, B., Orban, G.A.: Human velocity and direction discrimination measured with random dot patterns. Vis. Res. **28**(12), 1323–1335 (1988)
21. Donmez, B., Nehme, C., Cummings, M.L.: Modeling workload impact in multiple unmanned vehicle supervisory control. IEEE Trans. Syst. Man Cyber.-Part A Syst. Hum. **40** (6), 1180–1190 (2010)
22. Elfar, M., Zhu, H., Raghunathan, A., Tay, Y.Y., Wubbenhorst, J., Cummings, M.L., Pajic, M.: Platform for security-aware design of human-on-the-loop cyber-physical systems. In: Proceedings of the 8th International Conference on Cyber-Physical Systems, p. 93. ACM, April 2017
23. Simons, D.J., Ambinder, M.S.: Change blindness: theory and consequences. Curr. Dir. Psychol. Sci. **14**(1), 44–48 (2005)
24. Temple, J.G., Warm, J.S., Dember, W.N., Jones, K.S., LaGrange, C.M., Matthews, G.: The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. Hum. Factors **42**(2), 183–194 (2000)
25. Boussemart, Y., Cummings, M.L.: Predictive models of human supervisory control behavioral patterns using hidden semi-Markov models. Eng. Appl. Artif. Intell. **24**(7), 1252–1262 (2011)
26. Boussemart, Y., Cummings, M.L., Las Fargeas, J.C., Roy, N.: Supervised vs. unsupervised learning for operator state modeling in unmanned vehicle settings. JACIC **8**(3), 71–85 (2011)
27. Pallier, G.: Gender differences in the self-assessment of accuracy on cognitive tasks. Sex Roles **48**(5), 265–276 (2003)

Augmented Learning and Training

Mitigating Skill Decay in Military Instruction and Enemy Analysis via GIFT

Michael W. Boyce[1,2(✉)], Jeanine A. DeFalco[1,2,3], Robert C. Davis[4], Erik K. Kober[4], and Benjamin Goldberg[2]

[1] Army Research Laboratory, West Point, NY, USA
{michael.w.boycell.civ, jeanine.a.defalco.ctr}@mail.mil
[2] Army Research Laboratory, Orlando, FL, USA
benjamin.s.goldberg.civ@mail.mil
[3] Oak Ridge Associated Universities, Oak Ridge, USA
[4] Department of Military Instruction, United States Military Academy, West Point, NY, USA
{Robert.Davis4, Erik.Kober}@usma.edu

Abstract. This paper will review the collaborative effort of the Army Research Laboratory (ARL) and the Department of Military Instruction (DMI) at the United States Military Academy (USMA) to develop and execute a pedagogical pathway to validate the efficacy of mitigating skill decay in the content area of enemy analysis by way of the Generalized Intelligent Framework for Tutoring (GIFT). An identified area of concern in educating USMA cadets in the area of military preparedness, mitigating skill decay in military instruction from year to year is a necessary, yet time consuming task that is susceptible to inconsistent reinforcement. This paper will provide an overview on the progress of developing an empirically validated course that can be used to offset skill decay while supplementing DMI with reusable and consistent content, with the flexibility to provide adaptable content and assessments that is unique to the GIFT platform.

Keywords: GIFT · Skill decay · Military instruction · Enemy analysis

1 Introduction

1.1 Overview

The adaptive training research program strives to provide innovative instructional practices to facilitate and enhance the delivery and assessment of learners. At the center of this program is intelligent tutoring via the Generalized Intelligent Framework for Tutoring (GIFT), which, as an Intelligent Tutoring System (ITS), can provide tailored learning content based on learner proficiencies that are assessed through a course [19]. In establishing needs for adaptive tutoring systems to better support military tasks, Sottilare [18] notes that ITSs need to acquire learner data, assess learner state, and select an optimal instructional strategy to meet the learners' need. Applied to the military domain, the problem space is often complex, ill-defined, and highly subjective in nature. Therefore, this project seeks to work closely and collaboratively with military

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 171–182, 2018.
https://doi.org/10.1007/978-3-319-91467-1_14

educators such that the capabilities of GIFT can serve as an augmentation rather than a replacement for existing instructors.

1.2 Background

Within the Army, one of the most significant locales to connect with military educators is the United States Military Academy at West Point (USMA). The adaptive training research program has a long-standing relationship with USMA, and has recently developed a modeling and simulation cell located physically at West Point and staffed by two adaptive training scientists. This provides opportunities to better engage and integrate with the West Point students, faculty, and staff for a mutually beneficial relationship of meeting ARL and USMA goals.

1.3 Collaboration Efforts

Since the interest of this project was the complex military domain, going to the educators who teach military instruction was a natural pairing. The Department of Military Instruction (DMI) aims to provide the necessary military training to ready cadets from a military perspective. Most instructors within the department are active military personnel and as such, the department experiences a frequent turn-over rate. While the benefit of this is that cadets receive instruction fresh off the field, the limitation of this is that instructional design and learning objectives often lack necessary scaffolding and reinforcement from year to year. Therefore, there has been a growing consensus as to the importance of developing a GIFT course that not only encapsulates necessary domain knowledge about military tactics necessary for military preparedness, but is flexible enough that can be adaptable and flexible for authoring modifications by instructors from year to year.

As such, after the development of a successful capstone working with cadets and GIFT to assess squad and platoon level military tactics [1], there was a desire to examine how technologies like GIFT could further impact military instruction. It is from this, that a new partnership was forged to provide the first steps in the use of GIFT in the military instruction curriculum.

Members from the Army Research Laboratory's (ARL) adaptive training research program began to have collaborative meetings with the departmental faculty of DMI. The goal of these meetings was to help identify a specific area where functionality such as that which exists in GIFT would be able to assist instructors in delivery of their classes. The aim was to find a problem and task which this combined team could investigate together. From the very beginning of these discussions, it was clear that retention was an area that could use some focus. With lesson content in military instruction being spaced across several semesters with summer training breaks in between, there is ample opportunity for cadets to forget content and therefore necessitating refresher training frequently as a part of classes. An area of interest for military instruction is augmenting existing classroom activities such that information is retained better by cadets, reducing the need for refresher training and improving assessment scores.

An additional challenge faced by the military learner is the amount of content expected to be learned. The amount of information is due to the structure of the course and the utilization of scaffolding (i.e., progressing learner knowledge incrementally with more and more complex problem sets), precisely mirroring the format of the Army Operations Order. If a learner (cadet) does not engage the content and instructor at the onset and remain engaged throughout, it will have a negative aggregate effect on their overall course grade. During the Mission Analysis portion of the course, the learners are learning three specific content areas: (1) the conceptual framework (explained later in the document); (2) the placement of the Mission Analysis input into the Operations Order; and (3) the language or formatting of language to generate content to into the Operations Order. One specific aspect of mission analysis is enemy analysis, which is the focus of this research.

The use of GIFT to support this augmenting existing classroom activities is specifically relevant to DMI and the Military Science (MS) instruction series due to changes in the way the class periods are being restructured. The course, which is 1.5 credit hours, has traditionally been organized in 40 class periods with 55 min per period, equating to 3600 min over the course of the semester. This is being modified for the 2019 academic year to reflect 30 class periods at 75 min per class period. It is documented in the literature that there is a correlation between the amounts of time a learner is exposed to content, and the proficiency in accomplishing a task [6].

As mentioned above, there is a large breadth of material that instructors must cover to achieve ultimate course outcomes. These outcomes are represented in the progression for a learner from the previous course (MS100), to the current course (MS200), to the following courses (MS300 and Cadet Leadership Development Training - CLDT). The ability to have an increase in depth of content, and an increase in number of repetitions to facilitate better retention via conventional pedagogical techniques does not currently exist, which drives DMI to look for Live, Virtual, Constructive, and technologically advanced means (like intelligent tutoring via GIFT) to increase the efficacy and customize the learning experience for each student. The approach of using GIFT in this application within a military classroom can be described as an epistemic process of the assessing and understanding individual student, identifying where the problem areas or gaps in knowledge-building exist, and reinforcing them with a technologically efficient means. GIFT is able to support the conjunction of the vast amount of information needed to be learned and the need to retain it over long periods of time, minimizing skill decay of enemy analysis concepts.

2 Skill Decay and Enemy Analysis

2.1 Identification of the Problem: Skill Decay

Skill decay occurs when skills are not used and the ability to execute suffers. The amount of decay varies in accordance to the task and their dependence on cognitive and psychomotor information elements [8, 17]. Deterioration of performance is further compounded if skills have not been reinforced or have been newly learned.

Performance is often determined by the level of experience someone has and how frequent they are trained on the task of interest [6].

Major dimensions that are often discussed in skill decay research include: length of non-use period, how much overlearning occurred, characteristics of the skill, testing methods of previous learning, type of retrieval, method of training, individual differences, and motivation [7, 8, 10, 15, 20]. There is well-established research that has shown a direct relationship to the rate at which forgetting occurs and the amount of controlled rehearsal associated with the task [14]. Additionally, skill decay contributes to loss of confidence in performing a necessary skill [7].

2.2 Refresher Interventions

One of the ways to minimize skill decay is the use of refresher interventions. Refresher interventions are techniques that assist in re-attaining skill proficiency after it was lost due to skill decay. It has been shown that different refresher interventions effect skill and knowledge retention differently [10].

Symbolic rehearsal is one such refresher intervention, defined as the visualization of a task without actually performing the task [10, 11]. Practice problems consist of examples of the course material applied to actual learning scenarios. Since symbolic rehearsal has been shown in previous studies [9, 10] to be as effective as practice for knowledge retention but not for skill retention, recent research has proposed including process visualization tasks to provide a procedural component as well [11]. To this end, the domain area of military instruction is well suited to determine whether symbolic rehearsal will mitigate skill decay, within the subdomain of enemy analysis.

2.3 Conceptual Overview of Enemy Analysis

Enemy analysis is an area that has been identified by USMA instructors as being particularly susceptible to skill decay for cadets. Enemy analysis is a complex topic because it requires a multistep process to be successful.

Enemy analysis requires the learning of many different symbolic aspects of the battlefield and representation of the enemy, as well as incorporating other elements, such as terrain and weather, as elements that need to be woven into a cohesive narrative as a part of their assignment, which is essentially a simulated briefing. This briefing is executed by cadets after approximately four weeks of instruction in the form of an operations order, which is graded and assessed by their instructors.

The briefing task revolves around building an operations order to support enemy analysis. The operations order contains information on maneuver, which provides information on each the enemy unit's task and purpose. The second part of the operations order consists of purposes, priorities, allocation of, and restrictions for fire support. It also contains information on intelligence, supplies, commander's intent, and protection. These are known as the warfighting functions. And it is these warfighting functions that cadets must not only achieve content mastery and retention to accomplish their classroom task, but they must carry this domain knowledge with them as they proceed to their next year, next level of military instruction coursework.

There are 6 warfighting functions: mission command, movement and maneuver, intelligence, fires, sustainment, and protection:

1. *Mission Command*: The mission command warfighting function that allows a commander to balance the art of command and the science of control in order to integrate other warfighting functions.
2. *Movement and Maneuver*: The movement and maneuver warfighting function moves and employs forces to achieve a position of relative advantage over the enemy and other threats.
3. *Intelligence*: The intelligence warfighting function assists in understanding the enemy, terrain, and civil considerations.
4. *Fires*: The fires warfighting function provide collective and coordinated use of Army indirect fires, air and missile defense, and joint fires.
5. *Sustainment*: The sustainment warfighting function provides support and services to ensure freedom of action, operational reach and prolong endurance.
6. *Protection*: The protection warfighting challenge preserves the force so the commander can apply maximum combat power to accomplish the mission.

According to Army Doctrine Reference Publication (ADRP) 3.0, warfighting functions are related tasks and systems united by a common purpose or objective that allow commanders to accomplish mission or training goals [5].

2.4 Assessment of Enemy Analysis

Currently, the enemy analysis content module spans over three lessons coinciding with three formative assessments and one summative assessment. The first assessment, cadets are provided a snapshot of an enemy squad on a given area of terrain with a specified task and purpose. Cadets are individually assessed on their ability to provide a doctrinal template of how that squad is comprised (also known as composition and depicted in the form of a line-wire diagram) via numbers of personnel and equipment.

Additionally, they have to demonstrate their understanding of how to graphically depict the same information on a map with four required components: breakdown of the squad-sized element into two team-sized elements; provide a task and purpose for each team, demonstrating that they are mutually supportive of one another (adhering to the concept of nesting) and graphically depicting the coinciding tactical mission task for each respective element; a depiction of the key weapon systems of each team and providing their proper orientation (i.e. RPG Variant, RPK, and PKM); and all labeling adhering to the standard as prescribed by ADRP 1-02 [4].

During the second assessment, they are paired in groups of two, given an entire mission analysis prompt (including terrain, light & weather, and enemy analysis) and asked to conduct analysis on each respective input, brief it to the instructor in front of their peer cohort, and receive constructive feedback (with a weighted grade) in an attempt to allow them a complete deliberate repetition prior to receiving the mid-term assessment with a 20% course grade weight value. Finally, with the current course design, the content transitions to offensive operations during the latter half of the semester with a culminating summative assessment that encapsulates enemy analysis, forcing cadets to demonstrate retrieval.

3 Collaborative Methodology

3.1 Understanding the Problem Space

After a series of meetings and sharing of documents between ARL researchers and USMA instructors, qualitative observations were conducted in CPT. Robert Davis's cadet classes. From this step in the methodology, we elicited the declarative knowledge elements associated with the content. For the purposes of enemy analysis, these include the baseline understanding of four fundamental tenets: composition, disposition, strength, and capabilities, which are defined below:

1. *Composition*: describes how an entity is organized and equipped - essentially the number and types of personnel, weapons, and equipment.
2. *Disposition*: refers to how threat/adversary forces are arrayed on the battlefield/battlespace. It includes the recent, current, and projected movements or locations of tactical forces.
3. *Strength*: is described in terms of personnel, weapons, and equipment. The most important aspect of strength when evaluating a regular force is to determine whether the force has the capability of conducting specific operations.
4. *Capabilities*: an analysis that must determine what the enemy is capable of doing against a friendly platoon during the mission. Such an analysis must include the planning ranges for each enemy weapon system that the platoon may encounter.

Members of the ARL research team took this content and translated it into a basic prototype GIFT course to demonstrate proof of concept (see Fig. 1).

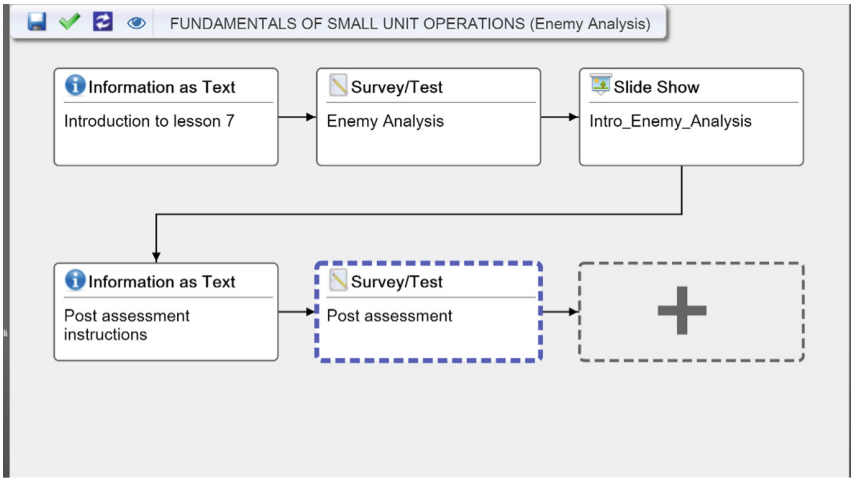

Fig. 1. Example of declarative knowledge course flow in GIFT

3.2 Understanding the Technology Space

The next step in the methodological process was determining how actual content was being delivered as part of a lesson. To accomplish this, qualitative observations were conducted in CPT. Davis's class. During this evaluation we identified the key concepts, domain knowledge, and learning objectives of enemy analysis, where these observations helped identify common patterns of misconceptions as well as help requests that cadets articulated during class time. The purpose of these observations was to help guide the design of the GIFT course in a more dynamic and adaptive manner.

In its current form, GIFT provides adaptive lesson capabilities based on individual differences (e.g., prior knowledge, motivation, grit, etc.) and real-time embedded assessments. GIFT applies a domain-agnostic pedagogical approach that is based on David Merrill's Component Display Theory (CDT [12]), where content is structured in a way to support the presentation of material (i.e., in the form of general rules of a domain and specific instances of those rules applied as seen by an example) and the assessment of material through dedicated question banks that are configured on a concept by concept basis. This theoretical approach was used in the design of GIFT's Engine for Management of Adaptive Pedagogy (EMAP [3]), which uses the CDT as the framework by which an instructor configures lesson material, with the design goal that the EMAP can apply across all notional cognitive problem domains.

In the context of enemy analysis, GIFT's EMAP provides the framework for an instructor to establish specified content (i.e., Rules and Examples) they want to present on a concept by concept basis, along with the tools to establish assessment questions and scoring parameters that drive performance outcomes. What this development supports is a closed-loop remediation model, where each individual cadet will receive a personalized experience based on their learner model and on the outcomes of the GIFT managed assessments. Following an assessment event, GIFT will either progress a trainee on to the next lesson activity, if all scoring thresholds were satisfactory, or GIFT will initialize a remediation loop that targets the specific concepts that scored below expectation. This model uses a focused-coaching strategy that targets only the concepts that require further instruction, so as to key in on each individual's strength and weaknesses.

An important note linked to the EMAP is its dependency on assessments to drive personalized remediation loops. For the initial implementation, the enemy analysis assessments in GIFT will utilize question bank approaches to infer the domain concepts that require further attention. In future iterations, there is room to incorporate more focused scenario-based exercises that leverage simulation environments. These practice events extend the assessment space by enabling more focused scenario-based exercises that focus on application of skill, rather than recall of knowledge.

Key Challenges Identified. In the first round of observations, CPT Davis engaged his cadets in discussions about the six warfighting functions. Importantly, CPT Davis had cadets work in groups to discuss one of the six warfighting functions, including identifying and utilizing information from the ADRP 3.0 doctrine, and describing in layman's terms what the information meant for enemy analysis [5]. What this activity demonstrated was not only that there should be a consideration for having a

functionality in GIFT for collaborative work, but more importantly to consider providing an opportunity for cadets within GIFT to teach each other. This pedagogical approach to learning is well aligned with dialogic teaching along the tradition of Dewey [2] and Vygotsky [21] where students are involved in the collaborative construction of meaning and share control over classroom discourse [16].

Observations also yielded information that there should be a consideration in the GIFT course to incorporate "help" functions in the form of hyperlinks or sidebar content for particularly complex ideas. This "help" function would be well aligned with the adaptive functionality of GIFT in recognition that not all learners need additional reinforcement and help in the same way and with the same frequency. Lastly, observations revealed the importance of incorporating within the enemy analysis GIFT course the need to include dynamic graphical supports such as maps and key legends. These two elements could serve a dual function: as a symbolic reinforcer as well as an additional assessment instrument.

Ultimately the final design of the Enemy Analysis course will be an iterative process that includes using key learning objectives articulated by DMI, a course design informed both by Merrill's Component Display Theory and educational psychology principles, where the content validity is established by current DMI instructors. Once this phase of the project is executed and validated, a longitudinal empirical study will be conducted to evaluate how refresher interventions improve the retention of military instruction compared to traditional learning methods.

4 Experimental Design

For the purposes of the anticipated experiment, three different cadet sections will be divided into three experimental groups across three time periods: Initial Training (IT), Refresher Intervention (RI), and Retention Assessment (RA) with two weeks of spacing in between – spanning a time from the summer before the year of instruction to the end of the fall semester. The same cadets will be tracked through all three time periods (See Fig. 2).

The experiment will examine how refresher interventions using GIFT can impact cadet performance. The two refresher interventions that will be manipulated will be practice problems (condition 1) and symbolic rehearsal (condition 2). The control condition will not include any refresher interventions.

To further support these refresher interventions in conditions 1 and 2, feedback will also be given via GIFT. One of the consistent findings through ITS research is that it is important to provide training activities that offer refresher interventions that also provide feedback tailored to the student [13]. Learning outcomes will be measured via a pre-posttest design, to determine if the refresher interventions had any impact on skill decay in conditions 1 and 2, independently in the IT, RI, and RA time periods, and longitudinally from IT to RA. Repeated measures Analysis of the Variance (ANOVA) will be used to analyze statistical significance of outcomes and interaction of trait metrics.

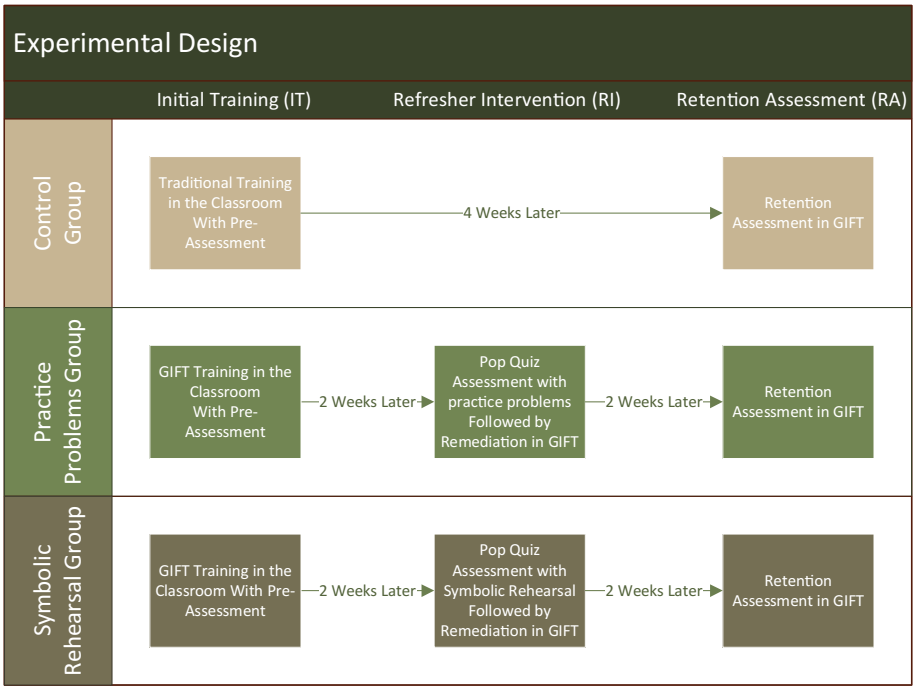

Fig. 2. Experimental design

4.1 Procedure

To help provide a more concrete idea of how the experiment would be executed, we have created a hypothetical procedure that the experiment would follow:

1. The cadet enters class and opens up a GIFT course, which is presented via the Web.
2. The cadet is shown large amounts of declarative knowledge elements in the form of bolded terms that the cadet will be asked to pay attention to.
3. The cadet will be provided with several example scenarios explaining how the knowledge elements are used. To make the task more challenging, more terminology will be interspaced in the text of the example scenarios.
4. In all conditions, a pre-assessment survey will be administered to establish baseline knowledge.
5. In the control condition, the cadet would receive traditional instructor-based training using PowerPoint slides.
6. In the practice problems condition, the cadet would be presented with a scenario and a series of multiple choice questions as to the proper definitions or courses of action based on that scenario.
7. In the symbolic rehearsal condition, the cadet would be asked to freely recall as many of the previously bolded terms as they can based upon the scenario.

8. In both conditions, the cadet would then receive feedback on the answer they provided with a justification as to why that was or was not the correct answer.
9. Two weeks after the administration of the test, the cadet will receive a surprise pop quiz on the content. The format of the quiz would match their experimental condition.
10. GIFT would then provide customized remediation based on content that the cadet got wrong. Each one of the bolded items from the original course would be tagged with one or more course concepts. Any course concept that a cadet receive more than 20% incorrect answers, he or she would receive remediation on.
11. Two weeks later, all conditions undergo a retention assessment in GIFT.

4.2 Establishing Research Questions

Research Questions. Based on previous research on refresher interventions and skill decay, the following research questions are proposed:

1. Can the use of refresher interventions improve the retention of military instruction content in comparison to traditional learning methods?
2. How do different types of refresher interventions impact assessment performance by cadets?

Hypotheses. These questions will determine whether we can reject or accept the following hypotheses:

1. The experimental conditions will have higher rates of retention than the control condition (supporting research question 1).
2. The practice experimental condition will have higher rates of retention than the symbolic rehearsal condition (supporting research question 2).

4.3 Next Steps and Future Work

The near term next steps include determining what resources and capabilities are needed to facilitate the research as well as an experimental research protocol to lay out specifics and responsibilities by the participating organizations. It will most likely require additional support in the creation of GIFT content to meet DMI's curriculum need.

In the longer term, the specific content being created will have to be developed incrementally. The first type of content is the declarative knowledge assessment that can be presented in GIFT and easily captured via the GIFT survey system. The second type of content is the ability to produce diagrams that are captured on paper. GIFT currently does not have existing functionality that can capture these diagrams, therefore an effort will be made to investigate the translation of line wire diagrams to GIFT course objects. Once the declarative knowledge elements and the line wire diagrams can be combined into a GIFT course, it will be more feasible to represent existing classroom activities, especially with an emphasis on group collaboration.

5 Conclusion

Although this project is still in its infancy, the synergy between ARL's Adaptive Training research program and USMA's Department of Military Instruction serves as a promising classroom use case of GIFT to assist in the mitigation skill decay. It will provide insight in the use of different types of refresher interventions and inspire new research questions for future investigation. It assists in identifying a path for longer range retention studies using GIFT and enhancing GIFT's capabilities to support military learners at USMA. An added benefit of this collaboration will be to identify ways that GIFT can assist the instructors to have a better sense of what is occurring within their classes with the goal of providing more efficient and effective learning for all.

Acknowledgements. This research was sponsored by the Army Research Laboratory Adaptive Training Research Program and the United States Military Academy Department of Military Instruction. The authors would like to thank all of the supporting colleagues in making this project possible. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Boyce, M.W., Rowan, C.P., Baity, D.L., Yoshino, M.K.: Using assessment to provide application in human factors engineering to USMA cadets. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10285, pp. 411–422. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_30
2. Dewey, J.: Experience and Education. Collier Books, New York (1967)
3. Goldberg, B., Hoffman, M., Tarr, R.: Authoring Instructional Management Logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). In: Sottilare, R., Graesser, A., Hu, X., Brawner, K. (eds.) Design Recommendations for Intelligent Tutoring Systems: Authoring Tools, vol. 3. U.S. Army Research Laboratory (2015)
4. Headquarters, Department of the Army: Terms and Military Symbols (ADRP 1-02) (2016)
5. Headquarters, Department of the Army: Unified Land Operations (ADRP 3-0) (2012)
6. Jastrzembski, T.S., Addis, K., Krusmark, M., Gluck, K.A., Rodgers, S.: Prediction intervals for performance prediction. In: Proceedings of the 10th International Conference on Cognitive Modeling, Philadelphia, PA, August 2010
7. Jenkins, S.M., Wills, K., Pick, S., Al-Kutubi, S.: Preventing decay: collaborative partnership between students and staff to prevent deterioration of dental undergraduate practical skills. Pract. Evid. Sch. Teach. Learn. High. Educ. **10**(1), 84–101 (2015)
8. Kim, J.W., Ritter, F.E., Koubek, R.J.: An integrated theory for improved skill acquisition and retention in the three stages of learning. Theor. Issues Ergon. Sci. **14**(1), 22–37 (2013)
9. Kluge, A., Frank, B.: Counteracting skill decay: four refresher interventions and their effect on skill and knowledge retention in a simulated process control task. Ergonomics **57**(2), 175–190 (2014)
10. Kluge, A., Frank, B., Maafi, S., Kuzmanovska, A.: Does skill retention benefit from retentivity and symbolic rehearsal?–Two studies with a simulated process control task. Ergonomics **59**(5), 641–656 (2016)

11. Kluge, A., Burkolter, D., Frank, B.: Being prepared for the infrequent: a comparative study of two refresher training approaches and their effects on temporal and adaptive transfer in a process control task. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56, no. 1, pp. 2437–2441. SAGE Publications, Los Angeles September 2012
12. Merrill, M.D.: The Descriptive Component Display Theory. Educational Technology Publications, Englewood Cliffs (1994)
13. Moxley, J.H., Ericsson, K.A., Scheiner, A., Tuffiash, M.: The effects of experience and disuse on crossword solving. Appl. Cogn. Psychol. **29**(1), 73–80 (2015)
14. Peterson, L., Peterson, M.J.: Short-term retention of individual verbal items. J. Exp. Psychol. **58**(3), 193 (1959)
15. Perez, R.S., Skinner, A., Weyhrauch, P., Niehaus, J., Lathan, C., Schwaitzberg, S.D., Cao, C.G.: Prevention of surgical skill decay. Mil. Med. **178**(suppl_10), 76–86 (2013)
16. Reznikaya, A., Gregory, M.: Student thought and classroom language: examining the mechanisms of change in dialogic teaching. Educ. Psychol. **48**(2), 114–133 (2013)
17. Siu, K.C., Best, B.J., Kim, J.W., Oleynikov, D., Ritter, F.E.: Adaptive virtual reality training to optimize military medical skills acquisition and retention. Mil. Med. **181**(suppl_5), 214–220 (2016)
18. Sottilare, R.A.: Adaptive Intelligent Tutoring System (ITS) research in support of the Army Learning Model—research outline (ARL-SR-0284). US Army Research Laboratory (2013)
19. Sottilare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.K.: The generalized intelligent framework for tutoring (GIFT). US Army Research Laboratory (2012)
20. Wisher, R.A., Sabol, M.A., Ellis, J., Ellis, K.: Staying sharp: retention of military knowledge and skills. Human Resources Research Organization Alexandria, VA (1999)
21. Vygotsky, L.S.: The genesis of higher-order mental functions. In: Wertsch, J.V. (ed.) The Concept of Activity in Soviet Psychology, pp. 144–199. Sharpe, Armonk (1981)

Developing Accelerated Learning Models in GIFT for Medical Military and Civilian Training

Jeanine A. DeFalco[1(✉)], R. Stanley Hum[2], and Michael Wilhelm[3]

[1] United States Military Academy, Oak Ridge Associated Universities, Oak Ridge, USA
jeanine.a.defalco.ctr@mail.mil
[2] Columbia University Medical Center, New York, NY, USA
rsh2ll7@cumc.columbia.edu
[3] University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
mwhilhem@pediatrics.wisc.edu

Abstract. This paper will discuss the protocol of an inter-institutional study between the Army Research Laboratory (ARL) and Columbia University Medical Center that seeks to identify pedagogical models that can be employed in the Generalized Intelligent Framework for Tutoring system (GIFT) to support the transfer of skills from training to operations in individual Soldiers within the domain of critical care, addressing topics in hemorrhage, airway compromise, and/or tension Pneumothorax. The scientific approach will include two studies. The first correlational study aims to examine the effect of human variability on learning, performance, retention, and transfer by using individual differences (e.g., personality traits, cognitive abilities, and motivation) as criteria to tailor individual training for Soldier learning needs. The second study will be an experiment to examine how the priming of analogical reasoning tasks effects the problem-solving outcomes of increasingly complex critical care case study content. The authors intend to incorporate the findings of these two studies to support the development of accelerating expert-level reasoning skills and strategies to achieve cognitive flexibility, one of two paths that has been identified as a way to accelerate proficiency.

Keywords: GIFT · Medical education · Accelerated learning

1 Introduction

1.1 Overview

Developing accelerated learning models in the Generalized Intelligent Framework for Tutoring (GIFT) [26] for medical military and civilian training is a two-phase, inter-institutional effort between the Army Research Laboratory, Columbia University Medical Center, the University of Wisconsin, and the United States Military Academy. The overarching objective is to explore how to support accelerated learning within the domain of critical medical care for Soldiers and civilians as delivered by GIFT.

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 183–191, 2018.
https://doi.org/10.1007/978-3-319-91467-1_15

1.2 Background

This two-study effort seeks to address ARL's Essential Research Area: Accelerated Learning for a Ready and Responsive Force, and contribute to an understanding of what factors, tools, and methods help individual Soldiers learn faster, perform at consistently higher levels, retain knowledge and skills longer, and transfer skills from training to operations at a higher rate. This effort will also address the gap of identifying pedagogical models that can be employed in GIFT to support the transfer of skills from training to operations in individual Soldiers within the domain of critical care, addressing topics in hemorrhage, airway compromise, and/or tension pneumothorax – three leading causes of battlefield deaths [16].

GIFT is a service-oriented framework of tools, methods and standards to author, manage, and assess computer-based tutoring instruction [26]. GIFT is being developed under the Adaptive Training Research Program at the Learning in Intelligent Tutoring Environments (LITE) Laboratory, part of the U.S. Army Research Laboratory - Human Research and Engineering Directorate (ARL-HRED). The goal of this inter-institutional project is to provide empirical findings to U.S. Army stakeholders to contribute to developing models for GIFT to support an accelerated learning pathway for expert-level medical education, as well as explore models that support the higher-order thinking processes of learners [8]. As noted in Hoffman et al. [8], to accelerate instruction requires not only an understanding of tasks that need be learned, but also an understanding of the learner and a delivery of instruction that optimizes the growth and development of expertise by the learner. Lastly, by expanding our participant pool to civilians, we seek to improve the generalizability of these findings and improve external validity, making this study relevant not only to U.S. Army stakeholders, but to civilian medical education institutions as well.

2 Theoretical Approach

2.1 Accelerated Learning

Within the field of accelerated learning, there has been a distinction made between efforts for accelerated learning for novices and accelerated learning that target the journeyman, or senior apprentice, on their way to an expert level [7, 12, 14]. Hoffman [7] identifies the basic proficiency categories of learners as follows:

Naïve: One who is ignorant of a domain.
Novice: Someone who is new – a probationary member who has had some "minimal" exposure to the domain.
Initiate: Someone who has been through an initiation ceremony – a novice who has begun introductory instruction.
Apprentice: One who is learning, a student undergoing a program of instruction beyond the introductory level. Traditionally one who is immersed in the domain by living with and assisting someone at a higher level.

Journeyman: A person who can perform a day's labor unsupervised, although under working orders. An experienced and reliable work who has achieved a level of competence.
Expert: The distinguished or brilliant journeyman, whose judgments are un-commonly accurate and reliable, whose performance shows consummate skill and economy of effort, who can deal effectively with certain types of rare or tough cases. Has extensive experience with subdomains.
Master: One who is a journeyman or expert qualified to teach those at a lower level. A member of an elite group of experts whose judgments establish regulations, standards, or ideals.

While novices are still in a process of synthesizing their understandings of a new domain, the journeyman/apprentice and expert are in process of utilizing those understandings [4, 10, 11]. This process has been called "cognitive readiness," which includes higher-order thinking competencies, such as reasoning skills, amongst others [15].

The cognitive readiness of experts includes not only training that involves mirroring cognitive tasks in real-world tasks, such as using case libraries and scenario-based learning [11] but more importantly it includes the development of a skill set that is used adaptively when facing associated problems or challenges [14]. Jung [14] identifies characteristics of experts that can be universally applied to all domains, which include:

- Possession of an extensive and highly organized domain knowledge.
- The capability of identifying the underlying structure of domain problems.
- Choosing and employing proper problem-solving skills and procedures for the problem at hand.

Essentially, in the effort of supporting expertise development, Jung [14] and Hoffman et al. [12] recommend fostering high-level reasoning skills. According to the Center for Advancement of Learning and Assessment [15], higher order thinking skills include critical, logical, reflective, metacognitive, and creative thinking, that are activated when individuals encounter unfamiliar problems, uncertainties, questions, or dilemmas. Within this framework, then, supporting an accelerated learning pathway to develop the cognitive skills of an expert can arguably be rooted in the development of creative thinking, specifically creative reasoning, which addresses a core element of cognitive readiness. Indeed, this approach falls well within recent thinking of education and training of Army personnel where creative thinking has been noted as both critical and necessary for successful leadership of the military [1].

2.2 Creative Thinking: Analogical Reasoning

Creative thinking includes the convergent process of identifying relevant items or schemas, and the divergent process of combining these items in novel ways [18]. Convergent thinking refers to deductive generation of a single, accurate, concrete, solution [24]. Divergent thinking, in contrast, requires the ability to create multiple, novel ideas [29]. Divergent thinking includes not only a freedom from functional fixedness [6] but it includes the ability to find different and original solutions to

problems and tasks [21]. Importantly, Weisberg [29] has argued that problem solving includes both a level of content expertise in a specific domain as well as strong creative thinking skills, which includes analogical thinking.

Echoing Weisberg's [29] analysis, Weinberger et al. [28] research has focused on creative reasoning – specifically on examining analogical reasoning. Their argument extends the theoretical into the practical evidence of the value of analogical reasoning, noting it as the basis of innovation in science and industry. However, in order for divergent and creative thinking to solve actual problems, solutions must be generated within certain constraints where the outcomes are viable [20]. As such, Weinberger et al. [28], highlight the notion that analogical reasoning is a good model for creativity in reasoning because it involves not only divergent thinking, but more practically it involves the use of sensible constraints.

Therefore, supporting creative thinking is not merely the generation of original and elegant solutions [2], but solutions that are sensible and viable to address complex, novel, ill-defined or poorly structured problems [20]. Within this context, then, creative thinking in this two-phase project will be operationalized within the framework of creative reasoning as constrained for the purposes of producing socially valuable products, and measured by way of divergent thinking tests, specifically analogical thinking tasks [3, 5, 28]. While creative reasoning is typically assessed according to domain specific products, the objective of this current research project will lay the groundwork for whether an instructional design that primes analogical reasoning tasks with sequentially complex case studies can be used as an effective pedagogical model across different domains to support expert level problem solving.

Accordingly, to assess divergent thinking includes seeking out first order relations to form valuable second-order relations that produce innovative solutions, and involves the ability to generate different and original solutions to problems and tasks in a problem context [21, 31]. After having conducted a review of the literature, the previously validated Analogical Finding Task Matrix [28] has been identified as the instrument by which the authors will measure divergent thinking.

2.3 Creative Thinking: Mental Rotation Tasks

Another capacity that has identified as relevant to high-level creative problem solving, and particularly relevant for learning in anatomy that is relevant to critical medical care, is a person's spatial ability [23]. The ability to manipulate metal imagery has not only been identified as a key factor in problem solving, but in memory as well [30]. Spatial ability and mental rotation has also been linked to success in surgical skill acquisition [27]. Most pertinent to this project is a current research interest in determining whether mental rotation can be improved through training as one methodology to improving performance in STEM [23].

There is a robust body of literature in the field of mental rotation testing [22, 25] More recently, Ganis and Kievit [5], constructed and validated a new set of three-dimensional shapes for investigating mental rotation processes that improves on the work of Shepard and Metzler [25]. One such improvement is the inclusion of shading cues that minimize error due to crowding (meaning, difficulty distinguishing edges of objects) and depth ambiguity (meaning, ambiguity as to the direction of the object from the perspective of

the viewer). The result of Ganis and Kievit's [5] work has resulted in a new set of 48 distinct mental rotations objects with rotated versions that include shading depth cues. Speed and accuracy are measured, then, to determine a person's level of spatial abilities. This metric will be particularly important to measure – not only to achieve a baseline assessment of a person's spatial abilities, but also will be informative as to the level of detailed imagery that will be narrated in the medical case study scenarios.

2.4 Creative Reasoning: Content Mastery

There is a body of research that maintains that another key element to creative thinking and reasoning is content mastery [13, 17, 29, 31]. The target domain for this project will be limited to medical education, specifically critical care, addressing topics in hemorrhage, airway compromise, and/or tension Pneumothorax, which are three leading causes of battlefield deaths. Focusing on these areas are in line with prior research [8, 9, 14, 19] that maintains accelerated learning methods should include leveraging computer technology to develop libraries, or case studies, that represent tough tasks and capture expert knowledge and skill. Ideally, future work in accelerated learning should include employing a library of "tough cases" that focus cognitive training on real work practice, highlighting the use of analogical strategies.

For this research, then, critical care case studies will be developed so that the successful completion of each case study, measured by way of an assessment instrument following each case study (still to be developed), will represent a different level of problem solving expertise from novice, to journeyman, to expert. Dr. R. Stanley Hum of Columbia Medical Center will develop these critical care case scenarios, and have these scenarios validated by critical care expert Dr. Michael Wilhelm of the University of Wisconsin's Medical Center, as well as by Major Angela Yarnell, Ph.D., of the United States Military Academy, whose expertise lies in medical education and includes critical care.

3 Experimental Designs of Study One and Two

3.1 Study One: Analogical Reasoning and Trait Correlational Study

For the first phase of this project, an initial correlational study will examine strengths of correlations between mental rotation/spatial ability; grit; analogical reasoning; personality types; and level of medical knowledge expertise. The goal is to determine what traits are relevant to superior analogical thinking skills and to use these outcomes to help inform the experimental design of study two. This information is key in designing an adaptive tutoring platform to tailor instruction for the individual medical military and civilian populations.

The first study will recruit approximately 128 participants through the United States Military Academy's (USMA) experiment sign up system (SONA-Systems) in collaboration with the Department of Behavioral Sciences and Leadership. Recruitment will also be conducted at Columbia Medical Center and Columbia University, seeking approximately 128 participants. This first study will be completed online as it consists of self-reporting questionnaires and surveys, and tasks that measure individual traits.

3.2 Study Two: Priming Analogical Reasoning Tasks and Problem Solving Medical Scenarios of Increasing Complexity

The second phase of this project will be a within-and between group design experiment. This experiment will examine how the priming of analogical reasoning tasks, and sequencing of the analogical reasoning with schematic content (medical definitions of related content), and scenario-based case studies of increasing complexity, effect the accuracy and speed of participants' problem-solving outcomes. This experiment will also examine how the other variables of personality type, grit, and level of medical knowledge expertise, moderate and mediate effects on the dependent variables of time and problem-solving outcomes of participants assigned to three experimental and one control condition groups.

Design. The design of this experiment will be a 3 × 1 experimental design with one control group and three experimental conditions. The intervention conditions (conditions one, two, and three, will prime the participant with one or both analogical reasoning task prior to reading the case studies and answering the post assessments of the case studies. All participants will be taking all the same instruments and only the manipulation of sequencing is different. The primary focus of this experiment will be to test the hypothesis that there will be a statistically significant difference between the learning outcomes in the experimental condition that primes participants with both mental rotation and analogical reasoning tasks when solving for the medical case studies – and the condition that primes only with the analogical reasoning task, the condition that primes only with the mental rotation task, and the control condition that does not prime analogical reasoning nor mental rotation tasks prior to solving the case studies.

Procedure. This second study will recruit 128 participants the United States Military Academy's (USMA) experiment sign up system (SONA-Systems) in collaboration with the Department of Behavioral Sciences and Leadership. Recruitment will also be conducted at Columbia University Medical Center, seeking approximately 40 participants. During the 1-h session with each participant, the participant will take, via the GIFT platform: a demographic survey, a pretest on critical care, a grit survey, a personality test, and an analogical reasoning task. Further all conditions will have post-tests after each case study that will assess their ability to successfully resolve the problems laid out in the case studies. Participants will be randomly assigned to one of four conditions on a critical care course (to be developed) that will be delivered via the GIFT platform, as follows:

1. *Experimental group one*: Sequence of mental rotation task, analogical reasoning, and schematics prior to scenario case studies, sequenced from novice to expert, post-test that evaluates problem solving of medical scenarios
2. *Experimental group two*: Sequence of analogical reasoning and schematic material prior to scenario case studies, sequenced from novice to expert; mental rotation task after post-test that evaluates problem solving of medical scenarios

3. *Experimental group three*: Sequence of mental rotation tasks and schematics prior to case studies of increasing complexity of novice to expert; analogical reasoning after post-test that evaluates problem solving of medical scenarios
4. *Control group*: Sequence only of schematics prior to case studies of increasing complexity from novice to expert; mental rotation and analogical reasoning after post-test that evaluates problem solving of medical scenarios.

The primary objective of this second study will be to test the hypothesis that there will be a statistically significant difference between the problem-solving outcomes in condition 1 vs. conditions 2 and 3 and the control condition. However, analyses will also be run to determine whether personality traits will function as a moderator and have an interactive effect on learning outcomes; to determine if Grit function as a mediator and has an interactive effect on problem solving outcomes; to examine whether analogical reasoning skill will function as a mediator and have an interactive effect on problem solving outcomes; and to see whether there will be a statistically significant difference in time of completion and accuracy of problem solving outcomes not only between conditions, but between novice and expert levels of medical students/practitioners.

4 Conclusion

In sum, the purpose of this inter-institutional study is aimed at developing a pedagogical model to support accelerated learning for the purposes of creating a learning pathway model that would accelerate the learning from journeyman to expert via GIFT. After determining what trait variables are more highly correlated to analogical reasoning and mental rotation, the authors will proceed to develop critical care case studies to further explore whether analogical reasoning and mental rotation tasks support expert critical care problem solving. We expect that in the second study will see a statistically significant effect for priming of analogical reasoning and mental rotation tasks in relationship to the fluency and speed of the problem-solving abilities of participants when solving the medical case studies, particularly as they increase in complexity. Further, we expect our post-analysis to provide evidence that can inform future examinations on how the engagement with analogical and creative reasoning tasks can be further capitalized upon to accelerate the learning process within the field of critical care medical education.

Acknowledgements. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number **W911NF-17-2-0152**. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Allen, C.D., Gerras, S.J.: Developing Creative and Critical Thinkers. Army Combined Arms Center, Fort Leavenworth (2009)
2. Besemer, S.P., O'Quin, K.: Confirming the three-factor creative product analysis matrix model in an American sample. Creat. Res. J. **12**(4), 287–296 (1999)
3. Cropley, A.J.: A note on the Wallach-Kogan tests of creativity. Br. J. Educ. Psychol. **38**(2), 197–201 (1968)
4. Fadde, P.J.: Instructional design for advanced learners: training recognition skills to hasten expertise. Educ. Technol. Res. Dev. **57**(3), 359–376 (2009)
5. Ganis, G., Kievit, R.: A new set of three-dimensional shapes for investigating mental rotation processes: validation data and stimulus set. J. Open Psychol. Data **3**(1), e3 (2015). https://doi.org/10.5334/jopd.ai
6. Guilford, J.P.: Creative Talents: Their Nature, Uses and Development. Bearly Limited, Buffalo (1986)
7. Hoffman, R.R.: How can expertise be defined? Implications of research from cognitive psychology. In: Williams, W.R., Faulkner, J.F. (eds.) Exploring Expertise, pp. 81–100. Macmillan, New York (1998)
8. Hoffman, R.R., Andrews, D., Fiore, S.M., Goldberg, S., Andre, T., Freeman, J., Fletcher, J. D., Klein, G.: Accelerated learning: prospects, issues and applications. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 54, no. 4, pp. 399–402. SAGE Publications, Los Angeles, September 2010
9. Hoffman, R.R., Feltovich, P.J., Fiore, S.M., Klein, G., Ziebell, D.: Accelerated learning (?). IEEE Intell. Syst. **24**(2), 18–22 (2009)
10. Hoffman, R.R., Fiore, S.M., Klein, G., Feltovich, P.: Accelerating the achievement of mission-critical expertise. Report to the Electric Power Research Institute, Palo Alto, CA (2008)
11. Hoffman, R.R., Militello, L.G.: Perspectives on Cognitive Task Analysis: Historical Origins and Modern Communities of Practice. CRC Press, Boca Raton (2012)
12. Hoffman, R.R., Ward, P., Feltovich, P.J., DiBello, L., Fiore, S.M., Andrews, D.H.: Accelerated Learning: Training for High Proficiency in a Complex World. Psychology Press, New York (2013)
13. Jaussi, K.S., Randel, A.E.: Where to look? Creative self-efficacy, knowledge retrieval, and incremental and radical creativity. Creat. Res. J. **26**(4), 400–410 (2014)
14. Jung, E.: Expertise development through accelerated learning: a multiple-case study on instructional principles. Doctoral dissertation, Indiana University (2016)
15. King, F.J., Goodson, L., Rohani, F.: Executive Summary: Definition. Higher Order Thinking, 22 October 2011. http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf
16. Kotwal, R.S., Montgomery, H.R., Kotwal, B.M., Champion, H.R., Butler, F.K., Mabry, R. L., Cain, J.S., Blackbourne, L.H., Mechler, K.K., Holcomb, J.B.: Eliminating preventable death on the battlefield. Arch. Surg. **146**(12), 1350–1358 (2011). https://doi.org/10.1001/archsurg.2011.213
17. Medeiros, K.E., Partlow, P.J., Mumford, M.D.: Not too much, not too little: the influence of constraints on creative problem solving. Psychol. Aesthet. Creat. Arts **8**(2), 198 (2014)
18. Mumford, M.D., Gustafson, S.B.: Creative thought: cognition and problem solving in a dynamic system. Creat. Res. Handb. **2**, 33–77 (2007)
19. Mumford, M.D., Medeiros, K.E., Partlow, P.J.: Creative thinking: processes, strategies, and knowledge. J. Creat. Behav. **46**(1), 30–47 (2012)

20. Mumford, M.D., Mobley, M.I., Reiter-Palmon, R., Uhlman, C.E., Doares, L.M.: Process analytic models of creative capacities. Creat. Res. J. **4**(2), 91–122 (1991)
21. Palmiero, M., Di Giacomo, D., Passafiume, D.: Divergent thinking and age-related changes. Creat. Res. J. **26**(4), 456–460 (2014)
22. Peters, M., Battista, C.: Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. Brain Cognit. **66**(3), 260–264 (2008). https://doi.org/10.1016/j.bandc.2007.09.003
23. Roach, V.A., Fraser, G.M., Kryklywy, J.H., Mitchell, D.G.V., Wilson, T.D.: Different perspectives: spatial ability influences where individuals look on a timed spatial test. Anat. Sci. Educ. **10**(3), 224–234 (2017)
24. Runco, M.A.: Creativity. Theories and Themes: Research, Development, and Practice. Elsevier Academic Press, Burlington (2007)
25. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science **171**(3972), 701–703 (1971). https://doi.org/10.1126/science.171.3972.701
26. Sottilare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.K.: The generalized intelligent framework for tutoring (GIFT). Concept paper released as part of GIFT software documentation. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED), Orlando (2012). https://gifttutoring.org/attachments/152/GIFTDescription_0.pdf
27. Wanzel, K.R., Hamstra, S.J., Anastakis, D.J., Matsumoto, E.D., Cusimano, M.D.: Effect of visual-spatial ability on learning of spatially-complex surgical skills. Lancet **359**, 230–231 (2002)
28. Weinberger, A.B., Iyer, H., Green, A.E.: Conscious augmentation of creative state enhances "real" creativity in open-ended analogical reasoning. PLoS ONE **11**(3), e0150773 (2016)
29. Weisberg, R.W.: Creativity: Understanding Innovation in Problem Solving, Science, Invention, and the Arts. Wiley, Hoboken (2006)
30. Yates, F.A.: The Art of Memory. Routledge and Kegan Paul, London (1966)
31. Zeng, L., Proctor, R.W., Salvendy, G.: Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity? Creat. Res. J. **23**(1), 24–37 (2011)

Experiential Intelligent Tutoring: Using the Environment to Contextualize the Didactic

Benjamin Goldberg(✉) and Michael Boyce

U.S. Army Research Laboratory, Orlando, FL 32826, USA
{benjamin.s.goldberg.civ,
michael.w.boycell.civ}@mail.mil

Abstract. With advancements in mobile computing technologies, extending training and education opportunities outside of traditional classroom environments is more feasible than ever. In this paper, we discuss the application of Intelligent Tutoring System (ITS) technologies to drive self-regulated educational experiences that incorporate physical elements of an environment to contextualize the concepts being instructed. This supports an experiential learning model that blends structured didactic instruction with interactions and assessments that utilize the physical environment.

Keywords: Intelligent tutoring systems · Experiential learning
Mobile application · Informal learning

1 Introduction

Intelligent Tutoring Systems (ITSs) are technology based learning solutions that provide personalized training and education experiences through Artificial Intelligence (AI) techniques grounded in learning science. These systems provide self-regulated learning opportunities through closed-loop inference procedures that contextualize behavior observed during interaction into actionable metrics that drive real-time pedagogical decisions across feedback and scenario/problem management dimensions. These pedagogical specifications are designed to augment training by addressing the needs of each individual as they relate to observed performance, inferred competency, and classified emotional reactions detected during a set of closed-loop assessment procedures [1].

The caveat with these processes is the dependency on the learning environment and data sources captured within, as well as the interfacing technologies made available to communicate and interact with the learner [2]. These dependencies are especially interesting when considering how ITS techniques are applied across mobile contexts outside of controlled classroom environments. With a push from the U.S. Army to extend ITS capabilities into military relevant simulation-based training spaces, this challenge is especially relevant. One challenge the authors address in this paper is the need to apply pedagogical practice in open live environments that vary from data interactions captured within desktop training applications (e.g., completing an exercise in VBS3) and controlled high-end simulation environments (e.g., marksmanship training in the Engagement Skills Trainer 2 [3]).

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 192–204, 2018.
https://doi.org/10.1007/978-3-319-91467-1_16

For this purpose, the Generalized Intelligent Framework for Tutoring (GIFT) is applied as a guiding function to conceptually design a mobile app training paradigm focused on experiential learning intertwined with just-in-time didactic instruction. GIFT is an open-source project built on a set of tools and methods used for the development, delivery, and evaluation of adaptive training applications [4]. This proposed open environment paradigm incorporates: (1) interpretation of mobile location and movement data over a cellular network as it relates to a set of specified tasks, objectives and concepts, (2) explicit feedback delivery mechanisms triggered by GIFT configured assessments for deliberate guidance purposes, and (3) the application of tailored training activities designed to mix real-world context with underlying conceptual understanding through personalized instructional activities. Two examples of this paradigm will be reviewed, with recommended guiding principles linked to the development and delivery of such approaches.

2 Blending the Physical Environment with Didactic Instruction

When taking into account the design of an experiential intelligent tutoring experience, there are a number of considerations that must be addressed that will inform development requirements. This includes defining and storyboarding the training model itself, along with any upfront assumptions, and defining the technological components that need to be in place with respect to data flow and instructional techniques.

2.1 Defining the Training Paradigm

The proposed instructional model is intended to leverage advancements in both mobile computing technologies and adaptive training capabilities to create a truly immersive learning experience that is focused on real-world context [5]. In this instance, rather than learn about facts, concepts, and procedures in a classroom environment, we are promoting an educational paradigm that embeds formal learning activities in informal settings that often require expert human instructors to lead the learning activities. Through development efforts to embed ITS functions in mobile applications, an individual's physical location can be used to trigger activities that use environmental features as contextualized learning content that would not be able to be replicated in a classroom or simulation-based setting [6].

The goal is to promote experiential learning by using one's surroundings as the learning environment itself. From a strictly conceptual standpoint, to promote this training model, a system will need the following functions at a minimum:

- Data – Performance, behavioral, and physiological inputs captured from the training environment.
- Triggers – Specified production rules and policies that serve as start- and end-triggers for contextualizing interaction based on location within the training environment.

- Content – Specified content and scenarios used as didactic exercises based on contextualized triggers.
- Assessments – Assessments designed for measuring performance and competency that merge the physical environment with the concepts being instructed didactically.

These functions, when combined, support an immersive self-regulated learning experience that intertwines content and assessment as one navigates through a configured environment. One assumption guiding initial model development is the dependency on map and location data to drive these instructional interventions. Accurate second-by-second tracking capability is critical to the utility of this approach. With these defined assumptions and dependencies, the next step is establishing tools and methods for an experiential intelligent tutor mobile application.

2.2 Defining the Technological Components

In implementing an experiential learning activity in an informal setting (i.e., 'in the wild'), there are a number of technology components that must be in place. This includes defining an architecture for task representation and data structures, establishing a data flow model that will guide inference procedures, the type of learning activities applied to drive training, how mobile technologies and applications apply, and how these tools and methods are used to author and configure the models and interactions used at run-time.

Tutoring Architecture. To frame the discussion, the proposed self-regulated training model will be conceptualized for implementation in GIFT. GIFT's architecture is modular and domain-agnostic, where its components can be configured to work across any conceivable training and education environment. GIFT is now implemented in a cloud-instance with modules and services hosted as an Amazon Web Service on a set of dedicated servers [7]. This cloud instantiation lays the ground work to support experiential intelligent tutoring by providing networked access to the core software modules that manage the training. A requirement is developing a data connection between GIFT Cloud and a cellular network for data tracking purposes, with GIFT's core modules being configured based on available data sources.

The core modules in GIFT establish foundational model representations required to drive tutoring interactions. This includes who is being trained (learner model), what is being trained (domain model), and how best to train (pedagogical model). For application across any disparate domain, a generalized schema was established in GIFT that guides authoring and configuration while developing adaptive scenario content. The schema is a hierarchical data structure used to define tasks to be executed, concepts to be applied within each task, and conditions for assessing the application of those concepts under the constraints of the task. This hierarchical breakdown is used to guide the data flow between modules.

Data Flow. During the execution of a training scenario, GIFT applies a data flow model that is used for instructional strategy selection. This inference procedure, called the Learning Effect Model (see Fig. 1) [8], is applied to contextualize raw system interaction data for assessment purposes. Interaction data (simulation, behavioral, and

is some cases physiological) is routed in real-time to the domain module where performance is inferred based on defined conditions across each established task concept. This data is used to generate performance states that are routed to the learner module for establishing an overarching learner state. The learner state combines performance information with relevant individual differences (i.e., affective states and traits). This resulting learner state is routed to the pedagogical module where policies are applied to determine a strategy selection. The strategy selection contains a request to either provide guidance (e.g., give a hint or prompt a learner), adapt the scenario (e.g., increase the complexity), request assessment (e.g., ask the learner a question), and do nothing. To complete the data flow of the Learning Effect Model, a strategy selection is then translated into a contextualized tactic that can be delivered within the specific learning environment one is interacting with. With that said, to support this inference model in any specified learning environment, including the natural environment, GIFT requires interoperability with external applications and interfacing technologies.

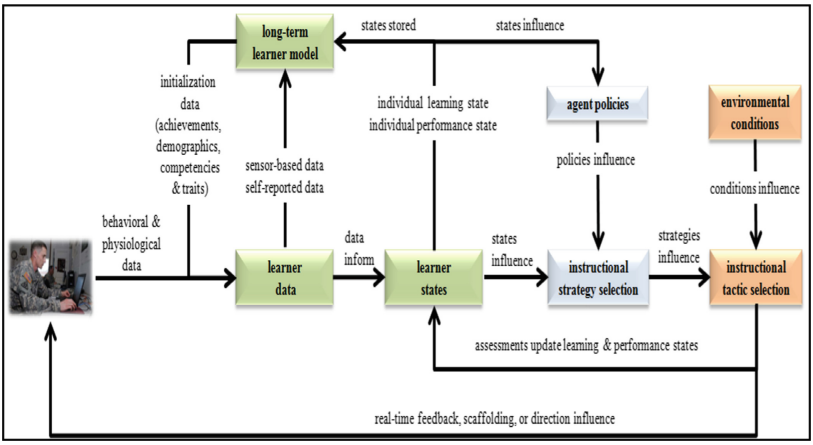

Fig. 1. GIFT's learning effect model (Sottilare….)

Interoperability. A major component of a generalized approach to developing ITS applications is being able to integrate and interoperate with a variety of learning platforms that incorporate elements across Live, Virtual, and Constructive (LVC) simulations. GIFT meets this need through the development of the Gateway Module, where specifications are made to link the architecture with existing non-adaptive systems [7]. The Gateway Module ultimately is designed to consume interaction data made available from a learning environment, where message protocols are specified and parsed for the data inputs required to inform the assessment conditions. A current example is seen in linking GIFT with VBS3. In this instance, the Gateway is configured to receive Distributed Interactive Simulation (DIS) data packets, where GIFT parses out location data, movement data, weapon data, and scenario object data, and routes these inputs into the Domain Module where performance states are computed. This injection of real-time data initiates the Learning Effect Model [8]. For all

learning environments integrated with GIFT, a Gateway Module is required that adheres to the specific message types generated during interaction. This dependency holds true when considering the utility of mobile interaction data to drive the effect chain. For this purpose, the development of a mobile application to support closed-loop inference procedures must be discussed.

Mobile Application. For the proposed learning model to extend into the natural environment, a mobile application is required. The mobile application serves multiple functions in this instance, where the device itself operates as a sensor, an input device, an interface, and a processor (see Table 1). As discussed above, the mobile application will require a Gateway Module to sync its inputs into the GIFT Cloud server where assessments and pedagogical decisions will be executed.

Table 1. Mobile application components in support of an experiential intelligent tutoring.

Mobile app components	Data type	Data function (examples)
Sensors	(a) GPS	(a) Location, movement, and orientation data
	(b) Accelerometer	(b) Body posture data
	(c) Activity tracker (watch/band)	(c) Physiological (ECG, GSR, etc.) and posture data via Bluetooth
Inputs	(a) Touch	(a) Manipulate touch required inputs (e.g., questions, maps, navigation)
	(b) Typing	(b) Natural language inputs to inform assessments and drive After Action Reviews
	(c) Voice	(c) Voice to text inputs to facilitate communication task requirements
Interface	(a) Html content	(a) Display web-browser supported content (e.g., web pages, videos, pdfs, etc.) as didactic content
	(b) Tutor user interface	(b) Display real-time feedback based on runtime assessments
	(c) Virtual humans	(c) Use agent technology to communicate with learner
	(d) AR/smart glasses	(d) Interface with smart glasses to display information over the environment

With data and interfacing variables identified, another component critical to a mobile application is the computing processor on the device itself. Because of the required linkage to a GIFT Cloud server, the onboard device processor is critical in executing data functions locally to reduce the bandwidth requirement, which will ultimately minimize network traffic. For this purpose, the mobile application might require customized GIFT modules running locally to process and classify data before sending it to the server for assessment. An example would include classifying workload or fatigue by processing the received physiological data in GIFT's sensor module to remove the need to stream raw sensor data into the cloud for classification purposes. With a mobile application in place that supports the necessary data flow to implement

the Learning Effect Model, the next element to address is learning activities and the role they serve in the experiential learning model.

Mobile Learning Activities. To facilitate the didactic instruction component of the experiential intelligent tutor, mobile learning activities are needed that facilitates the presentation of content, the delivery of assessments, and the management of feedback and remediation materials. In this use case, learning activities are initialized by some trigger related to the natural environment (e.g., learner came within 5 m of a designated way point). The mobile application will present content linked to the learner's physical location, with smartphone technologies providing multiple supported media formats. Following the presentation of content (which is optional), an assessment will be conducted that incorporates elements of the surrounding environment. This may take the form of a supported GIFT survey question (e.g., multiple choice, matching, T/F), along with other custom activities that leverage the mobile data and interfacing modalities (e.g., interacting with a map). While the mobile device can provide accurate learners' location within a learning environment, the embedded learning activities and assessments are the critical component to contextualizing the instruction. It's important that the activity leverages elements of the natural environment to better support accurate mental schemas of the concepts being instructed and its interacting parts.

Easy Authoring. With all technology components in place to support an experiential intelligent tutor, the remaining long-pole-in-the-tent is authoring and configuration. The main work performed in this capacity is establishing a domain model, captured in GIFT's Domain Knowledge File (DKF). The DKF is formulated from the GIFT schema, and is used to manage the configuration of assessments and instructional tactics as they relate the tasks, conditions, and standards being training within a specified scenario.

To ease the burden of authoring, GIFTwrap was developed as an intuitive authoring interface that enables a training developer to configure GIFT DKF assessment logic while interacting directly with a simulation-based mission editor [9]. GIFTwrap has currently been applied to support training exercises in constructive (e.g., Augmented Reality Sandtable; Amburn) and virtual simulations (VBS3), with primary assessments operating on location and movement oriented data as it relates to a specified terrain and map model to guide initial development efforts [6].

Current work is being performed to extend its application to support live training interaction data by linking GIFT to available map-based Application Programming Interfaces (API), such as GoogleMaps API. With this linkage, map and terrain data made available through Gateway specifications can be used to contextualize available mobile data types (see Table 1) around the tasks a learner will be asked to perform. What is required is re-conceptualizing the role of assessments in a live environment and how they might impact the learning experience. As discussed above, for a live training use case, an author will be responsible for configuring (1) triggers that initialize task interactions and (2) assessments that are used to guide coaching interventions and scenario adaptations. Examples of applying GIFTwrap to author triggers and assessments in a live-map environment are seen in Fig. 2. In this instance, GIFT is configured to trigger events when a learner reaches Points of Interest A-D (as designated by

entering the blue circle). In addition, GIFT can track your progress in navigating to the assigned points, where feedback can be provided if they wander too far off course (as designated by the orange corridors between points.

Fig. 2. GIFT wrap authoring interface showing four configured triggers used to drive learning activity initiation. (Color figure online)

While GIFTwrap can be used to author an array of assessments based on variations of available data sources, the use cases described below use GIFTwrap for the purpose of configuring a DKF to support the blended experiential learning model that contextualizes the natural environment with focused didactic content. For each case, we will present the triggers that drive interaction, the content associated with each trigger, and the underlying assessments and activities that will drive performance tracking and remediation practices.

3 Exemplar Use Cases

Below we present two distinct use cases that highlight the utility of an experiential intelligent tutor application. Each use case highlights different types of interaction and assessment strategies that can be applied to train a set of knowledge and skills that relate to the elements in a real-world task environment. The domains used to guide the discussion are land navigation and architectural design.

3.1 Use Case 1: Land Navigation and Terrain Association

A critical skill required for successful completion of a land navigation task is terrain association. This involves having an accurate mental model of contour lines represented on a topographical map and how those translate to their actual physical representation in the natural environment. This is essential when it comes to route planning for efficiently navigating terrain, as well as essential for tactical decision making by understanding land features and its impact on operational objectives.

To assist in the training of these concepts and skills, instructors often perform what is called a terrain walk. This is traditionally facilitated by a subject matter expert guiding a group of learners through a path with designated areas used to facilitate instruction. This is often a collaborative exchange where an instructor asks individual learners to point out land features and to describe their characteristics. The interactions are believed contextualize classroom instruction through direct observation of terrain's physical features, where tactical decision making can improve based on better mental terrain representations. In addition, as the group of learners move through the environment, they also have the opportunity to practice dead-reckoning procedures using a map and compass.

With a current need for SME support to facilitate, a terrain walk exercise is a perfect candidate for an experiential intelligent tutoring application for self-regulated use. By designating specified routes and activity points (see Fig. 2), and using mobile data for tracking purposes, a personalized terrain walk experience is based on triggered learning activities that incorporate content, assessments, and remediation materials. When a learner reaches a specified location, GIFT's Learning Effect Model is enacted to deliver an instructional activity. The activity is designed to replicate SME type interactions and to target the associated learning objectives linked with their point location on the terrain walk exercise. Along with location, orientation data can also be incorporated as an input to further enhance the assessment space by estimating field of view parameters and inferring what can and can't be seen (see Fig. 3).

With a learner's known location, and a SMEs input on the features of that location, targeted assessments and interactions can be triggered (see Fig. 4). Based on their location and their required understanding of topographical features, the GIFT Mobile App can assess a learner in two proposed formats. The first, which requires no additional development, utilizes GIFT survey authoring infrastructure to ask a question that requires observation of the environment to answer. Following, the second activity, which is an example of customized activity leveraging mobile map interfacing, requires the learner to identify the specific features on a map based on the topographic features. It's important to note that at each level of assessment, whether at the survey level or custom activity level, GIFT can provide feedback and remediation based on learner inputs. The goal is correct impasses and misconceptions, and to reinforce correct procedures and skill applications. Feedback is inherently linked through the Learning Effect Model, and configures and presents prompts based on observed shifts in performance states.

Through the GIFT Mobile instantiation of a terrain walk exercise, each learner can receive a personalized experience that is guided through a configured DKF. This self-regulated approach is believed to remove a requirement for instructor-led terrain

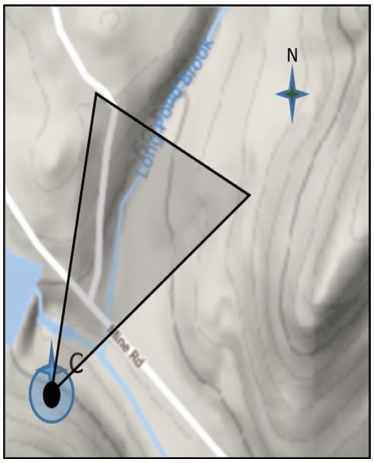

Fig. 3. Mobile app references orientation data as input for assessment and coaching (shaded triangle represents estimated field of vision) following triggered learning activity.

walks, where technology can facilitate the learning objectives defined within this exercise. Research is required to determine the utility of this approach and to collect training effectiveness oriented data to determine a Return on Investment of the produced training application [10].

3.2 Use Case 2: Architectural Design and City Planning

Beyond military application, these proposed technologies can apply in many public-sector domains, including education, training, and even tourism. Much like the terrain walk example, using known location data can be used in any data-accessible environment to contextualize instruction, including cities. An excellent domain for application on city streets is learning about building engineering and architectural design. A task might be to learn about the architectural history by visiting specified locations, viewing customized information authored by an instructor, and then completing assessment activities related to topic and objectives.

In the context of an experiential tutor, GIFT would be triggered as an individual is walking up to a specified location (see Fig. 5 for a GIFTwrap configured map of New York City with specified triggers). This is not to say that current app technology does not exist (i.e. the Discover NYC Landmarks App) but rather that GIFT can take it to a higher level of instructional assessment, while also enabling an instructor to easily build their own experiential content block by block, or even location by location. With the GIFTwrap tool, an instructor could conceivably teach about the architecture of Venice, Italy, in the classroom one day and then that very night produce an experiential intelligent tutor for use by students taking a summer trip.

As an example, let's assume a learner is visiting New York City and is approaching the Empire State Building in midtown Manhattan. As a learner walks up to the building, audio and video content can be cued via the GIFT mobile application (e.g.,

Fig. 4. Example mobile interactions for training terrain association in the wild (upper left: multiple choice question requiring touch input; upper right: map feature identification task requiring touch input; lower: html task feedback reinforcing learning objectives by linking content to physical environment)

voice comes on to say "You have reached the Empire State Building, the 5th tallest skyscraper in the United States, completed in 1931. It is built in Art Deco style which combined styles such as geometric figures of Cubism with the bright colors of Fauvism [11]. Check out this video that goes over the blueprint designs." The mobile app can also be configured to prompt the learner to look for specific features of the building that point out the unique nuances that went into the design, and use orientation data to guide

Fig. 5. GIFTwrap configured for experiential learning with the Empire State Building (from street [marker A] and observation deck [marker B]) and Chrysler Building [Marker C]) [x].

their gaze. From an instructional standpoint, an author is then responsible embedding instructional activities that include assessments for inferring understanding and competence. As seen in Fig. 5, you can even apply elevation data as a triggering condition, where instructional activities are to be enacted when a learner reaches the Empire State Building observation deck.

While we reviewed notional applications of this mobile approach to instruction in land navigation and building architecture, there are multiple use cases a learning technology of this nature can support. The overarching goal is all about contextualization. This is especially relevant with domains that incorporate numerous interacting elements that are difficult to replicate and ground in a classroom, or even simulated environment. Regardless of the domain, each experiential tutor works on the same data models, where guidelines can be established that empower any non-technical user to design and develop their own applications for use by a group of their choosing.

4 Future Work

4.1 Experiential Intelligent Tutoring Implementation

With a design in place, the next step is implementing the specified interactions based on GIFT's authoring workflows. This requires establishing a gateway connection with a mobile network and then programming condition classes (e.g., determine when a learner enters a location, measure distance traveled, etc.) based on the available data inputs. These condition classes are established in GIFT's source-code and referenced during authoring for the purpose of building scenario specific assessments. These components are then used to build a DKF that is applied to a specified scenario (see Fig. 6).

The DKF requires an ontological breakdown of tasks, concepts, and sub-concepts that are used to manage assessment and pedagogical configurations. Based on each

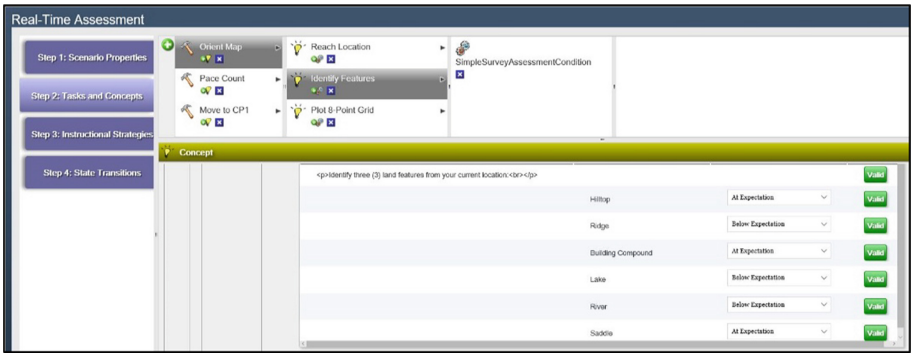

Fig. 6. GIFT's DKF (i.e., real-time assessment) authoring tool with three tasks configured for a terrain walk exercise.

task, concepts are defined that associate with configured assessments. Those assessments are then used to trigger the instructional activities described above, along with feedback messages that are based on performance outcomes. With a DKF in place, a full mobile course can then established and shared with an intended audience for wide use.

4.2 Social Media Component

Additionally, we are interested in a social media component, where students collaborate in groups where they can annotate and share information as they are moving through the environment. This format enables learners to take control of their learning where they can upload content of interest and rate the content of others. The strength of this interaction is the interconnectivity between the students, the environment, and the learning content. There is much research left to be done regarding the role of social networks in self-regulated informal learning settings.

References

1. Woolf, B.P.: Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing e-Learning. Morgan Kaufmann, Burlington (2009)
2. Sottilare, R., Goldberg, B., Brawner, K.W., Holden, H.: Modular framework to support the authoring and assessment of adaptive computer-based tutoring systems. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL (2012)
3. Goldberg, B., Amburn, C., Ragusa, C., Chen, D.W.: Modeling expert behavior in support of an adaptive psychomotor training environment: a marksmanship use case. Int. J. Artif. Intell. Educ. (2017). https://doi.org/10.1007/s40593-017-0155-y
4. Sottilare, R., Brawner, K.W., Goldberg, B., Holden, H.: The generalized intelligent framework for tutoring (GIFT). In: Best, C., Galanis, G., Kerry, J., Sottilare, R. (eds.) Fundamental Issues in Defense Training and Simulation, pp. 223–234. Ashgate Publishing Company, Burlington (2013)

5. Kolb, D.A., Boyatzis, R.E., Mainemelis, C.: Experiential learning theory: previous research and new directions. In: Sternberg, R.J., Zhang, L. (eds.) Perspectives on Thinking, Learning, and Cognitive Styles: The Educational Psychology Series, pp. 227–247. Erlbaum, Mahwah (2001)
6. Goldberg, B., Davis, F., Riley, J.M., Boyce, M.W.: Adaptive training across simulations in support of a crawl-walk-run model of interaction. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10285, pp. 116–130. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_8
7. Sottilare, R.A., Brawner, K.W., Sinatra, A.M., Johnston, J.H.: An updated concept for a generalized intelligent framework for tutoring (GIFT). Aberdeen Proving Grounds, MD (2017)
8. Sottilare, R.A., Ragusa, C., Hoffman, M., Goldberg, B.: Characterizing an adaptive tutoring learning effect chain for individual and team tutoring. In: Interservice/Industry Training, Simulation & Education Conference. Orlando, FL (2013)
9. Davis, F.C., Riley, J.M., Goldberg, B.S.: Development of an integrated, user-friendly authoring tool for intelligent tutoring systems. In: Paper presented at the Proceedings of the 5th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym5) (2017)
10. Goodwin, G.A., Kim, J.W., Niehaus, J.: Modeling training efficiency and return on investment for adaptive training. In: 5th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym5) (2017)
11. Wikipedia Empire State Building Page. https://en.wikipedia.org/wiki/Empire_State_Building. Accessed 02 Mar 2018

Guided Mindfulness: Optimizing Experiential Learning of Complex Interpersonal Competencies

Richard L. Griffith(✉), Lisa A. Steelman, Nicholas Moon, Sherif al-Qallawi, and Nisha Quraishi

Florida Institute of Technology, Melbourne, FL 32901, USA
griffith@fit.edu

Abstract. This paper presents an artificially intelligent platform designed to enable experiential learning of complex interpersonal competencies. Called Guided Mindfulness, the platform supports learning from on-the-job experiences through guided questioning and reflection. The AI platform is described within the context of a year-long learning cycle. Theories of mindfulness are linked to experiential learning with a specific emphasis on how Guided Mindfulness is an improvement over traditional mindfulness interventions for this type of learning.

Keywords: Guided mindfulness · Mindfulness · Experiential learning
Self-regulation

1 Introduction

The technology of munitions has rapidly advanced over the last half century. Until the 1980s, airborne munitions were comprised largely of unguided gravity bombs. When dropped en masse, these munitions had an impact in the vicinity of a targeted area, but did not offer a high probability hit on a specific target. Thus, to neutralize a specific target, a large number of bombs, and perhaps additional bombing runs, were necessary.

This approach has been replaced with smart munitions. Smart bombs are precision guided, which increases the probability of hitting a desired target. This means fewer bombs, with smaller yields, are needed to achieve mission objectives.

Traditional mindfulness interventions have been under increasing study for adoption in the US military. However, they operate similarly to a dumb bomb approach to learning, where a broad state of mindfulness increases the probability of overall learning, but may miss the targeted learning most valued by the organization.

In this paper we will present a new approach to learning, Guided Mindfulness, which we believe will provide the smart munitions necessary for scalable, adaptable, and personalized development of military personnel in the complex interpersonal competencies necessary for success in the increasingly complex U.S. military milieu.

The United States Military must operate in challenging environments characterized by volatility, ambiguity, and uncertainty. Thus, the military values complex interpersonal skills such as adaptive thinking, adaptive performance [1, 2] and resilience [3].

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 205–213, 2018.
https://doi.org/10.1007/978-3-319-91467-1_17

Complex interpersonal skills can be acquired in the classroom; however, they are better learned through experience [4]. Experiential learning is the process through which knowledge is derived from, and tested through, interactions with the environment. This process relies heavily on reflection and introspection [5]. However, a notable limitation of experiential learning is that it is generally unstructured and therefore idiosyncratic to the learner [6]. To address these learning challenges, we will present the concept of a technology enhanced platform, referred to as Guided Mindfulness (GM), that optimizes experiential learning [7, 8].

The learning needs of people in complex, dynamic (VUCA) organizations can no longer adequately be addressed with standard classroom or even newer online learning techniques [9]. Our objective is to take advantage of new artificially intelligent technologies to enhance real-time learning of complex competencies. While traditional mindfulness has been linked to valued outcomes, the broad focus of the approach may not fit the targeted needs of the military. Instead, GM is narrowly focused on directed, relevant competencies that are driven by the strategic vision and human capital plans of the military training doctrine. While achieving mission related learning outcomes, the GM platform facilitates the second-order goal of strengthening self-regulatory mechanisms that may generalize beyond the focal learning outcomes.

1.1 The Guided Mindfulness AI Platform

The GM platform is a technology-assisted individualized approach to experiential learning that triggers event-based preparation and reflection to increase state mindfulness, train self-regulatory mechanisms, and improve complex skill acquisition. Using an artificially intelligent platform, the learner is directed through the learning experience with prompting questions and activities before, during, and after specific experiential learning events. This just-in-time learning approach involves pre and post assessment, preparation, reflection, and review (see Fig. 1) to facilitate the self-paced directed learning of any interpersonal competency or targeted complex skill. To illustrate how the GM platform will operate, we will present the stages of GM in a 1 year learning cycle characterized by the stages of initial assessment, event based learning, final assessment and review, and final assessment.

Initial Assessment. GM begins with an assessment of the skills that have been targeted for development. A standardized assessment of relevant competencies serves as a baseline that allows the organization to prioritize experiential learning around its talent strategy. This initial assessment is in the form of a customized 360° feedback tool. This 360° feedback provides the learner a baseline from which to understand and leverage strengths and improve upon developmental opportunities [10]. The GM platform incorporates the 360° feedback results and guides the learner through the process of interpreting the results. By accessing a dashboard, the learner can refer to these results throughout the learning cycle. The GM platform will specifically target these competencies and provide data regarding change during the event-based learning phase and final assessment.

Event-Based Learning. Event-based preparation and reflection is the central activity of the GM approach, and is facilitated by an artificially intelligent (AI) platform. In

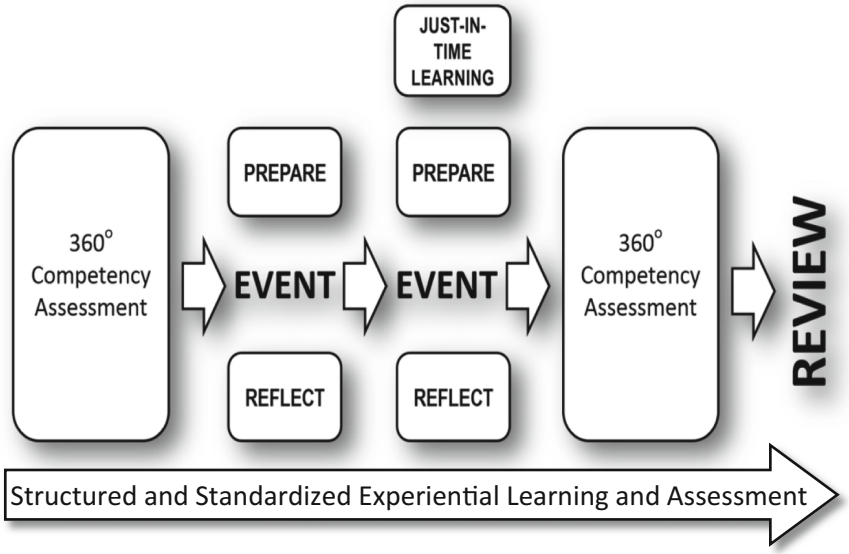

Fig. 1. Guided mindfulness experiential components

practice, the approach is expected to work as follows. The learner first identifies a learning opportunity in an upcoming event. The learner can note this opportunity in their calendar, and "invite" the AI entity to engage in GM related to the event. The GM platform would prompt the learner prior to the scheduled event and instruct her to think about the future event, what competencies are necessary for successful performance, her level of proficiency on those competencies, and possible barriers or roadblocks that may interfere with successful performance. We refer to this stage of the GM process as Prepare (Fig. 2). These questions are both competency-based and event-based and cover sufficient breadth as well as depth for the event. In other words, the questions are similar to the types of questions a coach might ask, thus the GM system is similar to electronic coaching or e-coaching [11]. Through these eliciting questions, the GM system guides the learner through preparation to anticipate and process the actions necessary for success. The learner's responses are the data that are captured and stored in a database for subsequent review.

Assessment and Review. Following the event, the AI entity would prompt the learner with questions requiring reflection on the event, referred to as Reflect (Fig. 2). The learner may be asked to self-assess performance during the event and indicate how the pre-identified competencies contributed to the outcome. In addition, the GM system may ask the learner to discuss the match between pre-event expectations and post-event insights. These types of questions prompt learners to fine tune their sensemaking and engage in simulation [7], deepening their experiential learning and modifying their mental models. Post-event reflection responses can also be collected and stored. Over time, the system gathers data over multiple events that can be sorted by factors such as

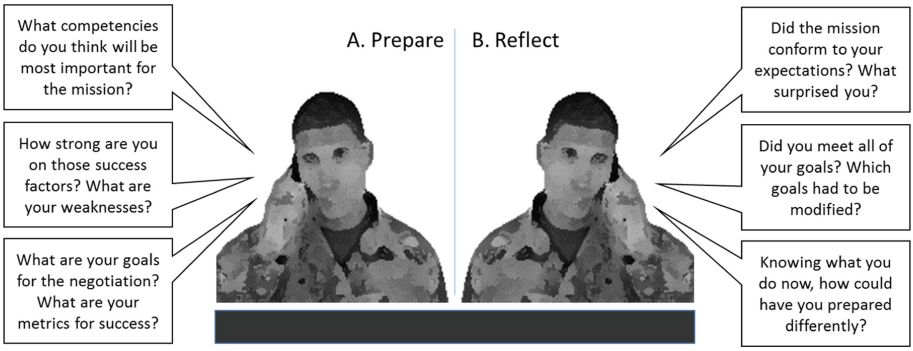

Fig. 2. Illustration of the preparation and reflection stages of guided mindfulness

event, competency, and problems areas. This data can be explored by the learner or aggregated and evaluated at a higher, organizational, level.

Final Assessment. At the end of the yearlong learning cycle, a final 360° feedback assessment is conducted. The GM platform will again compile the results, assist the learner in interpreting the results, and help the learner reflect on improvements as well as continued areas for development.

The combined stages of the GM platform serve to focus attentional resources on the targeted competencies in advance of and after learning opportunities. Thus, "Guided" in this sense does not connote providing a path to general mindfulness. Rather it refers to mindfulness of a specifically defined skillset. This approach differs substantially from traditional mindfulness approaches, which assume that once a state of mindfulness is achieved, the learner maintains a steady state of focused awareness of self and the situation. An implicit assumption is that under these conditions mindfulness generalizes across contexts. We believe this level of mindfulness is not sustainable for learning initiatives, or in some cases desirable in a military context. GM, on the other hand, tightly focuses on organizationally relevant competencies while providing the desirable side effect of strengthening overall self-regulation and thus enhancing other experiential learning opportunities.

1.2 Experiential Learning

Experiential learning is a process through which knowledge is derived from, and tested through, interactions with the environment. This process relies heavily on reflection and introspection [5]. There are many advantages to experiential learning, however, it is generally unstructured and idiosyncratic to the learner [6]. This means that what is learned and how much is learned depends on the individual leaner.

Classroom and other types of structured learning often do not provide sufficient opportunities to experience real-world problems and practice the skills needed to navigate complex environments [12]. Therefore, many professionals view experiential learning as the best approach for complex skill acquisition, particularly for developing complex interpersonal and coping skills such as adaptability [2], leadership [5, 13],

cross-cultural competence (e.g. [14]) and resilience [3]. Learning complex skills through experience (rather than a classroom) has several distinct advantages. First, learning is inherently more relevant to the individual because it is based on unique personal experiences rather than generic one-size-fits all instructional materials. Second, learning from experience ensures a correspondence between newly acquired skills and real-world challenges, eliminating the transfer of training problem. Third, experiential learning is more variable than two dimensional training materials, which results in more integrated, generalizable, and permanent skill acquisition [15].

However, simply providing experiences does not guarantee that experiential learning occurs. Much of the effectiveness of experiential learning is dependent on the self-regulatory processes of the learner [16, 17]. Self-regulation, the inhibition or activation of affective, behavioral, and cognitive processes, allows the learner to focus attention, reflect, and achieve goals [4, 13]. However, personnel in extremely dynamic environments often do not have the spare cognitive resources needed for the self-regulatory activities required to produce effective experiential learning [9]. These obstacles to complex skill acquisition are magnified when the immediate needs of task completion in volatile and unpredictable environments overshadow the chance to learn and continuously improve.

What is needed is a solution that enables mobile, adaptive, moment-of-need access to skill development that can seamlessly train and transfer skills through a work unit or agency. Essentially, the GM platform "jumpstarts" the self-regulatory processes that are necessary to gain from experience. By prompting the learner, the platform can initiate meta-cognition that may not have occurred under non-augmented conditions. Based on a self-regulation view of learning, the GM approach integrates mindfulness techniques into individualized, adaptive, artificially intelligent training that should lead to improved learning outcomes over currently used training methods. Self-regulation theory suggests that a limited amount of resources are available in a given moment [18]. GM facilitates the efficient allocation of these resources, and tightly focuses mindful states associated with learning.

1.3 Mindfulness and Experiential Learning

Mindfulness is comprised of awareness and attention [19]. Awareness refers to a broad observation of the environment, letting stimuli flow through conscious awareness. Attention refers to the focus of that awareness on the present and focal target, without judgement or evaluation. In other words, mindfulness is the non-evaluative objective experience of the environment, rather than perceiving the world through the lens of previous experience, heuristics or other self-relevant and cognitive filters. In the workplace, mindfulness is viewed as a self-regulatory mechanism that works through decreasing the automaticity with which we interact with the environment, decoupling the self from events in order to experience them with less emotion, and increasing self-awareness [20]. A state of mindfulness enables an individual to reflect on an event in the here-and-now, without judgement, and integrate it to build increasingly robust and useful schema.

Reflection is a key process in experiential learning [5] and mindful engagement [21]. It is clear that it is not the provision or completion of an experience that matters to

learning but rather how individuals go through those experiences. Most people are action-oriented and accomplishment focused and tend to live their lives moving from one event to another without fully digesting or reflecting on them. The value of mindfulness is that it promotes non-evaluative reflection. There is some evidence that those with higher levels of mindfulness are better at learning and retention of material [22, 23]. Mindful attention may allow learners to not only deploy their current knowledge to an experience, but also to explore new possibilities which should promote greater learning [24]. In these ways mindfulness can have a positive impact on learning.

1.4 GM as an Improvement Over Traditional Mindfulness Interventions

GM directly impacts both awareness and attention. First, the assessment phase, reflective questioning, and improvement in targeted metrics should enhance the self-awareness of the learner. Over time, the learner will be exposed to quantitative and qualitative indicators of performance in the targeted domain. Thus, they will not only become aware of their current level of performance, but also get a sense of their rate of learning, barriers to learning, and potential moderators of outcomes. Second, the competency-based line of inquiry focuses self-regulatory resources on reflection related to the competencies designed in the system rather than on the idiosyncratic choices of the learner. This focused attention is the rationale behind the "Guided" nomenclature of GM.

The GM process of questioning and reflection is targeted to help learners get the most learning benefits for critical competencies from their on-the-job experiences. Therefore, establishing a targeted state of mindfulness at the outset of a learning experience through the GM Prepare phase should promote greater reflection and subsequent learning during the experiential phases. Moreover, ongoing prompts for reflection after a learning event should stimulate continued mindfulness and eventual learning from the experience. GM is targeted and narrow to specific competencies and learning experiences. It focuses attention on a particular stimulus, in this case an experience, as it is and decouples that attention from meta-awareness - the typical self-referenced evaluation and judgement of the situation. This is different from traditional mindfulness interventions that focus attention on a general stimulus such as breathing. These interventions have been used to promote more general health and well-being [25], whereas the GM approach is developed to contribute to unbiased reflection and subsequent learning of specific competencies from on-the-job experiences.

In essence, the GM platform serves as a coach that is aware of your baseline competency level and learning progress. This coach helps you make sense out of potential opportunities, and focus on opportunities to learn, practice and improve critical interpersonal and coping skills. Rather than simply transmit declarative knowledge like a trainer, the GM platforms prompts metacognitive routines such as simulation and reflection which should result in deep learning [26].

1.5 GM and Self-regulation

Mindfulness is theorized to improve learning through the mechanisms of increased attention, greater cognitive capacity and cognitive flexibility, and decreased emotionality [24]. GM is targeted toward specific self-regulated learning opportunities through the identification of relevant learning events, real-time reflection, and event-based probing questions. GM should enhance experiential learning by impacting self-regulatory processes directly relevant to the controlled processing required for skill acquisition in complex environments. Specifically, GM will improve experiential learning through several intervening self-regulatory processes including self-awareness, situational awareness, social awareness, and sensemaking [7].

GM enhances self-awareness by prompting learners to assess their own skills and competencies for the situation at hand via controlled, in –the-moment, non-judgmental processing. Reflection has been linked to a change in self-perspective and sense of self [21]. GM will also reduce automaticity and thereby increase self-knowledge. GM will improve situational awareness and facilitate understanding of the context and environment. Focusing the learner's attention on situational factors will result in better contingency planning and adaptability, as alternatives will be more easily activated. Social awareness refers to the ability to recognize tacit social cues in order to understand individual and group dynamics and interact effectively [27]. GM will promote social awareness through mindful attention and non-emotional processing of the social environment, and thus enable better management of social relationships. Self, situational, and social awareness will enable sense making which is the process by which people infer meaning from an event and decide on a future course of action [28].

The GM mindful preparation and reflection process focuses the learner's attention on the salient features of a situation which will enable self, situational, and social awareness. This focused attention on target competencies will help the learner derive meaning and ultimately incorporate new information into his or her knowledge structures on skills that will have the most impact on performance.

1.6 Implications and Benefits of the Guided Mindfulness Approach

The proposed GM platform is a flexible, agile, and scalable approach for complex skill acquisition that provides a number of benefits to the learner and his or her unit/agency. First, the GM approach is competency neutral, and therefore can be applied to any complex skills domain (e.g. leadership, negotiation, adaptive performance, cross cultural competence, etc.). In addition, if competency models change, a GM system can be easily modified without making drastic changes to its architecture. While content-based approaches to learning (e.g., classroom learning) need regular modifications to remain current, an event-based reflection approach is relatively content free, resulting in less regular instructional design costs. The GM approach will enable learners to get better at what matters to an organization via the competency-based system. This is in contrast to more general mindfulness interventions which may be associated with improvements in well-being but not necessarily the competencies most mission critical.

Second, rather than being restricted to a classroom or schedule, the proposed GM system would be an agile learning platform. The technology underlying the GM system

could be accessed on a PC, tablet, or smartphone though cloud-based technology. Not only does this allow for rapid relevant learning, but it also reduces training hardware and maintenance costs. This approach is also scalable. Learning need not be limited by the restricted number of seats or instructors in formal training situations.

Finally, the GM platform trains learners to think a particular way-to ask questions and reflect on experiences and answers. This should improve learning specific competencies as well as promoting lifelong learning via improved self-regulatory processes. In other words, traditional mindfulness is a likely result of the GM process, along with competency-specific skills.

1.7 Conclusion

Guided Mindfulness is intended to be a flexible, agile, and scalable approach for improving complex skill acquisition that adds structure to the reflection and mindfulness processes critical for experiential learning. We believe Guided Mindfulness can serve as a tool to optimize the experiential learning necessary to hone complex skills like adaptive performance, and will achieve superior results when compared to traditional mindfulness interventions. In sum, GM is more likely to impact the targeted competency in a precision-guided fashion, whereas traditional mindfulness results in broad learning enhancement, but may miss the mark on essential mission critical skills, much like the iron gravity bombs of an earlier age. To develop the agile leaders for tomorrow's complex and volatile environment, U.S. Military must leverage the best training content available - real-world experiences. Furthermore, they must do so in a manner that is flexible, agile, scalable, and leads to long-term change. The concept of Guided Mindfulness meets these requirements.

References

1. Charbonnier-Voirin, A., Roussel, P.: Adaptive performance: a new scale to measure individual performance in organizations. Can. J. Adm. Sci./Revue Canadienne des Sciences de l'Administration **29**(3), 280–293 (2012)
2. Pulakos, E.D., Schmitt, N., Dorsey, D.W., Arad, S., Borman, W., Hedge, J.W.: Predicting adaptive performance: further tests of a model of adaptability. Hum. Perform. **15**, 299–323 (2002)
3. Bartone, P.T.: Resilience under military operational stress: can leaders influence hardiness? Mil. psychol. **18**(S), S131 (2006)
4. McCall, M.W.: Leadership development through experience. Acad. Manag. Exec. **18**, 127–130 (2004)
5. Kolb, D.A.: Experiential Learning: Experience as the Resource of Learning and Development. Prentice Hall, Englewood Cliffs (1984)
6. Raelin, J.A.: Work-Based Learning: The New Frontier of Management Development. Prentice Hall, Upper Saddle River (2000)
7. Griffith, R.L., Steelman, L.A., Wildman, J.L., LeNoble, C.A., Zhou, Z.E.: Guided mindfulness: a self-regulatory approach to experiential learning of complex skills. Theor. Issues Ergon. Sci. **18**, 147–166 (2017)

8. Griffith, R.L., Sudduth, M.M., Flett, A., Skiba, T.S.: Looking forward: meeting the global need for leaders through guided mindfulness. In: Wildman, J.L., Griffith, R.L. (eds.) Leading Global Teams, pp. 325–342. Springer, New York (2015). https://doi.org/10.1007/978-1-4939-2050-1_14
9. Day, D.V.: Difficulties of learning from experience and the need for deliberate practice. Ind. Organ. Psychol. **3**, 41–44 (2010)
10. Bracken, D.W., Rose, D.S., Church, A.H.: The evolution and devolution of 360 degree feedback. Ind. Organ. Psychol. **9**, 761–794 (2016)
11. Rossett, A., Marino, G.: The art of feedback: if coaching is good then e-coaching is. Train. Dev. **59**, 46–49 (2005)
12. London, M., Mone, E.M.: Continuous learning. In: Pulakos (ed.) The Changing Nature of Performance: Implications for Staffing, Motivation, and Development, pp. 119–153 (1999)
13. McCall, M.W.: Recasting leadership development. Ind. Organ. Psychol. **3**, 3–19 (2010)
14. Ng, K.Y., Van Dyne, L., Ang, S.: From experience to experiential learning: cultural intelligence as a learning capability for global leader development. Acad. Manag. Learn. Educ. **8**, 511–526 (2009)
15. Gupta, A.K., Govindarajan, V.: Cultivating a global mindset. Acad. Manag. **16**, 116–126 (2002)
16. Kanfer, R.: Self-regulatory and other non-ability determinants of skill acquisition. In: Gollwitzer, P.M., Bargh, J.A. (eds.) The Psychology of Action: Linking Cognition and Motivation to Behavior, pp. 404–423. Guilford Press, New York (1996)
17. Sitzmann, T., Ely, K.: A meta-analysis of self-regulated learning in work-related training and educational attainment: what we know and where we need to go. Psychol. Bull. **137**, 421 (2011)
18. Baumeister, R.F.: Ego depletion and self-control failure: an energy model of the self's executive function. Self Identity **1**(2), 129–136 (2002)
19. Brown, K.W., Ryan, R.M.: The benefits of being present: mindfulness and its role in psychological well-being. J. Pers. Soc. Psychol. **84**, 822–848 (2003)
20. Glomb, T.E., Duffy, M.K., Bono, J.E., Yang, T.: Mindfulness at work. Res. Pers. Hum. Resour. Manag. **30**, 115–157 (2011)
21. DeRue, S.D., Ashford, S.J.: Power to the people: where has personal agency gone in leadership development? Ind. organ. psychol. **3**(1), 24–27 (2010)
22. Calma-Birling, D., Gurung, R.A.: Does a brief mindfulness intervention impact quiz performance? Psychol. Learn. Teach. **16**(3), 323–335 (2017)
23. Bonamo, K.K., Legerski, J.P., Thomas, K.B.: The influence of a brief mindfulness exercise on encoding of novel words in female college students. Mindfulness **6**(3), 535–544 (2015)
24. Good, D.J., Lyddy, C.J., Glomb, T.M., Bono, J.E., Brown, K.W., Duffy, M.K., Baer, R.A., Brewer, J.A., Lazar, S.W.: Contemplating mindfulness at work: an integrative review. J. Manag. **42**, 114–142 (2016)
25. Kabat-Zinn, J.: Mindfulness-based interventions in context: past, present, and future. Clini. Psychol. Sci. Pract. **10**, 144–156 (2003)
26. Rekalde, I., Landeta, J., Albizu, E., Fernandez-Ferrin, P.: Is executive coaching more effective than other management training and development methods? Manag. Decis. **55**, 2149–2162 (2017)
27. Hilton, R.M., Shuffler, M., Zaccaro, S.J., Salas, E., Chiara, J., Ruark, G.: Critical social thinking and response training: a conceptual framework for a critical social thinking training program. (ARI research report). Army Research Institute for the Behavioral and Social Sciences, Arlington (2009)
28. Weick, K.E., Sutcliffe, K.M., Obstfeld, D.: Organizing and the process of sensemaking and organizing. Organ. Sci. **16**, 409–421 (2005)

Curriculum for Accelerated Learning Through Mindfulness (CALM)

Anna Skinner[1(✉)], Cali Fidopiastis[1], Sebastian Pascarelle[2], and Howard Reichel[2]

[1] Design Interactive, Inc., Orlando, FL 32817, USA
{anna.skinner, cali.fidopiastis}@designinteractive.net
[2] In-Depth Engineering Corp., Fairfax, VA 22030, USA
{sam.pascarelle, howard.reichel}@indepth.com

Abstract. The military training community is faced with the daunting task of providing each and every warfighter with basic, journeyman and advanced training courses – using media and methodologies that permit rapid, efficient learning and transfer of the learning to a wide range of operational tasks. There is a need for a methodology and metrics to assess the best combinations of learning techniques that can be applied across various types of military training systems and a training testbed with which to assess individual and group characteristics that can accelerate the speed of learning, increase comprehension and retention, and improve transfer of training to performance on operational tasks. Accelerated learning is comprised of two primary components: accelerating the learning pathway and accelerating the learning process. The Curriculum for Accelerated Learning through Mindfulness (CALM) theoretical model seeks to combine these two components and to close the loop between the two, providing real-time correlation of training performance to cognitive state metrics and subsequent adaptation of training content (e.g., complexity/difficulty, modality, scaffolding) in order to maintain an optimal and accelerated state of learning.

Keywords: Accelerated learning · Brain sensors and measures
Effects of stress & cognitive load on performance
Measuring and adapting to individual differences
Sensor integration to characterize operator state

1 Introduction

1.1 Challenges in Optimizing Military Training

The military training community is faced with the daunting task of providing each and every warfighter with basic, journeyman and advanced training courses – using media and methodologies that permit rapid, efficient learning and transfer of the learning to a wide range of operational tasks. For example, the navy has expressed the need for each sailor to improve his or her training efficiency and learning retention for complex naval tasking by a factor of two times or more. Therefore, there is a need to discover accelerated learning techniques through valid and reliable metrics and the comparison of human behavioral techniques that support transfer of training to the complex

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 214–226, 2018.
https://doi.org/10.1007/978-3-319-91467-1_18

operational environment. Optimized training requires a notable change in the learning paradigm – a training paradigm in which the training system assists the learner in an effort to enter and maintain an optimal cognitive state for learning. Such a state should provide for enhanced skill acquisition and more rapid retrieval of information from memory, as well as the ability to sustain focused attention for longer periods of time and deeper levels of information processing (Wickens and Hollands 2000).

The problem faced is multifaceted in that accelerated learning pedagogy has not been explicated for complex training systems such as those found in the military. Further, optimal cognitive states vary across learning and performance domains, as well as within and across individuals. Even for the same task, optimal states have been shown to differ across individuals. For example, it has been shown that some individuals perform best when in higher states of arousal than others. Studies of accelerated learning are also complicated by findings by training experts that trainees often do not arrive at formal training sessions prepared to learn. Trainees may arrive in various states of arousal, fatigue, and anxiety based upon events of their personal and professional lives (e.g. marital problems, financial issues, multiple conflicting demands at work), leading to distraction and difficulty in assimilating complex training content. Further, trainees with less ambient arousal and anxiety are more likely to successfully focus and absorb the material presented. Research has shown that trainees who enter training with a negative brain state exhibit sub-par performance and learning as evidence by the measured modulation of encoding and retrieval processing within their memory systems (Margraf and Zlomuzica 2015). However, most individuals are not aware when their minds are closed to learning, and a suboptimal trainee learning state may not be apparent to an instructor.

1.2 Optimal Cognitive States for Learning

Research in the field of Neuroergonomics (the study of how the brain performs in operational environments), indicates that the type of sustained attention necessary for skill learning and safety monitoring while engaged in difficult operational tasks (Parasuraman 2000) is enhanced through techniques such as mindfulness. Self-regulation and self-management of the brain networks that underlie mindfulness lead to enhanced brain connections necessary for improving data processing and memory retrieval (Shapiro et al. 2006) and enhanced executive control over inflexible biased response (Teper and Inzlicht 2013). Thus, for deeper, accelerated learning of complex occupational skills, military personnel need to be engaged, involved, and mindful before and during the training process. The ability to improve engagement during training holds the potential to increase the transfer of training, leading to enhanced readiness and success on the across a wide range of military operational domains and tasks.

Current research also suggests that the external (learning material) and the internal (mental states) aspects of a learner should be in balance or in predictable order for optimal learning to occur (Csíkszentmihályi 2014, p.211). Csíkszentmihályi (2014) has further characterized this optimal learning state as a “flow” state - a state of heightened focus and immersion in activities such as art, play and work. Mindfulness may act to clear and awaken one’s mind, leading to reduced anxiety and improved access to

attentional networks. Entering flow state may then be enhanced increasing focused attention and heightening information processing in that domain. To maintain an optimal state of learning, the training should have clear tasks, optimal challenges, and provide clear and immediate feedback. In addition, the training must match the learner's current knowledge state and simultaneously challenge the trainee through a cooperative learning strategy such that he or she can perform the tasks alone with insight to problem solving beyond the formal instruction (Vygotsky 1987).

Thus, there is a need for a methodology and metrics to assess the best combinations of learning techniques that can be applied across various types of military training systems and a training testbed with which to assess individual and group characteristics that can accelerate the speed of learning, increase comprehension and retention, and improve transfer of training to performance on operational tasks.

1.3 Adaptive Training and Augmented Cognition

Vygotsky (1987) described the concept of the zone of proximal development (ZPD) as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (p. 86). Essentially, the ZPD describes what one needs to learn with assistance of an expert, such as in a train-the-trainer paradigm, or what the learner can do without assistance. ZPD requires monitoring of in-situ performance and providing assistance or feedback from an expert trainer or a team of peers, either simulated or live. Metrics for assessing mindfulness and monitoring flow currently consist of self-report questionnaires and interviews that are non-invasive and inexpensive; however, memory bias and task interruption affect the reliability and validity of these methods. To demonstrate efficient and effective methods that increase the sailor's learning speed, comprehension, and performance requires an ability to measure cognitive and affective states of the learner throughout the training experience. Therefore, there is a need to discover accelerated learning techniques through valid and reliable metrics and the comparison of human behavioral techniques that support transfer of training to the complex operational environment.

In addition to subjective measures of flow such as the 36-item Flow State Scale (FSS) (Jackson and Marsh 1996), several more recent research efforts have been undertaken to use psychophysiological measures to detect and characterize flow state using a wide range of sensor technologies (Berta et al. 2013). One such study demonstrated that psychophysiological data and pupil dilation characteristics were significantly different while using Facebook, an activity thought to be highly engaging over extended periods of time, as compared to induced stress and relaxation conditions on multiple linear and spectral indices of somatic activity. The psychophysiological state evoked in the Facebook condition was characterized by high positive valence and high arousal, which have been used to define Core Flow State (Mauri et al. 2011). Research has also indicated that flow experiences may combine subjectively positive elements with physiological elements associated with strainful tension and mental load based on heart rate variability and salivary cortisol (Keller et al. 2011). Similary, Nacke and Lindley (2008) demonstrated video gameplay scenarios designed for

combat-oriented flow experiences demonstrated measurable high-arousal positive affect emotions, which were consistent across both subjective (questionnaire responses) and objective measures collected (electroencephalography, electrocardiography, electromyography, galvanic skin response and eye tracking).

Hou and Fidopiastis (2017) developed an adaptive intelligent tutoring methodology for accelerating learning for unmanned system operators that extended the research and development of the Defense Advanced Research Projects Agency's (DARPA) Augmented Cognition (AugCog) program. The purpose of the AugCog training architecture was to evaluate best practices for designing interfaces that facilitated trainee learning by maintaining the trainee in the optimal state of learning (Nicholson et al. 2007; Hou and Fidopiastis 2017). The successful mapping of psychophysiological human state changes in adaptive tutoring systems has been demonstrated by the AugCog research community (St. John et al. 2003; Palmer and Kobus 2007). Specifically, cognitive and affective state changes of the learner are measurable using psychophysical metrics such as electroencephalography (EEG), heart rate variability (HRV), and electrodermal response (EDR) to identify cognitive workload (Sciarini and Nicholson 2009), engagement (Berka et al. 2007), and negative emotional states (Vartak et al. 2008). However, for proper effectiveness or training transfer to the field environment, the training system must adapt the learning content base on the skill level of the learner and optimize the learner states of engagement, performance, cognitive workload, and affect throughout the training experience (Hou and Fidopiastis 2014, 2017).

1.4 Accelerated Learning Operational Definition and Application

Accelerated Learning has been operationally defined as "The reduction of learner time required to meet learning objectives in a training event (Hoffman et al. 2010, p. 400)". According to Hoffman et al. accelerated learning is comprised of two primary components: accelerating the learning pathway and accelerating the learning process. The former involves enabling the learner to cover and master material in less time, while the latter involves employing instructional design principles that increase learner engagement with the material. The remaining sections of this paper present a theoretical approach and initial instantiation of a method by which to apply and operationalize this definition of accelerated learning within the context of a military training task.

2 Curriculum for Accelerated Learning Through Mindfulness (CALM)

2.1 CALM Theoretical Model and Testbed System

The Curriculum for Accelerated Learning through Mindfulness (CALM) theoretical model seeks to combine the two components of accelerated learning as defined by Hoffman et al. (2010): accelerating the learning pathway and accelerating the learning process. This model has been instantiated within a prototype testbed system involving a mindfulness intervention that prepares the sailor for learning and a methodology to

assess the flow state of the learner throughout a contextually relevant Naval learning experience using psychophysiological measures. This model is based on the premise that a trainee who is mindful (engages in focused attention) and in the flow (meeting the learning material with appropriate cognitive challenge and skill level) will maintain optimal cognitive and affective states that prepare his brain to learn more effectively and efficiently, produce higher and longer retention of the training material, and facilitate transfer of that training to the operational environment.

Accelerating the Learning Pathway. This component of accelerated learning involves enabling the learner to progress through and master material in less time. Within the CALM theoretical model this is achieved via a paradigm that uses an unconventional application of Item Response Theory (IRT) to drive an adaptive trainer that automatically determines the appropriate training and testing level of difficulty for a student based on his or her *in-situ* measured ability level. This component of the CALM testbed consists of an existing adaptive training system, the Adaptive Gaming Environment for Submarines (AGE-S), which provides computer-based Electronic Warfare (EW) Support (ES) operator proficiency instruction and assessment relevant to Electronic Support Measures (ESM).

Each training module within the AGE-S system contains a set of instructional materials, which consist of text, pictures, sound files, and short movie clips to convey the basic concepts underlying the skills to be taught. Following the instructional material, the module contains several groupings of questions. Each question consists of an EW scenario captured from an Advanced Submarine Tactical ESM Combat System (ASTECS) emulator either as a static picture, a sound file, or a movie file, plus a multiple choice question to go with the scenario. The groups of questions range from very simple, covering the most basic concepts, to intermediate difficulty, where part-task skills are tested and the number and complexity of sonar emitters is increasing, to the most difficult questions, which contain movies and complicated scenarios that closely parallel whole-task skill testing. Questions are binned into five levels of difficulty (Easiest, Easy, Medium, Hard, Hardest). The adaptive engine measures a student's performance on the questions and determines whether to provide questions of the same, easier, or harder level of difficulty. Each time a question is answered incorrectly, process feedback is provided in the form of a brief video clip containing step by step instructions, similar to a worked example, detailing the way in which the trainee should have completed the exercise to arrive at the right answer. When a question is answered correctly, the system indicates that the right answer was provided, and the next question is immediately presented, with no additional feedback.

As such, adept students will move quickly through the material, without pausing for additional instruction or feedback, progressing to increasingly harder questions and completing the module in less time, while less proficient students will take longer to progress through the course material, but will receive the extra instruction they need along the way to successfully complete the module.

Accelerating the Learning Process. This component of accelerated learning involves employing instructional design principles that increase learner engagement with the material. Within the CALM theoretical model, during training content development and validation, this is achieved by using psychophysiological measures to objectively

assess cognitive engagement across multiple individuals while interacting with instructional and assessment materials having varying levels of difficulty. This component of the CALM prototype system consists of a Testbed for Intelligent Tracking and Real-time Assessment of Trainee Engagement (TITRATE), comprising a physiological sensor suite and validated analytic algorithms for evaluating a subject's cognitive state in real-time while completing AGE-S assessment questions across all five levels of difficulty. These algorithms emphasize detection of optimal levels of engagement for a particular individual using a comparison to baseline and comparison to performance outcomes methodology. The current TITRATE prototype system uses the *B-Alert*^{tm} X10 (Advanced Brain Monitoring, Carlsbad, CA) mobile sensor system, a commercial off the shelf (COTS) technology that provides nine channels of high-quality EEG, plus one optional channel for Electrocardiography (ECG), Electromyography (EMG), or Electrooculography (EOG). The current prototype also uses a previously developed and validated engagement metric, which uses discriminant function analysis (DFA) methods to derive a four-class quadratic DFA model to distinguish high engagement, low engagement, distraction, and sleep onset classifiers.

Closing the Loop. Additionally, the CALM theoretical model advances current training paradigms by integrating a mindfulness module (MindMod) that prepares the trainee for learning or assists the trainee in returning to an optimized cognitive state for learning when engagement levels become too low. The MindMod component of the CALM prototype system consists of a series of mindfulness exercises, including guided meditation, which can be selected based on trainee preference, and which will be objectively assessed for efficacy under future research.

Finally, the envisioned CALM testbed system will include an Adaptive Driver to Augment Performance and Training (ADAPT) Application Programming Interface (API) that enables real-time correlation of training performance to cognitive state metrics and subsequent adaptation of training content (e.g., complexity/difficulty, modality, scaffolding) in order to maintain an optimal and accelerated state of learning.

2.2 CALM System Proof of Concept Functionality Testing

Proof of concept functionality testing was conducted in order to verify the ability of the prototype testbed system to (1) collect clean and complete physiological sensor data sets; (2) accurately synchronize physiological data timestamps to task related timestamps, task events, and relevant contextual information (e.g., start and stop of each question, question level of difficulty, answer accuracy); (3) assess the sensitivity of the TITRATE sensors and algorithms with respect to detection of potential changes in user engagement while completing the AGE-S module questions at varying levels of difficulty; (4) assess the sensitivity of the TITRATE sensors and algorithms with respect to detection of potential differences in cognitive state between a regular meditator and a novice meditator; and (5) assess the sensitivity of the TITRATE sensors and algorithms with respect to detection of potential changes in cognitive state before and after a brief meditation session. Functionality testing data were collected for two individuals, one who has maintained a dedicated meditation practice over many years and one having no prior meditation experience. Both individuals, having had prior training and

experience in using the AGE-S training system, completed a total of 100 AGE-S questions, across all five levels of difficulty, as well as a five-minute meditation session half-way through the testing session. The 100 questions were broken into 10 blocks, each with a specified level of difficulty, ranging from Easiest to Hardest. As shown in Fig. 1, each participant completed 10 Easiest, 10 Easy, 10 Medium, 10 Hard, and 10 Hardest questions, followed by the five-minute meditation (indicated by the green arrow as question #51). Both participants then completed 10 Hardest, 10 Hard, 10 Medium, 10 Easy, and 10 Easiest questions. EEG and ECG were collected for both participants throughout the testing and meditation sessions.

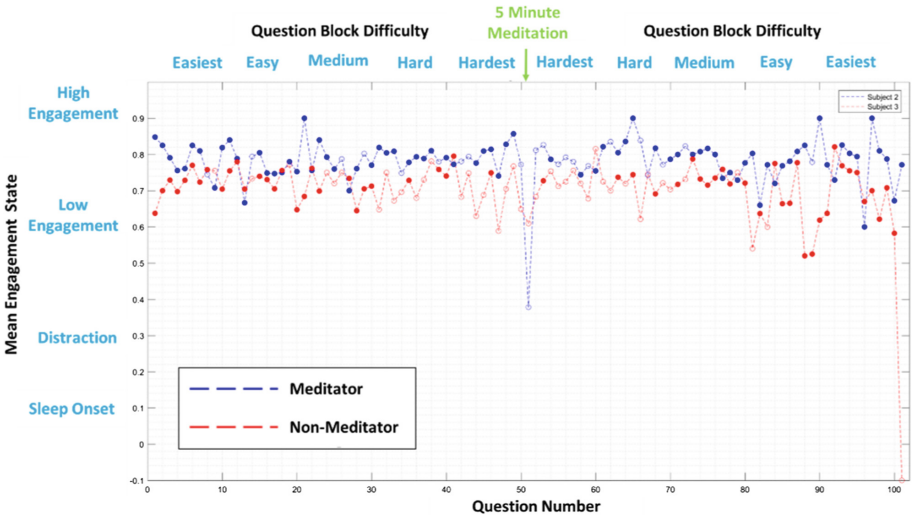

Fig. 1. Mean engagement state for the meditator and non-meditator across all 100 questions and the five-minute meditation session (represented as Question #51). (Color figure online)

Figure 1 provides average estimates of the highest probability EEG-based engagement state for the meditator (represented in blue) and the non-meditator (represented in red) across 100 AGE-S questions answered the five-minute meditation session (indicated with a green arrow and represented as Question #51). The engagement algorithm estimates the mostly likely state of the user, selected from Sleep Onset (SO), Distraction (DIS), Low Engagement (LE), or High Engagement (HE) on a second by second basis. The algorithm indicates the most likely state for each second with a specified value assigned to each state as follows: 0.1 (SO), 0.3 (DISS), 0.6 (LE), or 0.9 (HE). Given that each question took longer than one second to answer, and the times to complete each question varied, multiple state estimations were recorded for each question, and the estimated states may have changed over the course of each question. Initial analyses first evaluated the average estimated state over the course of each question and the five-minute meditation session. As such, the plotted points in Fig. 1 do not fall strictly at the specified level of a specific engagement state, but rather

indicate an average level across the four potential states for each question and during meditation.

In addition to displaying the engagement probability for each question answered, as well as during the five-minute meditation, Fig. 1 also indicates whether each question was answered correctly or incorrectly. The circular symbol on the line graph for each question is filled in if the question was answered correctly; if the symbol is empty, this indicates that the question was answered incorrectly.

The regular meditator (shown in blue), performed very well on the first 50 questions, answering only one incorrectly in the Easiest block, one incorrectly in the Easy block, two incorrectly in the Medium difficulty block, two incorrectly in the Hard block, and two incorrectly in the Hardest block. Upon beginning the second block of Hardest difficulty questions (starting with Question #52), immediately following the five-minute meditation, the regular meditator answered three questions incorrectly. In the subsequent Hard difficulty question block, he answered four incorrectly, followed by just one incorrect answer in the Medium difficulty question block, one incorrectly in the Easy block, and no incorrect answers in the final Easiest block of questions.

The non-meditator performed well on the first block of Easiest questions, getting only one incorrect, and on the Easy questions, getting two incorrect. However, he began to struggle more with the Medium questions, getting three incorrect, and then really struggled with the Hard and Hardest difficulty question blocks, answering seven and eight incorrect, respectively. Following the meditation session, the non-meditator continued to struggle, answering nine of the second block of Hardest questions incorrectly and seven of the Hard questions incorrectly. He then began to recover in performance during the Medium question block, answering only two incorrectly, and then maintained good performance throughout the remainder of the questions, answering just two Easy questions incorrectly, and none of the final block of Easiest questions incorrectly.

While, on average, both individuals remained somewhere between Low Engagement and High Engagement throughout the testing and meditation session, the meditator maintained consistently higher engagement levels than the non-meditator throughout all the question blocks, which may have contributed to better performance. For some questions, the meditator's average engagement value was 0.9, indicating that he was in a state of High Engagement the entire time he was answering that question (e.g., Questions #21, #65, #90, and #97). Notably, all four of those questions were answered correctly.

Interestingly, the meditator displayed markedly lower engagement levels during the five-minute meditation session (indicated as Question #51). This result is to be expected with a highly practiced meditator, who is likely able to go into very deep states of consciousness very quickly; those skilled in meditative practices are extremely good at disengaging from active thought processes and unintended distracting thoughts, even within a short meditation session. During longer sessions, these deeper states are very similar to sleep onset from an engagement perspective, and while experienced meditators can get into these states fairly quickly, practiced meditators typically bring themselves out of those states slowly at the end of a meditation session. In this case, immediately following the short (five-minute) meditation, the regular meditator returned to pre-meditation engagement level, but performed poorly much more poorly

on the Hardest block of questions as compared to his performance on the Hardest block preceding meditation. This may be an indication that he went into an extremely low engagement state during meditation and was still coming out of that state when attempting to answer the questions in the next block. This is to be expected with a highly practiced meditator.

The non-meditator not only exhibited lower engagement levels overall, for some questions, his average engagement state fell below the threshold for Low Engagement of 0.6, indicating that during at least part of the time while answering those questions, his highest probability cognitive state was Distraction (e.g., Questions #47, #81, #83, #88, and #89). The non-meditator's average engagement level during the meditation was markedly higher than the regular meditators, remaining in an average state of Low Engagement throughout the meditation exercise. This is typical for individuals inexperienced in meditation, who may find it very difficult to clear their minds, reducing focused attention and dismissing distracting thoughts.

In order to better represent the highest probability engagement state (i.e., the state that was most prevalent for each question), the mode value is presented in Fig. 2 for each question, as well as the mode value for the five-minute meditation session (again shown as question #51).

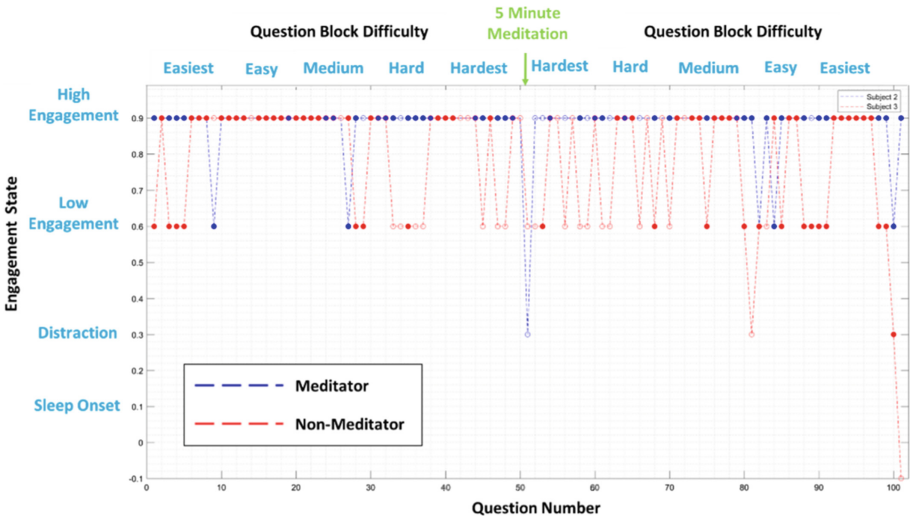

Fig. 2. Mode cognitive state (level of engagement) for the experienced meditator and non-meditator across all 100 questions and the five-minute meditation session (represented as Question #51).

Figure 2 provides a clearer picture of the cognitive state levels for both the meditator and non-meditator across the entire data collection session. The regular meditator was in a state of High Engagement for 95 of the 100 questions, dropping to Low Engagement for the remaining five questions, but only dropping below Low Engagement during meditation. Conversely, the non-meditator was only in a state of

High Engagement for 62 of the 100 questions, dropping to a state of Low Engagement for 37 questions; and furthermore, the non-meditator did not drop below Low Engagement during meditation, but did drop down into the Distraction state on Question #80.

3 Conclusions

Taken together, these data results indicate very clear differences between the two individuals, with the regular meditator maintaining higher engagement overall and better performance overall, particularly for the harder difficulty questions before the meditation session. It is possible that the non-meditator performed more poorly due to lower engagement, but it is also possible that the meditator simply has greater knowledge of the subject matter and the AGE-S question types. Most interestingly, the two participants demonstrated very different engagement levels during the meditation session, with the meditator exhibiting much lower engagement while completing the meditation session. As such, it appears that the meditator was better able to maintain higher levels of focused attention over a long period of time while answering questions, but was also able to quickly drop into a state of lower engagement while meditating, possibly due in part to his regular meditation practice.

While investigators (Kabat-Zinn 1990, 1998; Shapiro and Schwartz 2000; Teasdale 1999; Segal et al. 2002), provide descriptions of what mindfulness may be, the field lacked an operational definition from which to create testable hypotheses and determine the utility of such a construct in everyday life (Bishop 2002). Mindfulness, for the purpose of this effort, is operationally defined as a basic human ability to regulate the focus of attention toward current actions and events, without influence of personal affective states, such as when experiencing anxiety or arousal (Bishop et al. 2004). A mindful brain state potentially translates into learning through the breaking of brain habits that keep the brain inflexible to learning new concepts and strategies important for performing tasks in dynamically changing operational environments, such as military contexts (Langer 2000). Langer (2000) also suggests that current curricula are setup for mindless learning through repetition and single exposure. The brain is working in mindless mode when it (1) operates out of an inflexible habit and (2) when it biases to a particular perspective. In each case, the brain is not allowed to extend learning when the context of application changes.

Inherent in the mindless mode of brain operation during training is the lack of control or self-regulation of focused attention (i.e. engagement) by the trainee. Research in the operational environment that utilizes brain state measures such as electroencephalography (EEG) indicate that mindfulness improves the capability of the brain to engage in the type of focused and sustained attention necessary for skill learning and safety monitoring (Parasuraman 2000). Other benefits of improved self-regulation and self-management of the brain networks within the training context include enhanced neural connections necessary for improving data processing and memory retrieval (Shapiro et al. 2006) and enhanced executive control over inflexible biased response (Teper and Inzlicht 2013). The ability to improve engagement during training holds the potential to deepen learning such that the transfer of skills training to the military

context is quicker and more sustained, leading to enhanced readiness and success on the battlefield.

There are potentially two components of mindfulness relevant for studying the effects of mindfulness on training complex military skills: (1) acquiring skills in the self-regulation of attention and (2) adopting an acceptance toward life's experiences (Bishop et al. 2004). The concept of "staying in the moment" relies upon the brain's ability to process the immediate experience without distraction. Distraction and boredom are cognitive states that research suggests inhibits or preoccupies the attentional network, and therefore negatively impacts learning (Eastwood et al. 2012). A person who is more in control of their focused attention can potentially shift attention quickly back to the immediate task if the brain becomes distracted. There are more brain resources to process the training material. Additionally, the attitude one has toward the tasking of the current moment can affect how deeply one learns. For example, curricula that do not foster perspective taking (passive) and are repetitive (boring) can create a context where trainees cannot maintain attentional focus or switch from where the brain wanders back to the task (Posner 1980). Consequently, the training information is not deeply process such that learning takes longer and transfer of training is improbable (Langer and Moldoveanu 2000). Following a flow state methodology where the learning material and the internal mental states of the trainee are kept in balance during training may be a better predictor of accelerated learning and training effectiveness than other constructs such as cognitive load (Fidopiastis 2011). Future research is needed to formalize and validate this theoretical model within an experimental paradigm that allows for further exploration of the effects of various instructional design and adaptive training techniques, as well as various mindfulness interventions, on both performance and learner engagement.

Acknowledgements. This work was funded by a Phase I Small Business Innovation Research (SBIR) contract (N00178-17-C-1133).

References

Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat. Space Environ. Med. **78**(5), B231–B244 (2007)

Berta, R., Bellotti, F., De Gloria, A., Pranantha, D., Schatten, C.: Electroencephalogram and physiological signal analysis for assessing flow in games. IEEE Trans. Comput. Intell. AI Games **5**(2), 164–175 (2013)

Bishop, S.R.: What do we really know about mindfulness-based stress reduction? Psychosom. Med. **64**(1), 71–83 (2002)

Bishop, S.R., Lau, M., Shapiro, S., Carlson, L., Anderson, N.D., Carmody, J., Segal, Z.V., Abbey, S., Speca, M., Velting, D., Devins, G.: Mindfulness: a proposed operational definition. Clin. Psychol. Sci. Pract. **11**(3), 230–241 (2004)

Csikszentmihalyi, M.: Toward a psychology of optimal experience. In: Flow and the Foundations of Positive Psychology, pp. 209–226. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_14

Eastwood, J.D., Frischen, A., Fenske, M.J., Smilek, D.: The unengaged mind: defining boredom in terms of attention. Assoc. Psychol. Sci. **7**(5), 482–495 (2012)

Fidopiastis, C.M.: Theoretical transpositions in brain function and the underpinnings of augmented cognition. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) FAC 2011. LNCS (LNAI), vol. 6780, pp. 153–158. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21852-1_19

Hoffman, R.R., Andrews, D., Fiore, S.M., Goldberg, S., Andre, T., Freeman, J., Fletcher, J.D., Klein, G.: Accelerated learning: prospects, issues and applications. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Sage, CA, vol. 54, no. 4, pp. 399–402. SAGE Publications, Los Angeles (2010)

Hou, M., Fidopiastis, C.M.: Untangling operator monitoring approaches when designing intelligent adaptive systems for operational environments. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2014. LNCS (LNAI), vol. 8534, pp. 26–34. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07527-3_3

Hou, M., Fidopiastis, C.: A generic framework of intelligent adaptive learning systems: from learning effectiveness to training transfer. Theoret. Issues Ergon. Sci. **18**(2), 167–183 (2017)

Jackson, S.A., Marsh, H.W.: Development and validation of a scale to measure optimal experience: the flow state scale. J. Sport Exerc. Psychol. **18**(1), 17–35 (1996)

Kabat-Zinn, J.: Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness. Delacorte, New York (1990)

Kabat-Zinn, J.: Meditation. In: Holland, J.G. (ed.) Psychoncology, pp. 767–779. Oxford University Press, New York (1998)

Keller, J., Bless, H., Blomann, F., Kleinböhl, D.: Physiological aspects of flow experiences: skills-demand-compatibility effects on heart rate variability and salivary cortisol. J. Exp. Soc. Psychol. **47**(4), 849–852 (2011)

Langer, E.J., Moldoveanu, M.: The construct of mindfulness. J. Soc. Issues **56**(1), 1–9 (2000)

Margraf, J., Zlomuzica, A.: Changing the future, not the past: a translational paradigm shift in treating anxiety. EMBO Rep. **16**, 259–260 (2015)

Mauri, M., Cipresso, P., Balgera, A., Villamira, M., Riva, G.: Why is Facebook so successful? Psychophysiological measures describe a core flow state while using Facebook. Cyberpsychol. Behav. Soc. Netw. **14**(12), 723–731 (2011)

Nacke, L., Lindley, C.A.: Flow and immersion in first-person shooters: measuring the player's gameplay experience. In: Proceedings of the 2008 Conference on Future Play: Research, Play, Share, pp. 81–88. ACM (2008)

Nicholson, D.M., Fidopiastis, C.M., Davis, L.D., Schmorrow, D.D., Stanney, K.M.: An adaptive instructional architecture for training and education. In: Schmorrow, D.D., Reeves, L.M. (eds.) FAC 2007. LNCS (LNAI), vol. 4565, pp. 380–384. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73216-7_43

Palmer, E.D., Kobus, D.A.: The future of augmented cognition systems in education and training. In: Schmorrow, D.D., Reeves, L.M. (eds.) FAC 2007. LNCS (LNAI), vol. 4565, pp. 373–379. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73216-7_42

Parasuraman, R.: Designing automation for human use: empirical studies and quantitative models. Ergonomics **43**(7), 931–951 (2000)

Posner, M.I.: Orienting of attention. Q. J. Exp. Psychol. **32**(1), 3–25 (1980)

Sciarini, L.W., Nicholson, D.: Assessing cognitive state with multiple physiological measures: a modular approach. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) FAC 2009. LNCS (LNAI), vol. 5638, pp. 533–542. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02812-0_62

Segal, Z.V., Teasdale, J.D., Williams, J.M., Gemar, M.C.: The mindfulness-based cognitive therapy adherence scale: inter-rater reliability, adherence to protocol and treatment distinctiveness. Clin. Psychol. Psychother. **9**(2), 131–138 (2002)

Shapiro, S.L., Carlson, L.E., Astin, J.A., Freedman, B.: Mechanisms of mindfulness. J. Clin. Psychol. **62**(3), 373–386 (2006)

Shapiro, S.L., Schwartz, G.E.: Intentional systemic mindfulness: an integrative model for self-regulation and health. Adv. Mind-Body Med. **16**, 128–134 (2000)

St. John, M., Kobus, D.A., Morrison, J.G.: DARPA augmented cognition technical integration experiment (TIE), no. TR-1905. Pacific Science and Engineering Group, Inc., San Diego (2003)

Teasdale, J.D.: Emotional processing, three modes of mind and the prevention of relapse in depression. Behav. Res. Ther. **37**, S53–S77 (1999)

Teper, R., Inzlicht, M.: Meditation, mindfulness and executive control: the importance of emotional acceptance and brain-based performance monitoring. Soc. Cogn. Affect. Neurosci. **8**(1), 85–92 (2013)

Vartak, A.A., Fidopiastis, C.M., Nicholson, D.M., Mikhael, W.B., Schmorrow, D.: Cognitive state estimation for adaptive learning systems using wearable physiological sensors. In: BIOSIGNALS (2), pp. 147–152 (2008)

Vygotsky, L.: Zone of proximal development. Mind Soc.: Dev. High. Psychol. Process. **5291**, 157 (1987)

Wickens, C.D., Hollands, J.G.: Attention, time-sharing, and workload. Eng. Psychol. Hum. Perform. **3**, 439–479 (2000)

Augmented Reality for Tactical Combat Casualty Care Training

Glenn Taylor[1(✉)], Anthony Deschamps[1], Alyssa Tanaka[1], Denise Nicholson[1], Gerd Bruder[2], Gregory Welch[2], and Francisco Guido-Sanz[2]

[1] Soar Technology, Ann Arbor, MI 48105, USA
{glenn, anthony.deschamps, alyssa.tanaka, denise.nicholson}@soartech.com
[2] University of Central Florida, Orlando, FL 32816, USA
{gerd.bruder, welch, frank.guido-sanz}@ucf.edu

Abstract. Combat Life Savers, Combat Medics, Flight Medics, and Medical Corpsman are the first responders of the battlefield, and their training and skill maintenance is of preeminent importance to the military. While the instructors that train these groups are exceptional, the simulations of battlefield wounds are extremely simple and static, typically consisting of limited moulage with sprayed-on fake blood. These simple presentations often require the imagination of the trainee and the hard work of the instructor to convey a compelling scenario to the trainee. Augmented Reality (AR) tools offer a new and potentially valuable tool for portraying dynamic, high-fidelity visual representation of wounds to a trainee who is still able to see and operate in their real environment. To enhance medical training with more realistic hands-on experiences, we are working to develop the Combat Casualty Care Augmented Reality Intelligent Training System (C3ARESYS). C3ARESYS is our concept for an AR-based training system that aims to provide more realistic multi-sensory depictions of wounds that evolve over time and adapt to the trainee interventions. This paper describes our work to date in identifying requirements for such a training system, current state of the art and limitations in commercial augmented reality tools, and our technical approach in developing a portable training system for medical trainees.

Keywords: Augmented reality · Tactical combat casualty care
Medical training · Moulage

1 Problem and Motivation

Combat Life Savers, Combat Medics, Flight Medics, and Medical Corpsman are the first responders of the battlefield, and their training and skill maintenance is of preeminent importance to the military. While the instructors that train these groups are highly rated medics, most simulations of battlefield wounds are typically very simple

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 227–239, 2018.
https://doi.org/10.1007/978-3-319-91467-1_19

and static. These might range from simple moulage to show some characteristics of the wound (essentially rubber overlays with fake blood painted on) to a piece of tape inscribed with the type of wound, with no physical representation of the wound itself. In many field-training exercises, each soldier carries a "casualty card" that, if they are nominated to be a casualty, tells the soldier/actor how to portray a wound named on the card. The card also tells the trainee what wound to treat.

While casualty cards themselves are relatively simple to use, the simplicity of the presentation often requires the instructor to describe the wound or remind the trainee during an exercise about the qualities of the wound that are not portrayed, including how the wound is responding to treatment. To simulate arterial bleeding, an instructor may spray fake blood on the moulage. This effort by the instructors is there to compensate for the low-fidelity simulation, and takes away from time that could be spent providing instruction. While relatively simple, even these simulations take time and effort to create, set up, and manage, before and during the training exercise. The preparation before each exercise and the overall compressed training schedule of a training course means that trainees get limited hands-on practice in realistic settings.

Augmented Reality (AR), especially the recent boom in wearable AR headsets, has the potential to revolutionize how Tactical Combat Casualty Care (TC3) training happens today. Augmented Reality can provide a unique mix of immersive simulation with the real environment. In a field exercise, a trainee could approach a casualty role-player or mannequin and see a simulated wound projected on the casualty. The hands-on, tactile experience combined with the simulated, dynamic wounds and casualty response has the potential to drastically increase the realism of medical training. To enhance Army medical training with more realistic hands-on training, we are working to develop what we call the Combat Casualty Care Augmented Reality Intelligent Training System (C3ARESYS). This paper outlines our work to date in identifying how AR tools could fit into, and augment, current US Army medical training. We first briefly cover the types of training that occur in the standard 68 W (Army Medic) course, and the types of injuries on which they are trained. We also briefly describe the task analyses we conducted related to medical training. Together these serve as a basis for identifying elements of training including some requirements that an AR-based training system would need to meet. We then describe our C3ARESYS concept, our anticipated approach, and challenges to developing and evaluating the system. In this work, we have evaluated current AR technologies on the market relative to the requirements we identified. While there are significant limitations to current AR systems, our approach works within the current limitations of current AR technologies, while anticipating future advances that we could leverage.

2 Background: Augmented Reality

AR typically refers to technology that allows a user to see a real environment while digital information is overlaid on that view. Heads-Up Displays (HUDs) such as in cockpits or fighter pilot helmets represent early work in AR, though typically these

overlays do not register with objects in the environment. Later work includes registering information with the environment for tasks ranging from surgery, to machine maintenance, to entertainment such as the addition of AR scrimmage lines in NFL football games, or the highlighting the hockey puck in NHL games. See [1, 2] for thorough surveys of augmented reality. As mobile devices (phones, tablets) have become more capable, augmented reality has become more mobile, with game examples such as Pokemon Go™, which provides an "AR view" option to show 3D renderings of game characters overlaid on top of camera views. More recently, wearable AR hardware has tended to focus on see-through glasses, visors, or individual lenses that allow for computer-generated imagery to be projected hands-free, while allowing the user to see the surrounding environment directly. Additionally, more sophisticated AR projections are registered with the real environment, where digital objects can be placed on real tables or seem to interact with real obstacles. It is these latter wearable, spatially aware technologies we focus on.

While the technology continues to improve, there are several limitations with current AR systems that have real implications in training, including limited computer processing power and limited field of view. We will cover these limitations, and their impact on training, throughout this paper in the context of a medic training application.

3 Related Work

The main method of hands-on medic training is through simulation. This often focuses on hands-on physical simulants, such as moulage overlaid on a simulated human casualty, either a mannequin or a human playing the role. Some training facilities use instrumented mannequins that can bleed, exhibit a pulse, and even talk. However, these systems, including the computers that enable them, are expensive, not very portable for field training and are not at every training site. There are also physical part-task training simulators, such as tools to teach proper tourniquet application that require purpose-built hardware. Examples include a computerized portion of a fake leg with fake blood (e.g., TeamST's T3 Tourniquet Task Trainer [3]), or instances with metaphoric cues – lights that go out when the tourniquet is properly tightened (CHI Systems' HapMed Tourniquet Trainer [4]).

There are also examples of digital simulations for training medics. For example, ARA's virtual reality medical simulation ("HumanSim: Combat Medic" [5]) provides game-like ways to view wounds and apply treatments. Rather than the trainee physically performing a treatment, this environment focuses on the procedures. The trainee in uses the mouse or keyboard to select some treatment; the game visuals then show that treatment happening, along with the effect of treatment. Instead of naturalistic cues about the wound or the casualty (e.g., such as feeling a pulse by putting fingers on a wrist), the game provides metaphoric cues (such as displaying the pulse on the screen). With more portable and more capable technology, Augmented Reality is starting to be

used in medical training, including Case Western Reserve University using Microsoft's Hololens™ for anatomy training [6], and CAE's VimedixAR ultrasound training system [7].

4 Domain and Requirements Analysis

Wounds and Procedures. To help define the scope of the system, we surveyed current training recommendations, manuals, and other TC3-related publications, and also interviewed instructors to get a broad view of medic training. Findings from recent conflicts identify particular distribution and mechanisms of wounds [8, 9], which are summarized in Table 1 below. More specifically, the Army Medical Department (AMEDD) Approved Task List (2016) gives the assessments and treatments that a trainee must know to become a medic. The TC3 handbook [10] also provides details of the types of injuries seen in recent conflicts, along with treatment procedures.

Table 1. Injuries in recent conflicts (from [8])

Main **distribution** of wounds: • Extremities: 52% • Head and neck: 28% • Thorax: 10% • Abdomen: 10%	**Types** of injuries: • Penetrating head trauma (31%) • Surgically uncorrectable torso trauma (25%) • Potentially correctible surgical trauma (10%) • Exsanguination (9%) • Mutilating blast trauma (7%) • Tension pneumothorax (3–4%) • Airway obstruction/injury (2%) • Died of wounds - infection and shock (5%)
Injury **mechanisms:** • 75% blast (explosives) • 20% gunshot wounds	

Along with identifying injuries, we worked to identify and document treatment procedures for these injuries using task analysis methods. We focused on three main sources for our task analysis: published documents (e.g., field manuals and related publications [9, 10]), interviews with SMEs, and observations of medic training. We conducted interviews with subject matter experts on our team, with instructors at the Pennsylvania National Guard Medical Battalion Training Site (MBTS), and with a medic at Fort Bragg, and also observed training at MBTS. These interactions helped us understand the spectrum of tactical combat casualty care, including the types of training that occurs in Army medical training, and details on particular treatments.

Along with scoping, the goal of our analysis was to identify specific wounds and related procedures that medics train for, so we could identify how an AR system could contribute to training. We looked broadly at medic training, and then looked more narrowly at selective examples to assess the level of detail required for an AR system. The Army's Tactical Combat Casualty Care training manual [10] includes step-by-step instructions about procedures. There are also previously published task analyses of treatments such as cricothyroidotomy [11, 12] and hemorrhage control [11].

For our purposes, we needed to identify not just the treatment procedures that a medic would perform, but also what the medic would perceive about the casualty and the wound to be able to perform some procedure. For this reason, our analysis was in the style of Goal-Directed Task Analysis (GDTA) [13], which captures the hierarchical nature of goals and tasks, along with decisions that must be made to perform the tasks, and the situational awareness requirements needed to make those decisions. Figure 1 shows an example of GDTA applied to a medical task. The uppermost goal is to perform an airway/breathing/circulation assessment, and a sub-goal is to perform a breathing assessment. Rectangular boxes connected by lines are the medic's goals and sub-goals. The rounded nodes beneath the task nodes contain decisions that must be made in order to perform the tasks. The rectangle beneath the decision identifies the situation awareness requirements needed to make those decisions. Per Endsley's approach to situation awareness (SA) [14], the three levels include: Level 1: immediate perception; Level 2: relating those perceptions to goals; and Level 3: projecting the current state into some future state.

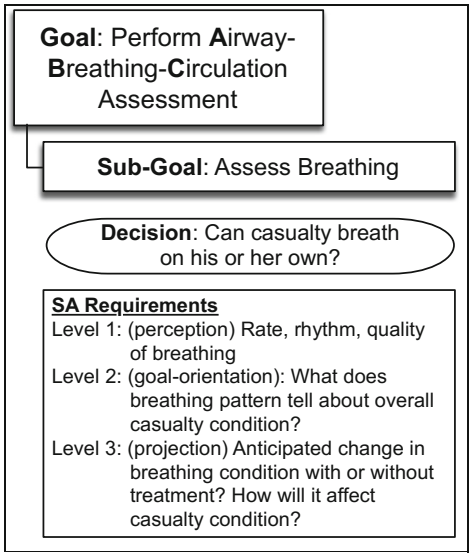

Fig. 1. Example Goal-Directed Task Analysis for assessing casualty breathing.

While many of these procedures are documented, not all of the documents or prior analyses included all of the elements that we needed for a GDTA. Thus, our effort included combining data from different sources to construct a more comprehensive task model with the level of detail needed to build a training system. For example, our task analysis for the process of controlling bleeding is a consolidation of the Cannon-Bowers, et al., task analysis of *Hemorrhage Control* [11] and the task *Apply a Hemostatic Dressing* task from the Soldier's Manual [10], supplemented with other related treatments from the Soldier's Manual and interviews with SMEs. The medical paper provided a rough outline of the task, along with some decisions to be made and SA requirements to perform the task; the Soldier's Manual provided a more detailed breakdown of the subtasks involved, but both needed additional detail for our design purposes.

This analysis has served a few purposes toward defining the requirements for a building an AR-based training system. First, the analysis captures the steps necessary to perform a treatment task, which can serve as the basis for an expert model to compare against trainee actions in an assessment process. Second, this same model can be used as the basis for automatically *recognizing* trainee actions, based on the atomic actions identified as the sub-tasks in the GDTA. Third, the Level 1 Situation Awareness Requirements define the cues that need to be present in a training environment to help

the trainee identify the injury and make decisions about treatment. (Levels 2 and 3 are products of the trainee's cognition but could also be used as part of assessing the trainee's skills or to provide additional feedback to the trainee.)

Types of Training. A good deal of training occurs in classrooms, but our focus was on hands-on, scenario-based medic training. Sometimes called "lane training," this type of training aims to cover different conditions and settings that medics will have to work in. At MBTS, the scenario-based training included dismounted patrols where the trainees had to care for wounded soldiers while under fire; indoor trauma aid stations where trainees had to triage, treat, and evacuate casualties; and mobile care where the trainees had to perform care while in casualty evacuation (CASEVAC) vehicles. In addition to the stress of treating casualties with life-threatening wounds, most of the scenarios included external stressors such as tight time schedules, extreme noise, or enemy fire to make the scenario more realistic to the trainee.

Role of Instructors. In addition to the wounds and procedures for treating them, a critical part of Army medic training today is the vital role of the instructors. Their presence, instruction, and participation during scenario-based training are especially important for a number of reasons. Because the baseline presentation of wounds is extremely simple and static (e.g., painted moulage or in some cases even less detail, such as a piece of tape with "amputation" written on it), the instructor must also provide to the trainee information about the wound and overall condition of the casualty – what it is, how it starts out, and how it changes over time. This may include giving verbal descriptions of the wound ("this is an amputation below the knee"), supplying vital signs that are not present in the casualty simulation, and describing the behavior of the casualty ("the patient is moaning in pain"). The instructor may also squirt fake blood on the wound to simulate arterial blood flow. Instructors are of course observing the trainee's treatments and other behavior as a way to assess trainee mastery of the tasks and performance under pressure. Instructors also inject dynamics into the training scenario, changing the difficulty in response to the trainee's behavior. They also provide instruction and direction during the scenario and lead after-action review sessions.

Technical Requirements. Based on the requirements given by the customer and our own analysis, we developed a list of stated and derived technical requirements that would help us define an AR-based training system to fit how medic training is currently done. These requirements cover a variety of categories such as *wound portrayal*, *hardware requirements*, *trainee interface*, and *instructor interface.* Table 2 below provides a subset of the roughly 40 high-level requirements we identified. These requirements guided our design of the system overall, which we cover in the next section.

Table 2. Requirements for outdoor lane training use (subset)

Req't #	Requirement description
Multi-modal augmented reality portrayal requirements	
AR1	System must overlay AR wounds on a casualty (human or mannequin) and those wounds must stay locked onto the correct position even with the trainee and/or the casualty moving
AR2	The system must portray the dynamics of wounds: blood flow, responses to treatment, etc.
Wearable hardware requirements	
HW1	The wearable system must fit with normal Soldier gear in outdoor lane training (i.e., when helmets are worn, with full rucks)
HW2	The wearable system must be ruggedized for outdoor lanes: the system must hold up to Soldier activities (running, diving, prone, etc.) and various weather conditions
Trainee interaction requirements	
TIR1	The system must recognize that the treatment is occurring with the right steps in the right order, with the right timing relative to the wound/casualty condition and to other treatments
TIR2	The system must recognize treatments that use instruments
Instructor interface requirements	
II1	Must enable instructor to get the same view of the casualty as the trainee, including any AR views
II2	Instructor must be able to get instructor-only views of the casualty; e.g., ground truth condition of the casualty
System and integration requirements	
SR1	The system must minimally be able to accommodate one casualty, with wounds, responses, etc.
SR2	The system must accommodate the use of part-task trainers (such as for intra-osseous infusion) when the procedure cannot be practiced on either mannequins or human volunteers

5 Technical Approach

The C3ARESYS concept focuses largely on the question of *training fidelity*. The centerpiece is the use of AR technology to enhance the visual aspects of training – portraying wounds in ways that not only look more accurate but also exhibit the dynamics of real wounds, including their progression over time and their responses to treatment. Because training is a multi-sensory experience, our approach leverages the moulage that is used today to provide the haptic sensations of wounds, while also exploring how it might be extended to provide richer training experiences. Figure 2 illustrates our C3ARESYS concept.

Given the complexity of potential models, the broad range of wounds, and the broad array of treatments performed by trainees, we chose to focus the design and development on the core AR modeling elements. This includes the visual display of wounds (and their dynamics), effective registration of the wound models on moving

casualties, as well as the tactile portrayal of wounds and other casualty information. Other future extensions could include automated treatment recognition and intelligent tutoring. In making this design choice, we must include an instructor in the loop to track the trainee's actions and provide feedback, but we aim to give the instructor tools to help him or her perform these tasks.

Fig. 2. Combat casualty care augmented reality intelligent training system (C3ARESYS) Concept (adapted from US Army photo).

5.1 System Design

C3ARESYS is composed of a number of technologies focused on enhancing the multi-sensory training experience. A high-level system view is given in Fig. 3. The main software component of C3ARESYS focuses on **Dynamic AR Modeling.** This component deals with producing a multi-modal rendering of a wound with appropriate cues relevant to the trainee. The **Casualty/Wound Tracker** determines where the wound (and related visual cues such as blood flowing from the wound) should be placed based on sensing the position of the casualty, moulage, and other cues. The **Multi-Modal Rendering Engine** renders visual and other wound effects such as the wound changing visually over time (e.g., based on treatments), audible and tactile cues associated with the wound (e.g., breathing sounds, pulse) based on parameters stored in the **Multi-Modal Wound Models** database. The **Physiology Modeling** module determines how the wound and the physiology of the casualty generally would evolve based on interventions by the trainee (or lack of intervention). We expect that the Physiology Modeling module will leverage current tools available, such as BioGears [15] or the Pulse physiology engine [16]. The input to the Physiology Modeling engine is a specification casualty's condition and of specific treatment (e.g., saline drip at

50 ml), which would then result in changes to physiological parameters of the casualty model (e.g., increased radial pulse). These inputs would come from an instructor who is observing the trainee's actions and entering the actions into an instructor interface (see below). The outputs of this engine (i.e., the collective set of parameters of the casualty model), combined with the Wound Models database, tell the rendering engine what to portray.

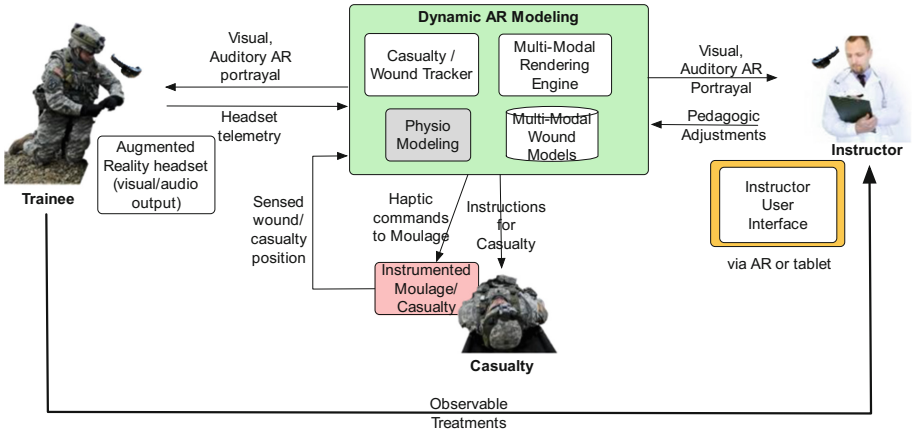

Fig. 3. High-level view of the C3ARESYS architecture.

The outputs of the Dynamic AR Modeling component will be rendered in a few ways: (a) visual and audio output through the AR systems worn by the trainee(s) and the instructor(s); (b) commands sent to the instrumented moulage to produce tactile cues; (c) instructions for the casualty. If it's a human volunteer, he or she might be told how to behave or what to say to portray the wound effects accurately (e.g., moaning in pain, being non-responsive, etc.). If the casualty is a mannequin, these instructions could go to a system that plays back audio recordings or generates speech from text. The system could also project AR overlays on instruments the trainees use, such as overlaying an animation on top of the blood pressure gauge to show the representative blood pressure of the casualty rather than whatever the blood pressure cuff would render from a live casualty or even a mannequin. Additionally, the Instructor's view through the AR glasses could include ground truth data that the trainee doesn't see, to help the instructor keep track of the condition of the casualties, for example.

The **Instrumented Moulage** component is standard moulage that we plan to augment in a few ways. The use of moulage by itself serves a few purposes. First, from an AR registration perspective, it provides the visual anchor to tell the AR system where to draw the wound. Without having some reference point, the AR visualization would float around independent of the position of the casualty. Second, it provides a

reference point to the trainee when using AR, both to tell the trainee where to look and also to give them a low-fidelity representation of the wound even when AR system is not tracking it. Third, it provides the tactile experience of the wound that AR by itself cannot provide. Typical user interactions with pure AR are, at this point, not rich enough to provide a haptic experience, and technologies like haptic gloves are still quite nascent in their development (not to mention that trainees typically wear surgical gloves during training). At least with today's training using typical moulage, the trainee gets some simulated version of how the wound feels.

In developing the Instrumented Moulage, we plan to explore the use of actuators (small motors) and sensors to provide an enhanced experience for trainees. We expect that the system could activate the moulage with specific patterns that simulate, for example, the casualty's pulse at the wrist or the feel of blood flowing. Sensors in the moulage could be used to identify treatments the trainee applies. The Instrumented Moulage system could be connected wirelessly (e.g., via Bluetooth) to the rest of the system.

Lastly, the **Instructor User Interface** provides a way for the instructor to participate in the training session. We envision that this interface could include an AR viewer to get views of the casualty, including the trainee perspective and an instructor-only, ground-truth perspective. This could be supplemented with a hand-held tablet-like device for making changes to the scenario, tracking trainee actions, or taking notes on trainee progress. Such a system would also help the instructor manage multiple training sessions simultaneously. These tools in concert could also be used to facilitate after-action reviews.

6 Challenges with Augmented Reality

There are several challenges with using augmented reality for practical applications, including medical training. We break down these challenges into four categories: field of view, visual tracking/processing power, form and fit, and user interaction.

Field of View (FOV). One of the most apparent when putting on wearable AR technology is the limited field of view. Most wearable technologies average around a 35° diagonal field of view. Besides taking away from an immersive experience, users often have to search around to find any AR objects placed in a scene, and large objects often get cut off by the FOV restriction. Some applications will guide the user with arrows or other indicators for where to look, but these can also distract from the user experience. Our use of moulage as a visual marker is in some ways an accommodation to this limitation. If the trainee looks away from the moulage, to outside of the core projection FOV, the digital wound model will disappear from the trainee's view. However, the moulage will remind the trainee where the wound is, and provides at least a lower-fidelity version of the wound.

Processing Power and Tracking. For AR applications where objects need to be registered with a location in space, those objects need to stay in place reliably while the user moves around. This is especially true in medic training, where the trainee is constantly moving around the casualty, and may even move the casualty around to

perform assessments and treatments. Reliable tracking is a function of the system sensing and processing the environment fast enough as the user moves relative to the target to keep the digital object locked in place. Vision-based tracking systems also require good lighting to be able to track the environment effectively.

In our first phase of work, we implemented some simple versions of marker-based tracking as a feasibility assessment of our design as well as a way to get hands-on experience with existing AR tools. Our initial testing used Microsoft HoloLens. Because there are several limitations to what the HoloLens provides to developers (in particular, no explicit object tracking), we had to add some extensions to be able to track these markers. We explored using different 3rd-party tools including OpenCV and Vuforia™ to recognize and track visual markers. Our first pass used OpenCV implemented on the HoloLens, using QR-style markers for tracking. The system was able to track the marker as the user moved around, while keeping the marker in view and while moving the casualty's arm side-to-side. However, movement induced noticeable lag when tracking the markers and trying to keep imagery in place. Figure 4 shows a version of the system using Vuforia running with the HoloLens. This was faster than OpenCV, but still had some lag issues. We have also done some hands-on testing with Osterhout Design Group's R7 glasses, with similar results with moving targets.

Fig. 4. Visual marker (top), and wound overlaid on the marker (bottom).

Form and Fit. The recent boom in AR wearables has opened many doors for how AR technologies might be used. However, the form that these systems take is often a bulky headset made of seemingly delicate components for the price. Many designers choose to put all the sensors, computing power, and power sources on board, which results in more weight carried on the user's head. For example, the current $3000 Microsoft HoloLens seems too fragile for military use and is too bulky to fit under a standard Kevlar helmet. Other designers go the route of having a separate connected device to provide battery and processing power (e.g., Meta2, Epson Moverio BT-300), thereby allowing the headset to be lighter.

Ruggedness is also a question. Medic trainees operate in many environmental conditions with a bunch of other gear. They might be treating casualty in heavy rain, or diving for cover to avoid (simulated) enemy fire. These uses risk damaging or breaking what is (so far) quite expensive equipment. Instructors may be unwilling or unable to spend money on such fragile equipment. This may limit many of the training use cases to those where the conditions are more suitable to the device. Some manufacturers are starting to address the issue of being tolerant to different environmental conditions (e.g., ODG R-7HL), but this is not a universal concern among hardware providers.

User Interaction. Perhaps more fundamental than the above is the lack of compelling interactions with AR objects. Processing power continues to increase year after year, as does battery size and efficiency, which will also contribute to more efficient, more compact devices. However, current user interaction tends to use traditional computing metaphors. The current state of the art for AR systems lies three main areas: speech interaction, head tracking to draw a cursor where the user is looking, and limited gesture recognition to capture simple interactions such as pinching or grasping objects. Speech recognition can be useful in the right conditions, but its utility is limited in our C3ARESYS application. Using one's head as a pointing device can become tiring, especially if the objects to interact with are small and require precision. Gesture recognition is often in the form of making selections or dragging objects around (Microsoft's "air tap") or giving commands (Augmenta's iconic hand shapes). Hand tracking and gesture recognition could be compelling and useful if related to objects themselves such as grasping and manipulating them naturally, but recognition of these inputs needs to be highly accurate, otherwise the user is left frustrated at the poor interaction. Some systems use hand-held controllers to manage user input, but these add additional gear that the user has to hold to operate, which takes away from the hands-free nature of wearable AR and does not fit with this medic training domain. None of these typical types of interactions are especially compelling to medic training; instead, we need ways for the trainee to interact directly with the AR wounds, which could include domain-specific interactions such as filling a cavity with gauze or putting pressure on the wound to stop bleeding. We will continue to explore interaction features such as hand tracking as the technology continues to improve.

Feedback to the user is also another area in which AR technology is lacking. Visual and audio feedback is typically the norm, as expected. However, as mentioned earlier, this hands-on medic domain relies on tactile sensations and haptic feedback to be realistic. Medics will feel for a pulse and will palpate a wound to assess its condition. Haptic gloves could be a solution, but current technology is fairly rudimentary, and they require their own power sources and computing. As mentioned, this is another reason we have chosen to stay with moulage: to provide the tactile sensation that AR currently lacks.

7 Summary

We have described the motivation, requirements, and design of a system we call the Combat Casualty Care Augmented Reality Intelligent Training System (C3ARESYS). The motto "Train as you fight" that is ubiquitous in the military is a main driver – working to improve the fidelity of hands-on medic training. Whereas today's trainees at best experience static moulage as a representation of a wound (and very often they are presented with much less than this), AR has the potential to provide a more representative multi-modal training experience. However, as we describe above, there are many limitations in current AR technology that have forced our hand in designing a system for near-term use. Field of view, processing power, fit, form, lack of ruggedness, and limited user interaction all have very tangible effects on our system design and how readily such AR technology can be used in the field. We have tried to make

design decisions that will enable us to build a prototype system today, while also being able to take advantage of AR technology as it improves in what is currently a very dynamic marketplace.

We have presented only a design here. Our next step in this work is to develop a working prototype that can be used for a limited set of treatment procedures. This will include the trainee's experience and tools for the instructor so that we can mirror the current instructor-in-the-loop training paradigm. Once we have developed a prototype system, we aim to conduct hands-on evaluations with medic instructors and trainees to get their feedback.

References

1. Azuma, R.: A survey of augmented reality. Presence: Teleoperators Virtual Environ. **6**(4), 355–385 (1997)
2. Azuma, R., et al.: Recent advances in augmented reality. IEEE Comput. Graph. Appl. **21**, 34–47 (2001)
3. TeamST: T3 Tourniquet Task Trainer (2018). http://www.teamst.com.au/torniquet-task-trainer.html
4. Systems C: HapMed Tourniquet Trainer (2017). https://www.hapmedtraining.com/
5. ARA: Combat Medic 3D Virtual Trainer (2015). https://www.ara.com/projects/combat-medic-3d-virtual-trainer
6. Hein, I.: Cadaverless Anatomy Class: Mixed Reality Medical School. Medscape (2017)
7. CAE: Vimedix (2017). https://caehealthcare.com/ultrasound-simulation/vimedix
8. Helwick, C.: US Army Reveals Trends in Combat Injuries. Medscape 2011, October 2016
9. USArmy Tactical Combat Casualty Care: Observations, Insights, and Lessons, 12-10 (2011)
10. USArmy, Soldiers Manual and Trainer's Guide: MOS 68 W Health Care Specialist, Department of the Army (2013)
11. Cannon-Bowers, J., et al.: Using cognitive task analysis to develop simulation-based training for medical tasks. Mil. Med. **178**(10:15), 15–21 (2013)
12. Demirel, D., et al.: A hierarchical task analysis of cricothyroidodomy procedure for a virtual airway skills trainer simulator. Am. J. Surg. **212**, 475–484 (2016)
13. Endsley, M.R., Jones, D.G.: Designing for Situation Awareness: An Approach to User-Centered Design, 2nd edn. Taylor and Francis, New York (2011)
14. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Hum. Factors **37**(1), 32–64 (1995)
15. ARA. BioGears (2018). https://www.biogearsengine.com/
16. Kitware. Pulse Physiology Engine (2018). https://physiology.kitware.com/

A Workload Comparison During Anatomical Training with a Physical or Virtual Model

Andrew Wismer[1(✉)], Lauren Reinerman-Jones[1], Grace Teo[1], Sasha Willis[1], Kelsey McCracken[1], and Matthew Hackett[2]

[1] Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA
awismer@ist.ucf.edu

[2] Simulation and Training Technology Center, Army Research Lab, Orlando, FL, USA

Abstract. Recent research argues for the supplementation of traditional anatomical training with emerging three-dimensional visualization technologies (3DVTs); however, little is known regarding the effect these technologies have on learner workload. In this experiment, sixty-one participants studied gross brain anatomy using either a plastic physical model (PM; $n = 29$) or models presented in virtual reality (VR; $n = 32$). Participants were fitted with a functional near-infrared spectroscopy (fNIRS) sensor, worn on the prefrontal cortex. fNIRS measures regional saturation of oxygen (RSO_2) and is indicative of workload. Participants then completed a pre-knowledge test on human brain anatomy. Participants were given 10 min to use the provided 3DVT to study 16 anatomical brain structures. Following the study period, participants completed additional surveys measuring workload, newly acquired anatomical knowledge, and cognitive resources used. Overall, anatomical knowledge increased at post-test and the change was no different between PM and VR conditions. Participants in the PM condition reported significantly higher levels of spatial workload, mental demand, and frustration. RSO_2 values suggest left hemispheric increases from baseline during learning for the VR condition, but decreases for the PM condition. No other measures revealed differences between the two conditions. These results provide support for the supplementation of traditional anatomical training techniques with virtual reality technology as a way of alleviating workload. Further research is needed to explain the link between workload and performance in anatomical knowledge acquisition.

Keywords: Physiological response · Workload
Methods and metrics for testing and evaluating augmented cognition system
Virtual reality · Anatomical training · fNIRS · Physical model
Visualization technologies

1 Introduction

When attempting to learn spatial information from two-dimensional displays, such as when students study gross anatomy using textbook images, a high level of workload is placed on the learner [1]. This workload has been associated with decreases in

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 240–252, 2018.
https://doi.org/10.1007/978-3-319-91467-1_20

knowledge acquisition [2]. Emerging three-dimensional visualization technologies (3DVTs) support a learner's understanding of spatial depth information by providing realistic representations of three-dimensional objects [3]. 3DVTs include physical models, virtual or augmented reality, and holographic displays. Indeed, numerous studies support the utilization of 3DVTs over traditional 2D displays [4–8].

Anatomical science is a domain that requires effective display of spatial information. Digital images are present not only in anatomical training and instruction, but also in medical diagnosis, pre-operative planning, and minimally invasive surgery [9]. Cadavers are commonly held as the gold standard for anatomical training as they enable hands-on experience with actual human tissue [10, 11]. However, cadavers can be costly to maintain, challenging to store, and require extra work for instructors [10–12]. For these reasons, 3DVTs are considered valuable supplements to traditional anatomical training.

A traditional supplement used in anatomical training is a physical model [12, 13]. Physical models replicate an anatomical system/structure using any variety of materials (e.g., plastic, fiberglass, clay). Physical models afford a learner hands-on experience with anatomical structures through rotation and often disassembly to aid in spatial comprehension. In general, physical models are easy to obtain, highly portable, and provide a useful tool to increase a person's base knowledge of anatomy at a low cost [14].

In addition, recent technological advances have increased the use of virtual reality in anatomical training. Virtual reality—the computer-generated simulation of three-dimensional objects/environments—provides capabilities similar to that of physical models for rotation, manipulation, and enhanced spatial understanding [15]. Much of the work to date on the use of virtual reality for knowledge acquisition in anatomical training has involved computer-based applications and modules (i.e., "desktop VR"). Research comparing these computer-based models to physical models has found benefits from using physical models for training in anatomical identification [16, 17]. While some work has been done with respect to more immersive virtual reality technologies (e.g., with head-mounted displays) for procedural training, much less has been done regarding the use of immersive VR for anatomical knowledge acquisition [18–20].

As physical and immersive virtual reality (VR) models share many features (e.g., both present 3D information, allow interaction and study of multiple views), an important criteria for evaluating their use for educational purposes is the level of workload they impose on the learner. Cognitive load theory [21] suggests increased workload is only detrimental if and when it exceeds a learner's working memory capacity. Currently, differences in detrimental workload imposed on a learner by 3DVTs such as physical or virtual reality models are unknown. Some 3DVT types (e.g., monoscopic 3D displays, digital holograms) have shown lower workload compared to 2D displays [22, 23], but this may not be true for physical and VR models.

The present work was designed to address the question: What are the differences in workload between physical and virtual reality models used for supporting knowledge acquisition in gross brain anatomy? This research question was addressed through both physiological and subjective measures of workload, with knowledge gain assessed through pre- and post- brain anatomy tests. The present experiment showed that

workload differences during anatomical knowledge acquisition may stem from limitations in the typical use of physical models compared to models presented in virtual reality.

2 Methods

2.1 Participants

Sixty-one students from the University of Central Florida (29 Males, 32 Females), between the ages of 18 and 28 (*Mdn* = 18, *IQR* = 1), completed the experiment for course credit. All participants provided written informed consent prior to participation and were at least 18 years old with normal, or corrected-to-normal, vision.

2.2 Experimental Design

Participants were assigned to the physical model (PM) or virtual reality (VR) learning condition.

2.3 Materials

Physical Model. The physical model was presented along with a label sheet defining the numbered structures on the model (see Fig. 1). The numbered labels were added by the researchers and color-coded to best match the colored regions of the virtual model in the VR condition. The physical model (6 × 5.5 × 5.5 inches) contained eight pieces and weighed 2.5 lb. The model could be examined as a whole or in any combination of its eight pieces.

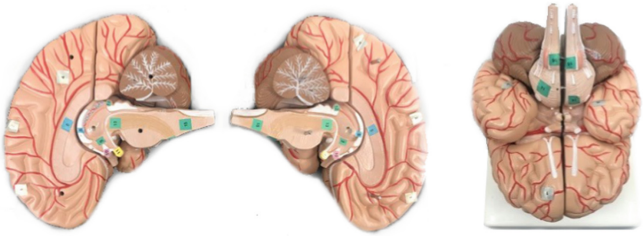

Fig. 1. Plastic physical brain model with eight removable pieces (Axis Scientific) and numbered structures.

Virtual Models. The HTC Vive virtual reality (VR) system includes a head-mounted display, two controllers (one for each hand), and two "light house" sensors that track the headset and controllers, and project them into the virtual environment. The Vive connects to a desktop computer and displays a virtual environment through SteamVR.

Within the virtual environment, two brain models were displayed on a table (see Fig. 2). One brain model showed the external view of the brain, with label sets corresponding to a ventral and lateral view, while the second model displayed labeled structures from the medial view of the brain. Label sets could be toggled on or off by the participant using the controllers. The VR system allowed the participant to fully rotate the brain models to study the brain structures and spatial relationships from different viewpoints.

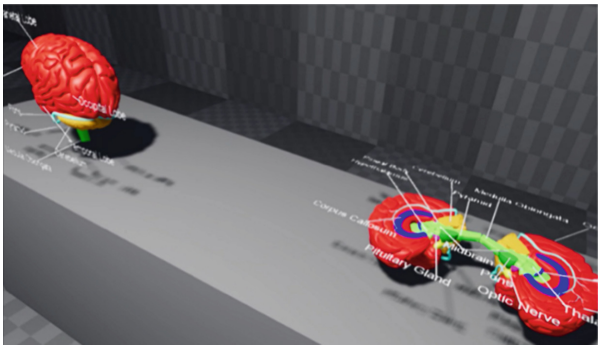

Fig. 2. Virtual brain models used in the VR condition.

2.4 Measures

Workload Measures

NASA-TLX. The NASA-Task Load Index (NASA-TLX) [24] was administered following the learning task to assess workload. Six items addressing workload (i.e., mental workload, physical workload, temporal workload, effort, frustration, performance) were presented on 100 point scales in 5-point increments. High scores on each scale indicate high workload, with the exception of the performance subscale in which high scores relate to "poor" perceived performance.

MRQ. The Multiple Resources Questionnaire (MRQ), developed by Boles and Adair [25], was also administered after the learning task to assess workload in terms of cognitive resources used while studying with the 3DVT. Eleven of the original 17 items were utilized for purposes of this experiment (see Table 1). Responses were measured on a 100-point scale, with 0 indicating no usage and 100 indicating extreme usage.

Performance Measures

Spatial Anatomy Test. A pre- and post-task Spatial Anatomy Test (SAT) was administered to evaluate knowledge gain. Accuracy and completion time were measured on identification (16 questions; one for each labeled brain structure), spatial knowledge (15 multiple-choice questions), and mental rotation questions (4 questions). Identification questions required the participant to select the correct label for a designated structure. Participants chose from a list of 32 brain structures (16 targets, 16 distractors). Four-alternative multiple-choice questions measured an understanding of spatial

Table 1. MRQ scores by 3DVT condition

	PM	VR	Test statistic	*p-value*
	M (*SD*)	*M* (*SD*)		
Manual	51.17 (27.08)	58.56 (22.98)	$t(59) = 1.15$	.254
Short-term memory	58.55 (32.45)	55.94 (26.36)	$t(59) = -0.35$	.730
Spatial attentive	82.31 (17.54)	77.94 (17.71)	$t(59) = -0.97$	.337
Spatial concentrative	74.66 (18.29)	58.47 (23.95)	$t(59) = -2.94$	.005*
Spatial emergent	66.45 (25.22)	48.88 (25.08)	$t(59) = -2.73$	.008*
Spatial quantitative	32.48 (29.62)	33.06 (25.32)	$t(59) = 0.08$	.935
Visual lexical	59.86 (26.23)	71.19 (25.05)	$t(59) = 1.72$	.090
Visual phonetic	44.00 (34.50)	36.50 (32.04)	$t(59) = -0.88$	.382
	Mdn (*IQR*)	*Mdn* (*IQR*)		
Spatial categorical	77.00 (36)	69.00 (32)	$U = 553.00$	.198
Spatial positional	90.00 (36)	69.00 (22)	$U = 628.50$	.017*
Tactile figural	60.00 (53)	30.50 (76)	$U = 574.00$	.112

Note. Asterisk (*) represent statistically significant group differences.

relationships between brain structures. In the mental rotation section, a target image of the brain model was provided along with four rotated images (two of which were mirror-images). The participants selected which two of the four new images were simple rotations of the target image. Brain images used for identification and mental rotation test questions were matched to the respective physical or virtual reality model condition.

Regional Saturation of Oxygen (rSO_2)

Changes in regional saturation of oxygen (rSO_2) in the left and right prefrontal cortex were measured using the Somanetics INVOS Cerebral/Somatic Oximeter through near-infrared light [26]. This non-invasive, indirect neuroimaging measurement, referred to as functional near-infrared spectroscopy (fNIRS), sheds light into cognitive functions such as workload [27].

2.5 Procedure

All participants provided written informed consent prior to participation. Each participant then completed a demographics survey and restrictions checklist, along with an Ishahara Color Blindness Test. These items did not serve as exclusion criteria; rather, they served to provide background information for use in later analyses. Next, the researcher fitted the participant with the fNIRS sensors. A five minute resting baseline was conducted as a reference for any changes in oxygenation during the experiment. The participant then completed the pre-task Spatial Anatomy Test (pre-SAT) to assess his or her prior knowledge concerning spatial brain anatomy. The participant was then given their assigned 3DVT (either the physical model or virtual reality system) and had ten minutes to use the technology to study the 16 labeled brain structures. The participant had the option to end the ten-minute study period early if they felt confident. Once the study time was complete, the participant completed a series of post-task

surveys on the computer, including the NASA-TLX, post-task SAT (identical to pre-task SAT but randomized order), and MRQ. Upon completion of the experiment, the fNIRS sensors were removed, and the participant was thanked, granted credit, and dismissed. The experiment took no longer than three hours to complete.

3 Results

3.1 Workload

NASA-TLX. Independent-samples t-tests were conducted for four of the six subscales of the NASA-TLX (i.e., Mental Demand, Temporal Demand, Effort, and Frustration) to examine the effect of each 3DVT on workload. Nonparametric Mann Whitney *U* tests were conducted for the two other subscales (i.e., Physical Demand and Performance), determined to come from non-normal distributions. Average scores on each subscale can be seen in Fig. 3.

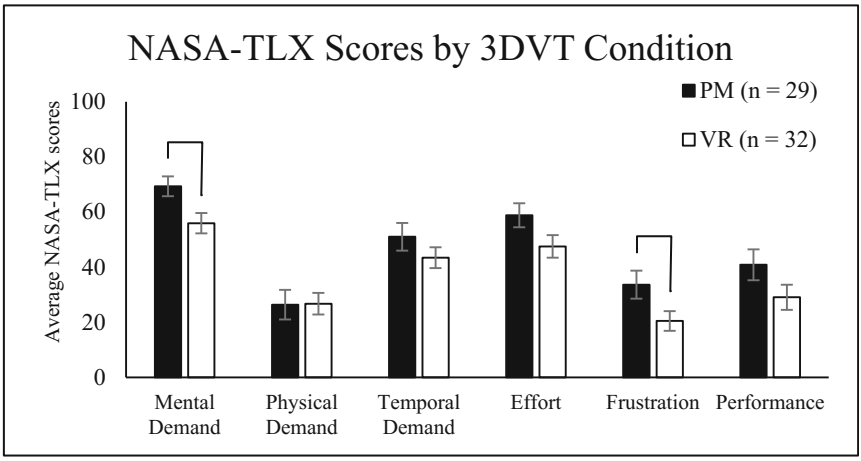

Fig. 3. NASA-TLX scores by 3DVT condition. Error bars represent standard error.

There was a significant difference between 3DVTs for Mental Demand scores, $t(59) = -2.56$, $p = .012$, with higher Mental Demand in the PM ($M = 69.31$, $SD = 19.31$) than VR condition ($M = 55.94$, $SD = 20.73$). There was also a significant difference between PM and VR conditions for Frustration, with higher Frustration in the PM ($M = 33.62$, $SD = 27.35$) than VR condition ($M = 20.47$, $SD = 20.26$). There was no significant difference between the two 3DVTs with respect to Physical Demand, $U = 415.50$, $Z = -0.706$, $p = .480$, $r = -0.09$, Temporal Demand, $t(59) = -1.22$, $p = .226$, Effort, $t(59) = -1.89$, $p = .064$, or Performance scores, $U = 569.00$, $Z = 1.53$, $p = .127$, $r = .20$.

MRQ. Independent-samples t-tests were conducted for 8 of the 11 included MRQ subscales (Manual, Short-Term Memory, Spatial Attentive, Spatial Concentrative, Spatial Emergent, Spatial Quantitative, Visual Lexical, and Visual Phonetic processes) to examine the effect of 3DVT on workload.

Nonparametric Mann Whitney *U* tests were conducted for the remaining three subscales (i.e., Spatial Categorical, Spatial Positional, and Tactile processes), which were determined to violate normality assumptions. See Table 1 for each of the 11 MRQ scale scores by 3DVT condition.

There was a significant difference between 3DVTs on Spatial Concentrative process scores, $t(59) = -2.94$, $p = .005$, with higher scores in the PM ($M = 74.66$, $SD = 18.29$) than VR condition ($M = 58.47$, $SD = 23.95$). There was a significant difference between 3DVTs on Spatial Emergent process scores, $t(59) = -2.73$, $p = .008$, with higher scores in the PM ($M = 66.45$, $SD = 25.22$) than VR condition ($M = 48.88$, $SD = 25.08$). There was a significant difference between 3DVTs on Spatial Positional process scores, $U = 628.50$, $Z = 2.38$, $p = .017$, $r = .31$, such that the PM condition ($Mdn = 90.00$, $IQR = 36$) had significantly higher scores than the VR condition ($Mdn = 69.00$, $IQR = 22$). No other significant differences were found between 3DVTs for the remaining subscales (all p's > .089).

3.2 Performance

Independent samples t-tests were conducted on pre-task SAT accuracy scores to investigate any differences in prior knowledge between groups. There was no significant difference between PM and VR conditions for overall, identification, or multiple choice SAT accuracy (all p's > .059). However, there was a significant difference between PM and VR conditions for mental rotation accuracy scores ($p = .039$).

2 (3DVT: PM, VR) × 2 (Testing time: pre, post) mixed factor ANOVAs were conducted on average SAT accuracy scores and completion times for the overall test and for identification, multiple choice, and mental rotation questions to examine differences in the level of spatial knowledge acquired between PM and VR conditions.

Level of Spatial Knowledge Acquired. There was a significant main effect of testing time on each of the following: average overall accuracy, $F(1, 59) = 629.23$, $p < .001$, $\eta_p^2 = .91$, average identification accuracy, $F(1, 59) = 445.67$, $p < .001$, $\eta_p^2 = .88$, average multiple choice accuracy, $F(1, 59) = 396.57$, $p < .001$, $\eta_p^2 = .87$, and average mental rotation accuracy, $F(1, 59) = 6.49$, $p = .013$, $\eta_p^2 = .10$. For each measure, post-task SAT scores were significantly higher than pre-task SAT scores (see Fig. 4).

There was no main effect of 3DVT condition on average overall, $F(1, 59) = 0.01$, $p = .942$, $\eta_p^2 < .01$, identification, $F(1, 59) = 2.23$, $p = .141$, $\eta_p^2 = .04$, or multiple choice accuracy scores, $F(1, 59) = 0.47$, $p = .494$, $\eta_p^2 = .01$. There was a significant main effect of 3DVT condition on average mental rotation accuracy, $F(1, 59) = 5.73$, $p = .020$, $\eta_p^2 = .09$, with higher accuracy in the PM ($M = 34.05$, $SD = 22.39$) than VR ($M = 20.31$, $SD = 22.39$) condition.

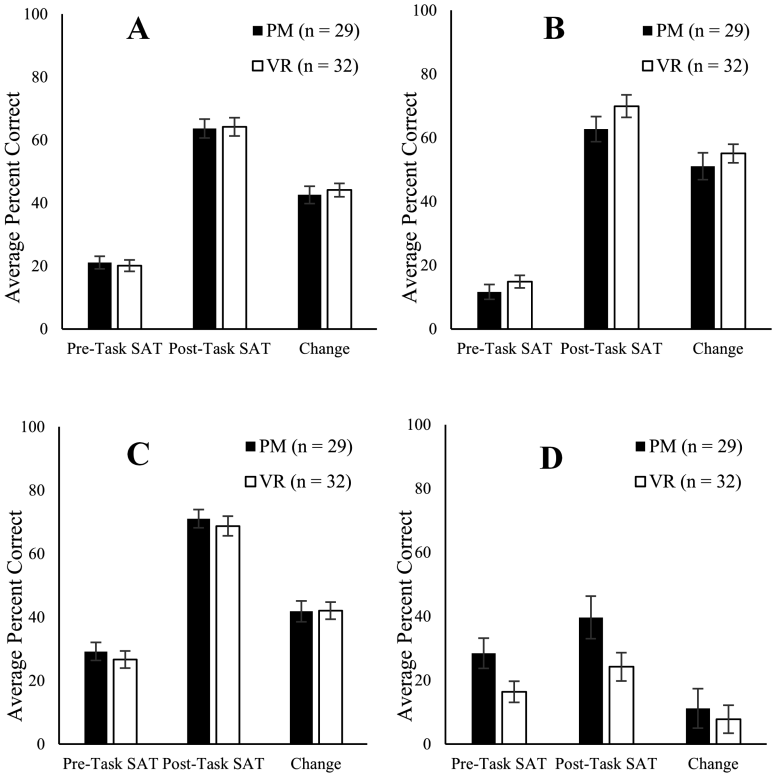

Fig. 4. Average overall (A), identification (B), multiple choice (C), and mental rotation (D) accuracy on SAT by 3DVT condition. Error bars represent standard error.

There was no significant interaction between testing time and 3DVT on any accuracy measure (all p's > .430).

Completion Times. There was a significant main effect of testing time on each of the following: average overall SAT completion time, $F(1, 59) = 12.06$, $p = .001$, $\eta_p^2 = .17$, multiple choice completion time, $F(1, 59) = 66.93$, $p < .001$, $\eta_p^2 = .53$, and mental rotation completion time, $F(1, 59) = 15.51$, $p < .001$, $\eta_p^2 = .21$. For overall and multiple choice questions, participants took significantly longer to complete the post-task SAT than the pre-task SAT. Conversely, participants took significantly longer to complete the pre-task SAT than the post-task SAT for mental rotation questions. There was no main effect of testing time on average identification completion time, $F(1, 59) = 0.22$, $p = .642$, $\eta_p^2 < .01$ (see Table 2).

There was no main effect of 3DVT condition on average overall, $F(1, 59) = 0.04$, $p = .836$, $\eta_p^2 < .01$, identification, $F(1, 59) = 1.31$, $p = .257$, $\eta_p^2 = .02$, multiple choice, $F(1, 59) = 0.02$, $p = .887$, $\eta_p^2 < .01$, or mental rotation completion times, $F(1, 59) = 0.87$, $p = .355$, $\eta_p^2 = .02$.

Table 2. Completion times on SAT by 3DVT condition and overall

Question type	Condition	Pre-task SAT	Post-task SAT	Change
		M (*SD*)	*M* (*SD*)	*M* (*SD*)
Overall	PM	400.32 (151.73)	488.59 (141.33)	88.27 (125.11)
	VR	428.54 (121.99)	447.95 (108.27)	19.41 (117.06)
Identification	PM	167.51 (82.51)	193.47 (59.53)	25.96 (89.40)
	VR	173.86 (65.43)	157.97 (49.66)	−15.90 (78.73)
Multiple choice	PM	134.82 (67.05)	215.22 (95.72)	80.39 (67.69)
	VR	155.90 (58.02)	198.90 (62.02)	43.00 (49.47)
Mental rotation	PM	97.99 (29.93)	79.90 (29.45)	−18.09 (27.67)
	VR	98.77 (23.12)	91.08 (29.64)	−7.69 (23.43)

Note. Completion Times are an average summation of all the questions in a category type (e.g., all 15 multiple choice questions).

There was an interaction between testing time and 3DVT for overall, $F(1, 59) = 4.93$, $p = .030$, $\eta_p^2 = .08$, and multiple choice completion times, $F(1, 59) = 6.15$, $p = .016$, $\eta_p^2 = .09$. Overall test and multiple choice questions showed significantly longer completion times for post- than pre-test. Average completion times were longer for the VR condition compared to the PM condition at pre-test with the reverse trend at post-test. There was no significant interaction between testing time and 3DVT for identification, $F(1, 59) = 3.78$, $p = .057$, $\eta_p^2 = .06$, or mental rotation completion times, $F(1, 59) = 2.52$, $p = .118$, $\eta_p^2 = .04$.

3.3 Regional Saturation of Oxygen (RSO_2)

A 2 (3DVT: PM, VR) × 2 (Testing Time: pre, post) × 2 (Hemisphere: left, right) mixed factor ANOVA was conducted on average rSO_2 values during pre- and post-task SAT. There were no significant main effects or interactions among the included variables (all p's > .225). A 2 (3DVT: PM, VR) × 2 (Hemisphere: left, right) mixed factor ANOVA was conducted on average rSO_2 values during the learning task. There was no main effect of hemisphere on average rSO_2 values during the learning task, $F(1, 57) = 1.88$, $p = .175$, $\eta_p^2 = .03$. There was a main effect of 3DVT condition, $F(1, 57) = 9.17$, $p = .004$, $\eta_p^2 = .14$, with higher change from baseline in VR ($M = 1.78$, $SD = 3.51$) than PM ($M = -0.65$, $SD = 2.99$). Critically, there was a significant interaction of hemisphere and 3DVT condition, $F(1, 57) = 6.90$, $p = .011$, $\eta_p^2 = .11$. Bonferroni pairwise comparisons revealed a significant hemispheric difference in the PM condition ($p = .008$), but not in the VR condition ($p = .367$). Pairwise comparisons revealed a significant difference in rSO_2 change from baseline in the left hemisphere between 3DVT conditions ($p = .001$), while this difference between 3DVTs did not reach statistical significance for the right hemisphere ($p = .053$). See Table 3 for average rSO_2 change from baseline values by 3DVT condition and hemisphere.

Table 3. Average RSO_2 change from baseline values by 3DVT condition and hemisphere

		PM (n = 27) M (SD)	VR (n = 32) M (SD)
Pre-task	Left	0.63 (2.33)	−0.05 (2.73)
	Right	0.27 (2.93)	0.31 (3.88)
Learning	Left	−1.23 (2.87)	1.97 (3.72)
	Right	−0.06 (3.05)	1.60 (3.35)
Post-task	Left	−0.25 (2.64)	0.24 (4.59)
	Right	0.78 (2.38)	0.04 (2.24)

4 Discussion

The present experiment provides a workload comparison between a physical model (PM) and models presented in virtual reality (VR) for supporting knowledge acquisition in anatomical training. While participants in both three-dimensional visualization technology (3DVT) conditions showed similar levels of knowledge gain, the VR condition decreased test completion time (pre- to post-test) to a greater extent than the PM condition. The PM condition was found to impose a higher degree of workload in terms of mental demand, frustration, and spatial processes, while average RSO_2 values suggested higher workload in the VR condition.

These results differ from previous studies comparing physical models to computer-based virtual models which showed benefits for physical models in anatomical knowledge acquisition [16, 17]. Thus, the more immersive virtual reality condition explored here may provide a closer match to the use of tangible, physical models. Benefits of physical models (e.g., tangible, portable) must be weighed against benefit of models presented in virtual reality (e.g., immersive, readily accessible library of models to access online).

The workload differences found in the present experiment may stem from different presentation formats between the physical and virtual reality models. The virtual reality model had structure labels fixed to the models (see Fig. 3) that could be toggled on and off. The physical model differed in that the model was labeled by numbers 1-16 with structure identifiers listed on a sheet of paper next to the model. Thus, studying with the physical model required an extra mental step to connect numbered labels to structure identifiers. It is possible that this difference is responsible for the higher workload seen in the physical model condition. Future work could better match the information presentation formats to provide a more direct comparison between 3DVT types.

Previous studies have shown workload affects a person's ability to learn from a 3D anatomical model [28–30]. The workload differences here were not associated with detriments in knowledge gain, but rather in test completion time. This suggests that the higher workload with the physical model relative to virtual reality models may not have been high enough to hinder overall knowledge gain. In other words, when the workload associated with a knowledge acquisition task is moderate at most, workload differences among 3DVT types may manifest in completion or response times rather than

knowledge gain. When task load increases, the selection of 3DVT may become more important for knowledge acquisition. This question is left open to future work.

In sum, educators and trainers should be aware of the capabilities and limitations of 3DVTs to ensure they do not impose a level of workload that hinders knowledge acquisition. The present experiment suggests that the selection of 3DVT for supporting anatomical knowledge acquisition may be made on factors such as cost, accessibility, and interest since minor differences in workload did not hinder learning. Still, further research is needed to better understand the link between workload and performance in spatial knowledge acquisition.

Acknowledgements. This research was sponsored by the U.S. Army Research Laboratory (ARL) and was accomplished under Cooperative Agreement Number W911NF-15-2-0011. The views and conclusions contained in this document are those of the author's and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

References

1. Said, C., Shamsudin, K., Mailok, R., Johan, R., Hanaif, H.: The development and evaluation of a 3D visualization tool in anatomy education. EDUCATUM J. Sci. Math. Technol. **2**, 48–56 (2018)
2. Khalil, M., Paas, F., Johnson, T., Payer, A.: Interactive and dynamic visualizations in teaching and learning of anatomy: a cognitive load perspective. Anat. Rec. Part B: New Anatomist **286B**, 8–14 (2005). https://doi.org/10.1002/ar.b.20077
3. Young, J., Sewell, J.: Applying cognitive load theory to medical education: construct and measurement challenges. Perspect. Med. Educ. **4**, 107–109 (2015). https://doi.org/10.1007/s40037-015-0193-9
4. Müller-Stich, B., Löb, N., Wald, D., Bruckner, T., Meinzer, H., Kadmon, M., Büchler, M., Fischer, L.: Regular three-dimensional presentations improve in the identification of surgical liver anatomy – a randomized study. BMC Med. Educ. **13** (2013). https://doi.org/10.1186/1472-6920-13-131
5. Ruisoto, P., Juanes, J., Contador, I., Mayoral, P., Prats-Galino, A.: Experimental evidence for improved neuroimaging interpretation using three-dimensional graphic models. Anat. Sci. Educ. **5**, 132–137 (2012). https://doi.org/10.1002/ase.1275
6. Petersson, H., Sinkvist, D., Wang, C., Smedby, Ö.: Web-based interactive 3D visualization as a tool for improved anatomy learning. Anat. Sci. Educ. **2**, 61–68 (2009). https://doi.org/10.1002/ase.76
7. Hilbelink, A.: A measure of the effectiveness of incorporating 3D human anatomy into an online undergraduate laboratory. Br. J. Educ. Technol. **40**, 664–672 (2009). https://doi.org/10.1111/j.1467-8535.2008.00886.x
8. Hackett, M., Proctor, M.: Three-dimensional display technologies for anatomical education: a literature review. J. Sci. Educ. Technol. **25**, 641–654 (2016). https://doi.org/10.1007/s10956-016-9619-3
9. Escobar, M., Junke, B., Holub, J., Hisley, K., Eliot, D., Winer, E.: Evaluation of monoscopic and stereoscopic displays for visual–spatial tasks in medical contexts. Comput. Biol. Med. **61**, 138–143 (2015). https://doi.org/10.1016/j.compbiomed.2015.03.026

10. Ghosh, S.: Cadaveric dissection as an educational tool for anatomical sciences in the 21st century. Anat. Sci. Educ. **10**, 286–299 (2016). https://doi.org/10.1002/ase.1649
11. Habbal, O.: The state of human anatomy teaching in the medical schools of gulf cooperation council countries: present and future perspectives. Sultan Qaboos Univ. Med. J. **9**, 24–31 (2018)
12. Baskaran, V., Štrkalj, G., Štrkalj, M., Di Ieva, A.: Current applications and future perspectives of the use of 3D printing in anatomical training and neurosurgery. Front. Neuroanat. **10** (2016). https://doi.org/10.3389/fnana.2016.00069
13. McMenamin, P., Quayle, M., McHenry, C., Adams, J.: The production of anatomical teaching resources using three-dimensional (3D) printing technology. Anat. Sci. Educ. **7**, 479–486 (2014). https://doi.org/10.1002/ase.1475
14. Yammine, K., Violato, C.: The effectiveness of physical models in teaching anatomy: a meta-analysis of comparative studies. Adv. Health Sci. Educ. **21**, 883–895 (2015). https://doi.org/10.1007/s10459-015-9644-7
15. Chavan, S.: Augmented reality vs. virtual reality: what are the differences and similarities. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **5**, 1947–1952 (2018)
16. Khot, Z., Quinlan, K., Norman, G., Wainman, B.: The relative effectiveness of computer-based and traditional resources for education in anatomy. Anat. Sci. Educ. **6**, 211–215 (2013). https://doi.org/10.1002/ase.1355
17. Preece, D., Williams, S., Lam, R., Weller, R.: "Let's Get Physical": advantages of a physical model over 3D computer models and textbooks in learning imaging anatomy. Anat. Sci. Educ. **6**, 216–224 (2013). https://doi.org/10.1002/ase.1345
18. Seo, J.H., Smith, B.M., Cook, M., Malone, E., Pine, M., Leal, S., Bai, Z., Suh, J.: Anatomy builder VR: applying a constructive learning method in the virtual reality canine skeletal system. In: Andre, T. (ed.) AHFE 2017. AISC, vol. 596, pp. 245–252. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60018-5_24
19. Izard, S., Méndez, J.: Virtual reality medical training system. In: Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM 2016 (2016). https://doi.org/10.1145/3012430.3012560
20. Marks, S., White, D., Singh, M.: Getting up your nose. In: 2017 Symposium on Education on SIGGRAPH Asia - SA 2017 (2017). https://doi.org/10.1145/3134368.3139218
21. Paas, F., Renkl, A., Sweller, J.: Cognitive load theory and instructional design: recent developments. Educ. Psychol. **38**, 1–4 (2003). https://doi.org/10.1207/S15326985EP3801_1
22. Foo, J., Martinez-Escobar, M., Juhnke, B., Cassidy, K., Hisley, K., Lobe, T., Winer, E.: Evaluating mental workload of two-dimensional and three-dimensional visualization for anatomical structure localization. J. Laparoendosc. Adv. Surg. Tech. **23**, 65–70 (2013). https://doi.org/10.1089/lap.2012.0150
23. Hackett, M.: Medical holography for basic anatomy training. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), pp. 1–10 (2013)
24. Hart, S., Staveland, L.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 139–183 (1988). https://doi.org/10.1016/s0166-4115(08)62386-9
25. Boles, D., Adair, L.: The multiple resources questionnaire (MRQ). Proc. Hum. Factors Ergon. Soc. Ann. Meet. **45**, 1790–1794 (2001). https://doi.org/10.1177/154193120104502507
26. León-Carrión, J., León-Domínguez, U.: Functional near-infrared spectroscopy (fNIRS): principles and neuroscientific applications. Neuroimaging - Methods (2012). https://doi.org/10.5772/23146
27. Peck, E., Afergan, D., Yuksel, B., Lalooses, F., Jacob, R.: Using fNIRS to measure mental workload in the real world. Hum.–Comput. Interact. Ser. 117–139 (2014). https://doi.org/10.1007/978-1-4471-6392-3_6

28. Garg, A., Norman, G.R., Spero, L., Maheshwari, P.: Do virtual computer models hinder computer learning? Acad. Med. **74**(Suppl 10), S87–S89 (1999). https://doi.org/10.1097/00001888-199910000-00049
29. Roach, V.A., Brandt, M.G., Moore, C.C., Wilson, T.D.: Is three-dimensional videography the cutting edge of surgical skill acquisition? Anat. Sci. Educ. **5**(3), 138–145 (2012). https://doi.org/10.1002/ase.1262
30. Lisk, K., McKee, P., Baskwill, A., Agur, A.M.: Student perceptions and effectiveness of an innovative learning tool: anatomy glove learning system. Anat. Sci. Educ. **8**(2), 140–148 (2015). https://doi.org/10.1002/ase.1459

Shared Cognition, Team Performance and Decision-Making

Parole Board Personality and Decision Making Using Bias-Based Reasoning

Katy Hancock[1(✉)], Payton Brown[2], Antoinette Hadgis[2], Markus Hollander[3], and Michael Shrider[2]

[1] Criminal Justice Program, Murray State University, Murray, KY, USA
khancockll@murraystate.edu
[2] Sirius18, Melbourne, FL, USA
[3] Center for Bioinformatics, Saarland University, Saarbrücken, Germany

Abstract. Parole decision-making directly affects the lives of hundreds of thousands of individuals each year. While some research has been done on the offender and case factors that influence the parole decision, little research has been done on the characteristics of parole board members and how these influence the parole decision. Personality is one factor that influences the way individuals make decision.

This study simulates parole board members and their personalities, based upon the Myers Briggs Typology Inventory, and how these personalities impact the parole decision. The simulated parole board members used bias-based reasoning (BBR) in their decision-making process. BBR is a proprietary mathematical method for automating implementation of a belief-accrual approach to expert problem solving.

The results indicated that personality type was important for individual board members in the decision process. Specifically, the 'NT' subtypes were least likely to grant parole while the 'SF' subtypes were most likely to grant parole. Furthermore, parole boards composed of MBTI types likely ideal for careers on parole boards were less likely to grant parole.

These findings suggest that personality type is a key factor in parole decision-making and should be explored further. One important example is to examine the relationship between the MBTI makeup of a parole board and the accuracy of the parole decision. If there is an optimum mix of personalities for board effectiveness and efficiency, the offender, the justice system, and the community would benefit with regard to safety, possible financial savings, and the achievement of justice and correctional goals.

Keywords: Decision-making · Parole board

1 Introduction

The origins of parole in the United States can be traced back to the reformatories of the late 1800s. In these facilities, inmates could earn their way to an early release through good behavior, hard work, and participation in programming (Abadinsky 2011). From its inception, parole decision making was largely based upon a clinical model, meaning

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 255–272, 2018.
https://doi.org/10.1007/978-3-319-91467-1_21

the decision was based on the expertise and experience of the decision makers (Rhine et al. 2016). Due to criticisms regarding the inherent subjectivity and arbitrariness of such decisions, parole decision making has undergone major reforms, most notably since the 1970s (Abadinsky 2011). Standardized assessment tools and legislation creating guidelines for decision-making, though not without their problems, have resulted in a more transparent and objective parole decision-making process.

Today, parole structures and adjudication vary widely among the states, and a number of studies have examined how parole decisions are made. Indeed, as these decisions are necessarily made with limited information, involve the freedom of the potential parolees, and potentially impact the safety of the public, it is essential to fully understand and optimize the parole process. However, while researchers have studied the offender and case factors used to make parole decisions, there is a dearth of research regarding how the individual and collective characteristics of parole board members can influence decision-making. Personality is most notably a factor that influences decision-making. The current study will simulate parole board members and their personalities in order to explore how personality might impact these critical decisions.

2 Background

Parole is the early supervised release from a term of incarceration and includes a set of agreed upon conditions (U.S. Department of Justice (USDOJ) 2015). These conditions include standard conditions, such as regular meetings with a parole officer and refraining from associating with felons, and may also include specific conditions, such as random drug tests and treatment for drug offenders or internet restrictions for cybercriminals. The purposes of parole are threefold: to assist the offender with reintegration into society, to protect society from the offender, and to prevent needless imprisonment of those who are unlikely to commit future crime (USDOJ 2015).

In the United States, there were approximately 870,500 individuals on parole at yearend 2015, a 1.5% increase from the previous year (Kaeble and Bonczar 2017). Nationally, parole boards grant parole in roughly 43% of cases; this, however, varies greatly by state, with a low of 0% of cases being paroled (Illinois) and a high of 87% of cases being paroled (Arkansas and Nebraska) (Alper et al. 2016). The proportion of parolees who either completed their supervision or were granted early release (termed the exit rate) increased in 2015 to 54 per 100 parolees (Kaeble and Bonczar 2017). Parolees are mostly male (87%) and are most likely under supervision for a violent (32%) or drug (31%) crime. In addition, the parole population is overwhelmingly either White, non-Hispanic (44%), Black (38%), or Hispanic (16%) (Kaeble and Bonczar 2017). While these demographic and offense statistics do not match those of the general U.S. populations, they are somewhat representative of those of the prison population from which this group largely comes.

2.1 Parole Boards

Parole granting decisions happen in two main ways. One is termed mandatory parole, and is calculated based upon the amount of time served, incorporating good time[1], as set in state statute. The other is termed discretionary parole, and is granted by parole boards. The current study will focus on discretionary parole.

Parole Board Composition and Structure. As stated earlier, parole overall varies widely from state to state; this is true also for discretionary parole specifically. State parole boards range in size from a low of 3 members (Alabama) to a high of 17 members (New York). However, the average size of parole boards is about 7 members, with 30 state parole boards having between 5 and 8 members, inclusive. Some parole boards are only responsible for making decisions to release offenders or revoke their parole while others are also responsible for the supervision of offenders in the community (Abadinsky 2011). In addition, parole board members are typically appointed by the governor. However, in a few states, appointment may be by the Board of Corrections, the state attorney general, or a combination of these (Abadinsky 2011; Kinnevy and Caplan 2008).

Finally, parole boards vary with regard to their qualifications, which is a source of much criticism. Only a few states actually have any specific professional qualifications (Abadinsky 2011; Paparozzi and Caplan 2009). The joining of selection by appointment with no specific qualifications is said to insert politics into parole board selection and also to result in board members that are not qualified for the job.

Eligibility and Case Type. Another way that parole varies by state is through when offenders become eligible for parole. Many states have requirements about how much of the sentence the offender must have served before he or she becomes eligible for parole; this ranges from one-third to 85% (Alarid and Del Carmen 2011). These minimums may become even lower when good time is included; some states may also give credit for time served in jail prior to the sentence (Alarid and Del Carmen 2011). If an offender is denied parole, states vary as to when that offender will come up for parole again.

In addition, some states may require that the offender have a place to live and a job already established before they can be released (Abadinsky 2011; USDOJ 2015). Another way parole differs by state is in the types of cases that are eligible for parole. For example, some states have abolished parole for violent offenders or for serious repeat offenders. Other states only consider misdemeanors for discretionary parole. Still others only consider those convicted prior to a certain date; these states are typically moving to a mandatory parole model.

Parole Board Adjudication. The process by which decisions are made by parole boards is also determined by state. Boards may be divided and assigned to different parts of the state (Alarid and Del Carmen 2011). Parole boards vary with regard to how

[1] "Good time" refers to days, months, or years accumulated by the offender through avoidance of institutional rule-breaking and/or by participating in programming; this time is subtracted from their maximum sentence.

Table 1. Board sizes and adjudication

Board size	
3–4	6
5	13
6	4
7	13
8–11	8
12–17	5
Adjudication	
Majority	29
Unanimous	7
Other	8
Unknown	5

many make a quorum as well as how many have to meet for violent versus non-violent offenses. For example, certain types of serious offenses, such as sex offenses, may require a full board review (Alarid and Del Carmen 2011). Most states require either a majority or a unanimous decision; however, there are a few which specify a different number of votes. Moreover, some states have in-person hearings to inform the parole decision, while others have paper reviews where the offender is not present. See Table 1 for a breakdown of parole composition and adjudication.

Parole boards will have access to the offender's case file which contains case and background information, and quite probably a risk assessment. In fact, the vast majority of parole authorities (88%) report using some type of risk assessment in their decision-making (Kinnevy and Caplan 2008). This risk assessment calculates the offender's probability of parole failure, meaning reoffense or violating a condition of their parole. Typically in the in-person hearings, the offender, the prosecutor, law enforcement, the direct victim will be invited to make a statement, either oral or written (Alarid and Del Carmen 2011). Some states may allow the offender's family to be present. Using information from the case file and the hearing, the parole board will make their decision.

2.2 Decision-Making

A decision is a choice between alternatives (Houston 1999). In order to choose between alternatives, a decision maker applies decision rules to information (Stojkovic et al. 2015). Decision rules are criteria used to process information in order to make a choice and come in a variety of forms (Stojkovic et al. 2015). They can be quantitative in nature, such as a calculated risk score, and they can be clinical in nature, such as judgments based upon experience; individuals may even be unaware of their decision rules, such as a person who is unconsciously biased against minorities (Stojkovic et al. 2015). In the case of parole decision-making, decision rules include parole guidelines

created by the legislature or the correctional authority, judgements based upon training and education, standardized risk assessments, and, most importantly for this study, rules from personality traits. The information used by parole members for decision-making include the offender's case file as well as any oral or written statements made, as outlined above.

According to March and Simon's theory of bounded rationality, in order for decisions to be fully rational, decision-makers must have two things: perfect information and the appropriate amount of time to fully process and make the decision (March and Simon 1958). In real life, however, information is never perfect and decisions typically have a deadline. As such, March and Simon posited that decisions are not made rationally, but on the basis of bounded rationality. Due to a lack of information and time, decision made through bounded rationality will be acceptable, rather than optimal. "Satisficing" is a term used to describe attaining acceptable (versus ideal) results based upon incomplete information in a limited amount of time (Stojkovic et al. 2015).

For parole decision-making, the information used will definitely be incomplete. For example, information may be missing from an offender's case file, victims and family members may not make a statement and some institutional behaviors may have gone undetected. Indeed, the offenders themselves are incentivized to hide information about themselves when they make their statement at the hearing, a prime example being an offender faking remorse or a commitment to change. In addition, risk assessments used by parole boards are far from perfect predictors. In fact, one of the most common risk assessments, the LSI-R (Kinnevy and Caplan 2008) was found in one study to have a 30% false positive error rate (Hemphill and Hare 2004). Another popular risk assessment, the COMPAS, was found to have a 70% accuracy in predicting general rearrests, but did poorly in predicting future violent behavior (Zhang et al. 2014). Other scholars have also outlined the predictive and validity issues with risk assessments (Desmarais et al. 2016).

Moreover, parole caseloads and the time available to hold hearings will limit the amount of time parole boards have to process information from each case. The average parole board caseload in 2006 was approximately 35 cases for each working day (Kinnevy and Caplan 2008). Therefore, due to imperfect information and limited processing time, parole board members must "satisfice" when they decide whether or not to grant parole. Because the reality of parole decision situations makes bounded rationality a necessity, improving the amount of information available and fully understanding the decision rules used (such as those derived from personality traits) are critical to optimizing parole decision-making.

Factors Impacting Parole Decision-Making. A number of studies have been conducted to determine which factors are most important in determining the parole decision. These factors are usually offender and case factors. In a 2008 national survey, parole board chairpersons reported the nature and severity of the current offense, the offender's prior record and risk assessment, the offender's institutional disciplinary record and program participation, previous parole adjustment, and victim input as the most important factors in the parole decision (Ruhland et al. 2016). In addition, the

parole chairpersons overwhelmingly (87%) stated that they either agreed or strongly agreed that risk assessments were essential in parole decision making (Ruhland et al. 2016, p. 8).

Other studies have found similar results, concluding that parole authorities consider offense seriousness, institutional misconduct, and parole readiness (Huebner and Bynum 2006), institutional behavior and risk of reoffense (Carroll et al. 1982), and risk of future crime (Carroll 1978) as most important. While victims do not often take advantage of the opportunity to make a statement, there is evidence that when they do, these statements are related to a lower likelihood of the board granting parole (Bernat et al. 1994; McLeod 1989; Morgan and Smith 2005; Parsonage et al. 1994; Smith et al. 1997). In addition, the offender's age, history of drug and alcohol use, percentage of the sentence that has been served, and parole hearing demeanor are often considered as well (Abadinsky 2011; Kinnevy and Caplan 2008; Ruhland et al. 2016). In the current study, the simulated parole board members will use the following factors for their decision-making: severity of the offense, prior record, institutional record, rehabilitative program completion, risk of re-offense, age of the offender, victim statements, offender remorse, and abuse of alcohol and drugs. These factors make up the information upon which the decision rules of the different personalities will operate.

Personality and Decision-Making. Because personality is said to impact choices, values, and reactions (Myers 1962), it seems logical that personality will act upon the different parole factors to influence parole decisions. Indeed, one of the few studies that focused on parole board members rather than offender and crime characteristics found that parole decision making was an outcome of interactions within the parole board (Conley and Zimmerman 1982).

Probably the most well-known and widely used personality test is the Myers-Briggs Typology Indicator (MBTI) (Furnham 2017). The theory behind the MBTI is that differences in the way people take in information (perception) and make conclusions about that information (judgement) relate to the variations in values, reactions, choices, and behaviors (Myers 1962). Under the MBTI, there are 4 basic preferences that make up a person's personality: between extroversion and introversion (E or I), between sensing or intuition (S or N), between thinking or feeling (T or F), and between judging or perceiving (J or P). (Myers 1962; The Myers & Briggs Foundation 2017).

The preference between extroversion and introversion refers to whether an individual prefers to focus on the outer world or their inner world. The preference to focus on basic information or interpret and add meaning is the sensing/intuition preference. The preference between an initial focus on logic and consistency or people and circumstances is the thinking/feeling preference. Finally, the judging/perceiving preference is between having things decided or being open to new information (Myers 1962; The Myers & Briggs Foundation 2017). Thus, a person who prefers to focus on their inner world, to focus on basic information, to focus on logic and consistency, and who prefers to have things decided would be said to have an ISTJ personality. The simulated parole board members have been coded so that they have various types of the 16 MBTI personalities that will function as decision rules for them.

3 Methodology

The current study simulates parole board members and their personalities in order to explore how personality might impact parole adjudication. There are 16 different sets of code that are used to represent the different personality types. A number of different trials were performed to represent different parole situations because, as stated above, there exists wide variation in parole board makeup and adjudication.

3.1 Coding

The simulated parole board members will utilize bias-based reasoning (BBR) in their decision-making process. BBR is a proprietary mathematical method for automating implementation of a belief-accrual approach to expert problem solving (Hancock 2012). It enjoys the same advantages human experts derive from this approach; in particular, it supports automated learning, conclusion justification, confidence estimation, and natural means for handling both non-monotonicity and uncertainty. Dempster-Shafer Reasoning is an earlier attempt to implement belief-accrual reasoning, but suffers some well-known defects (Lotfi paradox, constant updating of parameters, monotonic, no explicit means for uncertainty) (Boyen and Koller 1998). BBR overcomes these.

Pose a problem for a human expert in their domain, and you will find, even given no evidence, that they have an a priori collection of beliefs about the correct conclusion. For example, a mechanic arriving at the repair shop on Tuesday morning already holds certain beliefs about the car waiting in Bay 3 before she knows *anything* about it. As she examines the car, she will update her prior beliefs, accruing "bias" for and against certain explanations for the vehicle's problem. At the end of her initial analysis, there will be some favored (belief = large) conclusions, which she will test, and so accrue more belief and disbelief. Without running decision trees, applying Bayes' Theorem, or using margin maximizing hyperplanes, she will ultimately adopt the conclusion she most believes is true. It is this "preponderance of the evidence" approach that best describes how human experts actually reason, is fully in line with March and Simon's theory of Bounded Rationality, and it is this approach that is modeled in BBR.

For simplicity and definiteness, the reasoning problem will be described here as the use of evidence to select one or more possible conclusions from a closed, finite list that has been specified a priori (the "Classifier Problem"). Expert reasoning is based upon facts (colloquially, "interpretations of the collected data"). Facts function both as indicators and contra-indicators for conclusions. Positive facts are those that increase our beliefs in certain conclusions. Negative facts are probably best understood as being exculpatory: they impose constraints upon the space of conclusions, militating against those unlikely to be correct. Facts are salient to the extent that they increase belief in the "truth", and/or increase "disbelief" in untruth (Delmater and Hancock 2001).

A rule is an operator that uses facts to update beliefs by applying biases. In software, rules are often represented as structured constructs such as IF-THEN-ELSE, CASE, or SWITCH statements. We use the IF-THEN-ELSE in what follows.

Rules consist of an antecedent and a multi-part body. The antecedent evaluates a BOOLEAN expression; depending upon the truth-value of the antecedent, different

parts of the rule body are executed. The following is a notional example of a rule. It tells us qualitatively how an expert might alter her beliefs about an unknown animal should she determine whether or not it is a land-dwelling omnivore:

```
If  (habitat = land) AND (diet = omnivorous) THEN
INCREASE BELIEF(primates, bugs, birds)
INCREASE DISBELIEF(bacteria, fishes)
ELSE
INCREASE DISBELIEF(primates, bugs, birds)
INCREASE BELIEF(bacteria, fishes)
End Rule
```

If we have an INCREASE BELIEF function, and a DECREASE BELIEF function ("aggregation functions", called AGG below), many such rules can be efficiently implemented in a looping structure:

In a data store:

Tj(**F**i)	truth-value of predicate j applied to fact **F**i
bias(kj, 1)	belief to accrue in conclusion k when predicate j true
bias(kj, 2)	disbelief to accrue in conclusion k when predicate j is true
bias(kj, 3)	belief to accrue in conclusion k when predicate j false
bias(kj, 4)	disbelief to accrue in conclusion k when predicate j is false

Multiple rule execution in a loop:

```
IF Tj(F)=1 THEN           'if predicate j true for Fi...
   FOR k=1 TO K           'for conclusion k:
      Belief (k)=AGG(B(k,i),bias(k,j,1))       'true: ac-
crue belief bias(k,j,1)
      Disbelief(k)=AGG(D(k,i),bias(k,j,2)) 'true: accrue
disbelief bias(k,j,2)
   NEXT k
ELSE
   FOR k=1 TO K
'for conclusion k:
      Belief(k)=AGG(D(k,i),bias(k,j,3))      'false: ac-
crue belief bias(k,j,3)
      Disbelief (k)=AGG(B(k,i),bias(k,j,4)) 'false: ac-
crue disbelief bias(k,j,4)
   NEXT k
END IF
```

This creates a vector B of beliefs (b(1), b(2), …, b(K)) for each of the conclusions 1, 2, …, K, and a vector D of disbeliefs (d(1), d(2), …, d(K)) for each of the conclusions 1, 2, …, K. These must now be adjudicated for a final decision.

Clearly, the inferential power here is not in the rule structure, but in the "knowledge" held numerically in the biases. As is typical with heuristic reasoners, BBR allows the complete separation of knowledge from the inferencing process (Friedman et al. 1998). This means that the structure can be retrained, even repurposed to another problem domain, by modifying only data; the inference engine need not be changed. An additional benefit of this separability is that the engine can be maintained openly apart from sensitive data.

Summarizing (thinking again in terms of the Classifier Problem): When a positive belief heuristic fires, it accrues a bias $\beta > 0$ that a certain class is the correct answer; when a negative heuristic fires, it accrues a bias $\delta > 0$ that a certain class is the correct answer. The combined positive and negative biases for an answer constitute that answer's belief.

After applying a set of rules to a collection of facts, beliefs and disbeliefs will have been accrued for each possible conclusion (classification decision). This ordered list of beliefs is a belief vector. The final decision is made by examining this vector of beliefs, for example, by selecting the class having the largest belief-disbelief difference.

BBR can also incorporate variation in decision-making that is a result of uncertainty, termed bias-variability (Cover and Thomas 2001). Human decision-making is variable, and an individual, given the same input, may make two different decisions in two different situations. The introduction of some randomness, or bias-variability, into the decision-making can simulate the non-determinism inherent in human decision-making. Thus, each factor has some measure of variation in its code, using a subroutine; in the current study, the bias-variability is small enough that it only impacts decisions that would be considered "borderline" cases.

3.2 Parole Board Members

In the parole board situation, parole board members use a variety of factors (evidence) to make the decision to grant or not grant parole. In the current experiment, the simulated parole board members will review evidence regarding severity of the offense, prior record, institutional record, rehabilitative program completion, risk of re-offense, age of the offender, victim statements, offender remorse, and abuse of alcohol and drugs. Severity of the offense, prior record, institutional behavior, and risk of re-offense were considered the strongest factors, as per prior research.

As an example, consider severity of the offense. If an offender has been incarcerated for a relatively minor offense, then, all else being equal, that would increase the board members' belief that parole should be granted. Conversely, if an offender has been incarcerated for a severe offense, say murder, this would increase the parole board members' disbelief that parole should be granted. All else being equal, there is probably a low level of uncertainty about this decision. Conversely, while demeanor is a factor that parole board members consider, this is not only a more subjective factor than severity of offense (a legal concept) and considered less predictive of later behavior, but remorse and respect are easily faked by an offender. As such, the importance of this factor will be lower and the uncertainty for this factor is going to be higher.

Once all of the factors are considered, each parole board member will have a belief score between -1 and $+1$ as to whether parole should be granted. The threshold for parole was set at 0.4; scores higher than this resulted in granting parole, lower resulted

in denying parole. In this way, each parole member has a vote based upon their belief score; these votes can be combined in whatever way the method of adjudication for the current trial requires (i.e. majority, unanimous, etc.) to get the ultimate decision.

Personality. In addition, it is theorized that different personality types will vary with regard to how much different factors influence them. The personality variations are incorporated into the code by changing the amount of belief (□s) and disbelief (□s) each factor causes for each personality type. Moreover, the influence of uncertainty is also altered for each personality type. These personality variations resulted in 16 different sets of rules for parole decision making.

The following represent some examples of the thought process behind the coding of parole board members' responses to the parole decision factors. These examples are far from exhaustive.

Severity of Offense, Prior Record, and Risk of Reoffense. While most, if not all, individuals are likely to consider these as important (regardless of personality type), different types may react to these differently. For instance, 'feelers', 'intuitives' and 'perceivers' might be slightly more open to considering mitigating factors when it comes to severity of offense and prior record. In addition, it is possible that risk of reoffense is generally less important to 'feelers' compared to more emotionally salient factors like the offense itself and victim statements. 'Sensors' might be more likely to take the risk assessment score as is, whereas 'intuitives' might be more likely to extrapolate from the score, but might also be more likely to question the accuracy of the risk assessment.

Victim Statements and Offender Remorse. Victim statements and offender remorse might not be as important to board members as the severity of offense; the exception might be 'feelers' (as stated above) who place high value on personal and emotional factors, whereas 'thinkers' might be more inclined to put the more objective factors first.

Abuse of Substances. 'Intuitives' might have less of a problem with mild abuse, but see heavy abuse without treatment as indicative of future problems.

Variability in Decision-Making. When it comes to making choices, variability within individuals is possible. 'Perceivers' on average are going to exhibit higher variability in their responses than 'judgers', and, to a lesser degree, 'intuitives' more than 'sensors'.

3.3 Offenders

For this experiment, 500 offenders were generated with different scores on each of the factors to be considered by the parole board members. Severity of offense ranged from 1–5, prior record ranged from 1–6, institutional behavior 0–5, and risk assessment from 1–10. For all of these, lower scores are better. Additionally, program completion and victim statements were dichotomous as either present or not. Use of substances ranged from 0–2, with 0 being no history and 2 being a severe user. Demeanor was an abstract

scale ranging from 0–1, with 1 being a great demeanor. Finally, age ranged from 17–75, with the age makeup of the 500 offenders closely mirroring that of the overall U.S. prison population.

3.4 Experiments

Experiment 1: MBTI Type and the Parole Decision. For this experiment, each MBTI type made parole decisions as individuals. 1,000 trials were run in which each type made parole decisions for the 500 simulated offenders.

Experiment 2: Career-Based Parole Boards. One of the ways MBTI is used is to assist with career guidance. Different personality types seem fitted to certain types of careers. For this experiment, the personality types were divided into two groups and seven personalities were randomly selected from each group (reflecting the average parole board size):

Group 1: ISTJ, ESTP, ISTP, ISFJ, ENTJ, INTJ, INTP, INFP
Group 2: ESTJ, ESFJ, ESFP, ISFP, ENTP, INFJ, ENFJ, ENFP

Group 1 consists of personality types whose most recommended careers fall into criminal justice, psychiatry, counseling, rehabilitation, and social work (Schaubhut and Thompson 2008), careers which seem fitting for parole board members. Indeed, for those states that have requirements for parole board members, having experience in one of these fields is a common standard.

For the experiment, 1 Million simulations were run using the two groups. There were approximately 200 Group 1 Boards and 200 Group 2 boards and each board processed all 500 candidates 5 times. To adjudicate, the scores of each individual member (from summing the betas and deltas described above) were averaged to get an overall board score.

4 Results

4.1 Experiment 1

The results of the first experiment can be seen in Table 2. As shown, there was obvious variation in the parole decision-making of the MBTI types, with the ISFJ/ESFJ types granting parole about half the time while the INTJ/ENTJ types denied parole in almost all the cases. The ISFPs and ESFPs are most in line with the national parole grant rate as reported by Alper et al. (2016).

Looking at the differences in percentages granted/denied, there are clear gaps between sets of MBTI types; these are marked by lines in the table. The decisions seem to largely cluster into 4 groups: the NTs, the STs, the NFs, and the SFs. As such, it appears the strongest personality factors in parole decision-making are the middle two preferences: sensing/intuition and thinking/feeling. This is illustrated in Fig. 1.

Table 2. Parole decisions by MBTI type

Type	Granted	Denied
INTJ	2.20%	98.36%
ENTJ	2.20%	98.36%
INTP	2.40%	98.16%
ENTP	2.40%	98.16%
ISTP	8.20%	92.28%
ESTP	8.20%	92.28%
ISTJ	10.59%	89.85%
ESTJ	10.59%	89.85%
INFP	24.77%	75.46%
ENFP	24.77%	75.46%
INFJ	27.77%	72.42%
ENFJ	27.77%	72.42%
ISFP	43.54%	56.41%
ESFP	43.54%	56.41%
ISFJ	50.74%	49.13%
ESFJ	50.74%	49.13%

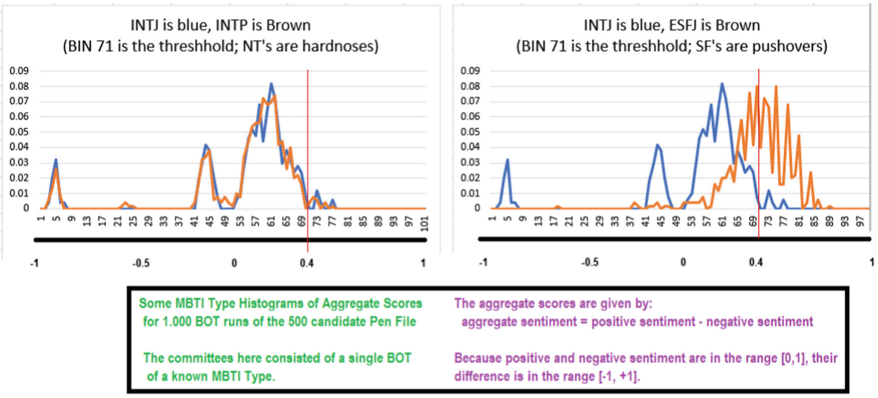

Fig. 1. Histograms of aggregate scores of NT/SF decisions (Color figure online)

What is also interesting is that, while the 'thinkers' and 'feelers' are clearly divided on parole, the 'sensors' and 'intuitives' are not. This distribution suggests that the combination of the two middle preferences is important. This makes sense, given that the two middle letters indicate how one gathers information and makes decisions.

What is perhaps most interesting is the judging/perceiving preference. Indeed, this preference seems to become more important in the parole decision as one moves down the table. For the SF personality types, the judging/perceiving preference seems more influential in their decision making; in contrast, the judging/perceiving preference seems to have little impact with the NTs. Furthermore, the Judging/Perceiving

preference seems to have a different relationship with the NT personality type, as with the NTs, 'judgers' are least likely to grant parole, while with all the other types, they are most likely to grant parole.

4.2 Experiment 2

The results of experiment 2 can be found in Table 3.

Table 3. Parole decisions by career groups

	Granted	Denied
Group 1	8.90%	91.10%
Group 2	26.47%	73.53%

As can be seen, Group 1, the career group that, on the surface, seems ideal for parole board membership, was unlikely to grant parole. This is not surprising in light of the results of experiment 1, given that 3 of the 4 NT types were in Group 1 while 3 of the 4 SF types were in Group 2. These results can also be seen in Fig. 2.

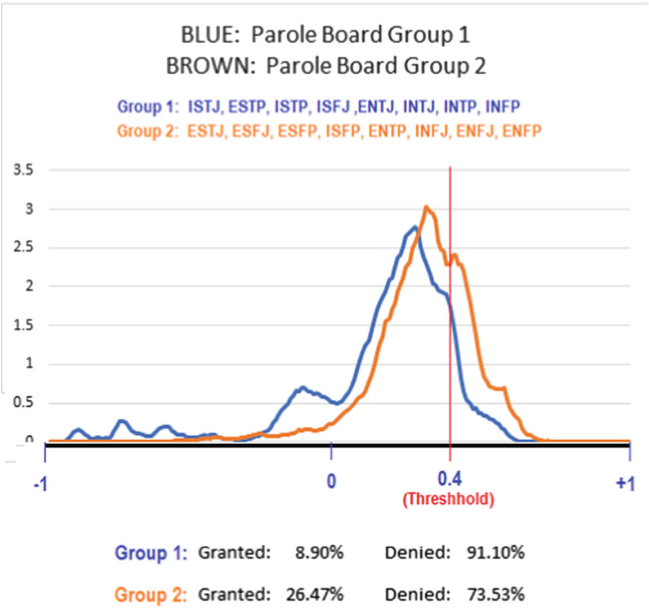

Fig. 2. Histogram of career group decisions (Color figure online)

In addition, Table 4 shows the correlation between the MBTI makeup of the boards and the overall board score. The strongest positive correlations are for ESFPs and ESFJs, indicating that the more individuals with these personality types, the higher the

Table 4. Correlations between score and board makeup

	Score
ESFP	0.2125
ESFJ	0.2115
ENFJ	0.1962
INFJ	0.1919
ISFP	0.1866
ENFP	0.1770
ESTJ	0.1009
ENTP	0.0762
ISFJ	−0.0878
INFP	−0.1121
ENTJ	−0.1899
ISTP	−0.1964
INTP	−0.1985
INTJ	−0.2027
ISTJ	−0.2037
ESTP	−0.2060

scores for the board, thus being more likely to reach the threshold to grant parole. In contrast, the strongest negative correlations were for ISTJs and ISTPs; boards with more of these individuals had lower scores and thus were less likely to grant parole. The weakest correlations were for the ENTPs and the ISFJs (which are, interestingly, exact opposites).

Table 5 shows the correlation between the overall board scores and the offender factors. The factors most strongly correlated to the board's overall score were the completion of a rehabilitative program, the severity of the offense, and the risk assessment. This is in line with what parole board chairs have said are the most important factors in the parole decision (Ruhland et al. 2016). The weakest correlation was for the use of substances.

Table 5. Correlations between board score and offender factor

	Score
Program completion	0.4703
Demeanor	0.2878
Substance abuse	0.0748
Age	−0.1602
Prior record	−0.2452
Institutional record	−0.2722
Victim statement	−0.2832
Risk assessment	−0.3607
Severity of offense	−0.3844

5 Discussion

The results of these experiments have suggested the importance of personality type to the parole decision-making process. As such, further research in this area is warranted, given the implications of parole decisions for social welfare, justice, and community safety.

Based upon experiment 1, there are clearly some MBTI types that are more likely to grant parole than others. Again, while the 'thinkers' and 'feelers' seem divided on parole, the 'sensors' and 'intuitives' are not. This relationship does make sense: 'thinkers' tend to put more weight on objective principles, hard facts, and logic, while 'feelers' put more weight on the personal or human aspect. It is not unexpected that the 'thinkers' would be "harsher" with regard to parole decisions and the 'feelers' more "lenient." In contrast, 'sensing' versus 'intuition' is more about how one orders and processes information, so it seems logical this would be less influential in the decision outcome and more related to how the individual arrived at the decision.

Moreover, the results showed that the "parole" career group was much less likely to grant parole than the other group. While there is some disagreement on career recommendations for MBTI types, three of the four NT types were in the career group. Due to these individuals' emphasis on logic, theory, and rationality, it is not surprising to find them in this group, as they are likely more rule and law oriented. Along the same lines, it makes sense that this group would be "harsher" with regard to parole. An implication of this is that drawing parole board members from a wider variety of career fields may be worth consideration. The greater assortment of careers could potentially diversify the personality makeup of the board, perhaps resulting in lower denial rates, if indeed that is found to be desirable.

The current study found a relationship between MBTI type and likelihood to grant or deny parole; this finding is value neutral. Consequently, one important topic to research in the future would be the connection between board member personality type and parole board effectiveness. One way this could be measured would be the accuracy of the parole decision-whether the individual who is paroled reoffends. It is possible that some MBTI types, because of the way they filter and prioritize different types of information, may be better able to discern an offender's potential for success upon release. Indeed, Sanchez (2011) studied MBTI types and their ability to detect lies; she hypothesized that ENTPs would be most accurate due to their focus on others (extroversion), their ability to see patterns of deceit cues (related to intuition), their focus on logic and analysis (thinking), and their less rigid thinking patterns (perceiving). Ultimately her results supported her hypothesis. The ability to detect lies would certainly be beneficial when interviewing a parole candidate.

This examination of effectiveness could also be extended to looking at the interplay between personalities on parole boards. There may indeed be an optimum mix of MBTI types on a parole board that yield more accurate or efficient decisions. For example, when it comes to 'feeling' versus 'thinking' dimension, the MBTI dimension correlates with the Big 5 agreeability dimension, where MBTI 'feelers' are far more agreeable than their 'thinker' counterparts (Furnham 1996). This may make 'feelers' generally more inclined to go with the group consensus, whereas 'thinkers' are generally more

inclined to ignore it and stick to their own opinions. However, both 'feelers' and 'thinkers' have their judgements to which they want to stick, and a 'feeler' can cling just as strongly to their values as a thinker to their reasons, as long as they feel strongly enough about it. In that case, a 'feeler' might be similarly unwilling to compromise. A 'thinker' on the other hand might think that it's more reasonable to compromise when they are not that certain about their own opinion. Similarly, a 'perceiver' might be generally more inclined to be laid back and open to adjust to the group consensus than a 'judger', who is generally more likely to stick to their principles. However, 'perceivers' have their principles, too, and can become very rigid about them. Thus, the ability to reach group consensus may be dependent not just on their attitude towards group consensus itself, but also on the strength of their conviction that they made the right decision. At the very least, the findings of this study indicate the need for more scrutiny regarding who is chosen for parole boards and stronger standards for selection.

It would also be valuable to examine in depth the interplay between the 'judging' preference and the parole decision. Experiment 2 showed an interesting relationship between the 'judging' preference and the NT personality type. The NT types tend to be logical, objective, analytical, and impersonal (The Myers & Briggs Foundation 2018). It is possible the more task-oriented, structured 'judging' preference amplifies the NT characteristics in ways that it doesn't with other types. This interaction could potentially result in seemingly "harsher" decisions.

Finally, it would be interesting to examine the relationship between personality and other criminal justice decisions, including judges, attorneys, and correctional officers. Indeed, while parole boards generally have limited time for each decision, police officers must often make split second decisions, so examining personality types in policing is also important. On the other side of the system, environmental conditions, genetics, and socialization may interact with personality type to influence the decision to commit crime.

6 Conclusion

Personality is a critical factor in decision-making, and this fact seems no different for parole boards. The factors influencing parole decisions are important as these decisions influence rehabilitation of offenders, the safety of the public, and the legitimacy of the criminal justice system.

References

Adabinsky, H.: Probation and Parole: Theory and Practice, 11th edn. Pearson, London (2011)

Alarid, L., Del Carmen, R.: Community-Based Corrections, 8th edn. Wadsworth Cengage, Boston (2011)

Alper, M.E., Reitz, K.R., Rhine, E.R., Watts, A.L., Robey, J.P.: By the Numbers: Parole Release and Revocation Across 50 States. Robina Institute of Criminal Law and Criminal Justice: University of Minnesota, Minneapolis (2016)

Bernat, F.P., Parsonage, W.H., Helfgott, J.: Victim impact laws and the parole process in the United States: balancing victim and inmate rights and interests. Int. Rev. Victimology **3**, 121–140 (1994)

Boyen, X., Koller, D.: Tractable inference for complex stochastic processes. In: Proceedings of 14th Conference on Uncertainty in Artificial Intelligence, Madison, WI, pp. 33–42 (1998)

Carroll, J.S.: Causal attributions in expert parole decisions. J. Pers. Soc. Psychol. **36**(12), 1501–1511 (1978)

Carroll, J.S., Wiener, R.L., Coates, D., Galegher, J., Alibrio, J.J.: Evaluation, diagnosis, and prediction in parole decision making. Law Soc. Rev. **17**(1), 199–228 (1982)

Conley, J.A., Zimmerman, S.E.: Decision making by a part-time parole board: an observational and empirical study. Crim. Justice Behav. **9**, 396–431 (1982)

Delmater, R., Hancock, M.: Data Mining Explained. Digital Press, Boston (2001)

Desmarais, S.L., Johnson, K.L., Singh, J.P.: Performance of recidivism risk assessment instruments in the U.S. correctional settings. Psychol. Serv. **13**(3), 206–222 (2016)

Friedman, N., Koller, D., Pfeffer, A.: Structured representation of complex stochastic systems. In: Proceedings of 15th National Conference on Artificial Intelligence, Madison, WI, pp. 157–164 (1998)

Furnham, A.: The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and the NEO-PI five factor model of personality. Pers. Individ. Differ. **21** (2), 303–307 (1996)

Furnham, A.: Myers-Briggs Type Indicator (MBTI). In: Encyclopedia of Personality and Individual Differences. https://link.springer.com/referenceworkentry/10.1007%2F978-3-319-28099-8_50-1. Accessed 30 Dec 2017

Hancock, M.: Practical Data Mining. CRC Press, Boca Raton (2012)

Hemphill, J.F., Hare, R.D.: Some misconceptions about the PCL-R and risk assessment: a reply to Gendreau, Goggin, and Smith. Crim. Justice Behav. 31, 203–243 (2004)

Houston, J.: Correctional Management: Functions, Skills, and Systems, 2nd edn. Nelson-Hall, Wokingham (1999)

Huebner, B.M., Bynum, T.S.: An analysis of parole decision making using a sample of sex offenders: a focal concerns perspective. Criminology **44**(4), 961–991 (2006)

Kaeble, D., Bonczar, T.P.: Probation and parole in the United States, 2015 [250230]. U.S. Department of Justice, Bureau of Justice Statistics (2017)

Kinnevy, S.C., Caplan, J.M.: Findings from the APAI International Survey of releasing authorities. Center for Research on Youth and Social Policy, University of Pennsylvania, Philadelphia, PA (2008)

K-L Divergence is one possible measure of inferencing error; see

Cover, T., Thomas, J.: Elements of Information Theory. Wiley, New York (2001)

March, J., Simon, H.: Organizations, 2nd edn. Blackwell Publishers, Hoboken (1958)

McLeaod, M.: Getting free: victim participation in parole board decisions. Crim. Justice **4**, 12–15/41–43 (1989)

Morgan, K., Smith, B.L.: Victims, punishment, and parole: the effect of victim participation on parole hearings. Criminol. Public Policy **4**(2), 333–360 (2005)

Myers, I.B.: The Myers-Briggs Type Indicator (Manual). Consulting Psychologists Press, Inc., Sunnyvale (1962)

Paparozzi, M.A., Caplan, J.M.: A profile of paroling authorities in America: the strange bedfellows of politics and professionalism. Prison J. **89**(4), 401–425 (2009)

Parsonage, W.H., Bernat, F.P., Helfgott, J.: Victim impact testimony and Pennsylvania's parole decision making process: a pilot study. Crim. Justice Policy Rev. **6**, 187–206 (1994)

Rhine, E.E., Petersilia, J., Reitz, K.R.: The future of parole release. Crime Justice **46**(1), 279–338 (2016)

Ruhland, E.L., Rhine, E.E., Robey, J.P., Mitchell, K.L.: The Continuing Leverage of Releasing Authorities: Findings from a National Survey. Robina Institute of Criminal Law and Criminal Justice, University of Minnesota, Minneapolis (2016)

Sanchez, T.: To catch a liar: a signal detection analysis of personality and lie detection. Poster Presented at the 39th Annual Western Pennsylvania Undergraduate Psychology Conference, April 2011

Schaubhut, N.A., Thompson, R.C.: MBTI Type Tables for Occupations. CPP Inc., Sunnyvale (2008)

Smith, B.L., Watkins, E., Morgan, K.: The effect of victim participation on parole decisions: results from a southeastern state. Crim. Justice Policy Rev. **8**(1), 57–74 (1997)

Stojkovic, S., Kalinich, D., Klofas, J.: Criminal Justice Organizations: Administration and Management, 6th edn. Cengage Learning, Boston (2015)

The Myers & Briggs Foundation. MBTI basics (2017). http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1. Accessed 30 Dec 2017

The Myers & Briggs Foundation. Function pairs (2018). http://www.myersbriggs.org/my-mbti-personality-type/understanding-mbti-type-dynamics/function-pairs.htm?bhcp=1

United States Department of Justice. U.S. Parole Commission: Frequently Asked Questions (2015). https://www.justice.gov/uspc/frequently-asked-questions

Zhang, S.X., Roberts, R.E.L., Farabee, D.: An analysis of prisoner reentry and parole risk using COMPAS and traditional criminal history measures. Crime Delinq. **60**(2), 167–192 (2014)

Validation of a Maritime Usability Study with Eye Tracking Data

Odd Sveinung Hareide[1(✉)] and Runar Ostnes[2]

[1] Royal Norwegian Naval Academy, Navigation Competence Centre, Norwegian Defence University College, Bergen, Norway
oddsveinung.hareide@sksk.mil.no
[2] Institute of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Aalesund, Norway

Abstract. The main objective of the navigation system on board a High Speed Craft (HSC) is contributing to safe operation, which is supported by a high degree of situation awareness for the navigator. On the modern HSC bridge, an increasing amount of displays and support systems has been introduced, with computers being networked and integrated information presented on Multi-Function Displays (MFDs). Eye tracking data in human-computer interaction is a valuable tool to identify challenges with design and user interfaces, and to better understand the workload of the subject. This paper presents and analyse two eye tracking data sets collected to validate a mid-life update of a HSC navigation system, and outlines the challenges when collecting eye tracking data in an operational environment. Data collection with Eye Tracking Glasses (ETGs) is proven to be a valuable tool, but the quantitative data needs to be supported by qualitative data to be unambiguous.

Keywords: Maritime · High speed · Navigation · Eye tracking data
Eye tracking glasses · Navigation system

1 Introduction

High speed navigation in littoral waters is a challenging task. Both civilian and military High Speed Crafts (HSC) are operating in speeds above 20 knots (37 km/h) and some exceeding 60 knots, making the safe and efficient conduct of the passage crucial.

To support the navigation process, the bridge is equipped with MFDs to facilitate the information management in the navigation system for the navigation team [1]. The navigation system is integrated and networked together, and information is typically presented and integrated on a MFD on the Electronic Chart Display and Information System (ECDIS), radar application and application with information about the ship propulsion and technical systems (conning). The Situation Awareness (SA) of the navigator is crucial in order to facilitate for the safe and efficient navigation, and the navigation system aims to support a higher degree of SA [2].

Several studies have highlighted the challenge with information overload for the navigation team [3–8], and raises the question whether a bridge design and layout supports the safe and efficient navigation of the vessel.

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 273–292, 2018.
https://doi.org/10.1007/978-3-319-91467-1_22

To better understand the task of navigation and what the navigator is addressing during a passage, eye tracking data can be collected and analysed. ETGs can provide sufficient freedom of mobility for the test participants, and has shown good potential in better understanding the task of the (HSC) navigator [9, 10].

Eye tracking data can be collected by using ETGs, and the use of ETGs has shown good potential in maritime usability studies [11–13]. Previous studies highlighted design and Graphical User Interface (GUI) issues on board the Skjold-class Corvette (Fig. 1) bridge navigation system [9, 11, 14], and these were corrected in a mid-life update [15]. This paper presents a pre- and post-mid-life update eye tracking data set collected to validate and support the findings in the pre mid-life update study.

The research question in the article is if eye tracking data collected from ETGs can be used to validate a design-review of a maritime HSC bridge.

1.1 Decision Making in High Speed Navigation

HSC navigation is most commonly conducted in a navigation team, consisting of two persons, the Officer of the Watch (OOW) and the Navigator, which share the tasks given to achieve safe and efficient navigation. Dependent on the confinement of the waters, weather and speed, the navigation team workload is high [16]. Safe navigation means that no incidents or accidents occur, while efficient navigation means that the speed potential of the vessel is utilized [17].

Figure 1 shows the Royal Norwegian Navy Corvettes, with speeds exceeding 60 knots (110 km/h or 70 mph).

Fig. 1. Skjold-class Corvette

The conduct of a safe passage with a HSC is a complex task, conducted in a sociotechnical system as a navigation team [18]. To support safe and efficient navigation, the navigation team uses a methodology to aid the decision making process and increase the SA, known as the phases of navigation [1] or Dynamic Navigation (DYNAV) [19, 20]. The conduct of safe and efficient planning is shown in Fig. 2, and is an iterative process.

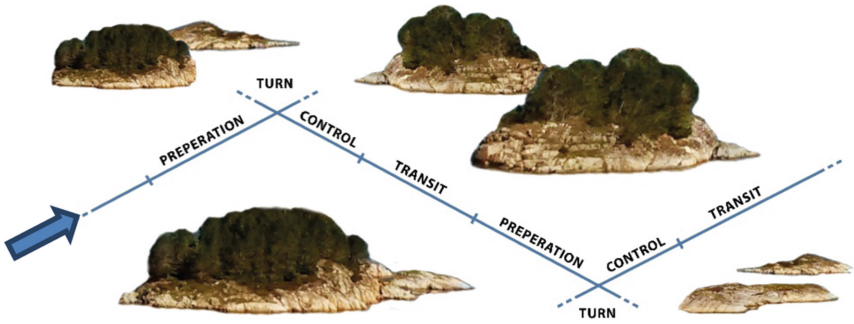

Fig. 2. Iterative process of (littoral high speed) navigation

In littoral waters there are multiple obstacles for navigation, making high speed navigation challenging. Each leg will vary in length, but as an example, a leg of one nautical mile (1 nm = 1852 m), will take 1 min to complete in 60 knots. In demanding waters, consecutive legs are often less than 0.5 nm in distance, making the decision process before the next leg less than 30 s.

In each phase of navigation, the navigator has a mental checklist to follow, and it is important that the navigators prioritize in order to have time to finish one phase before the next one starts, in order to maintain a high degree of SA. The navigator's SA consist of spatial-, task- and system awareness [6, 21], and the complexity of these factors affect the navigator's workload as shown in Fig. 3. Note that the bottom line in Fig. 3 is meant as examples, and is not complete.

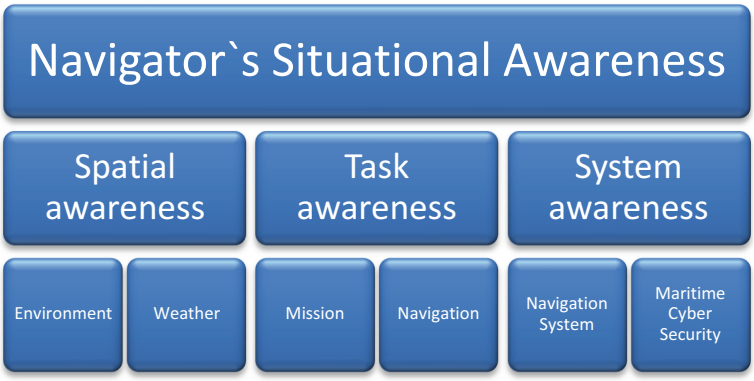

Fig. 3. Navigator's SA [21]

As navigation is conducted in a team, the communication skills is important to create and maintain a shared mental model in the navigation team, and the communication is mainly conducted in accordance with standard operating procedures. The integrated navigation information on the displays provide some of the basis of the navigation team shared mental model, however this information collection is non-verbal and could thus be interpreted differently by the operators [16].

1.2 Vulnerabilities in an Integrated Navigation System

Navigation systems on a modern HSC are networked, and the navigation sensors are integrated. The integrated information is presented on one or several MFDs, as shown in Fig. 4.

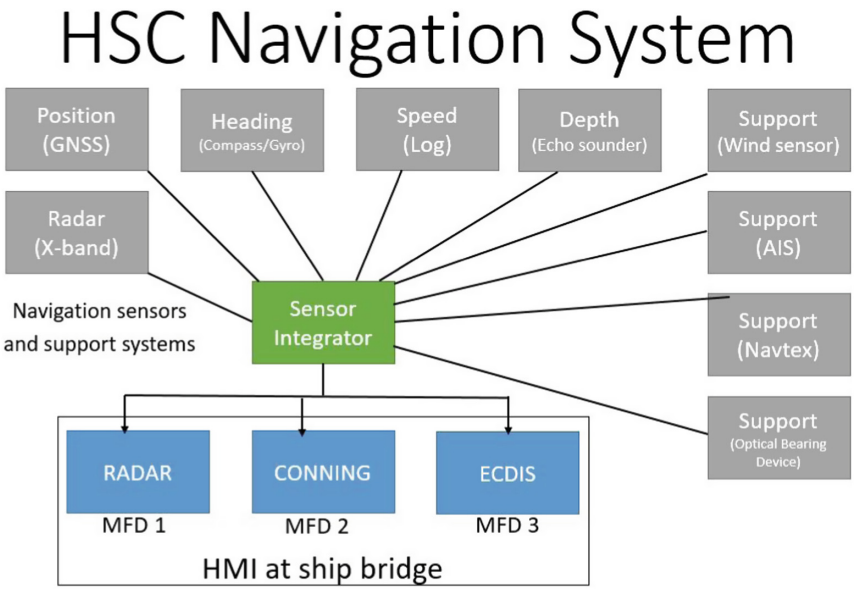

Fig. 4. Example of a HSC navigation system

The integration of navigation sensors in the navigation systems aims to contribute to improved SA for the navigator, and thus support the safe navigation of the vessel [2, 22–24]. This is partly conducted by presenting the near real-time position of the vessel on the ECDIS. The information from the position-, heading-, speed-, depth- and support sensors are integrated and presented on one of the MFDs on the ship bridge. The three main applications available for the navigation team is ECDIS, radar and conning.

The navigation system even on a relative small HSC vessel below 50 m is arguably a complex system in accordance to Redish [25], and there is a concern that the navigator does not hold a sufficient understanding of the navigation system they are

operating [26–28], known as system awareness in Fig. 3. This could lead to misinterpretation of information from the navigation system presented on the MFD.

Signal interference on the signal from a Global Navigation Satellite System (GNSS), intentional or un-intentional, can lead to Hazardous Misleading Information presented to the navigator [29]. There are several examples of jamming and spoofing of GNSS-signals [30–33], and the navigator needs to be aware of the vulnerabilities in the computer system in use [21].

1.3 Eye Tracking

Eye Tracking is the process of measuring the eye activities [34]. This could be performed by measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. An eye tracker is a device that can measure eye position and eye movement. ETGs is constructed in order to study human behaviour in real-world environments [35].

During the past years, eye tracking in Human-Computer Interaction (HCI) and usability studies/research has been more frequently used [36–41]. There has also been research and suggested frameworks for the use of eye tracking measurements when conducting usability evaluation at a ship's bridge [42].

Eye Tracking data from ETGs has been used to improve usability of bridge design [13, 43, 44], and the Graphical User Interface (GUI) and bridge layout of a HSC has been examined with ETGs in an earlier study [14, 15]. ETGs has been used as a tool to measure the efficiency of a navigator when conducting a passage [10], and in maritime bridge simulator assessments [45]. Nielsen and Pernice [40] find that the use of eye tracking data will aid the designers and software developers to better understand what people see and don't see, and ETGs has shown to be a useful tool in a framework to improve SA in demanding maritime operation training [12].

2 Methodology

The work presented in this article builds on earlier studies conducted prior to a mid-life update of the Skjold-class Corvette navigation system [9, 11, 14]. ETGs were utilized to better understand the visual attention of the navigator, in order to identify, and if possible correct, flaws in design and/or GUI. Tobii Pro Glasses 2 was used for the two data collections, and pros and cons with the use and different types of ETGs is laid down in earlier work [11].

2.1 Subjects

The participants were personnel in active service, mean age of 29 years (Standard Deviation (SD): 4 years), and a total of 13 subjects participated in the test conducting 19 runs. It would be beneficial with a higher number of test objects, but the amount of relevant personnel is limited. The RNoN has six Skjold-class in service, with two navigation teams on each vessel, thus 54.2% of available personnel participated in the data collection.

When recruiting personnel to the data collection, several challenges with the availability of relevant personnel was identified. The workload on personnel in active duty is high, and the data collection was not characterized as operational service, and therefore not given a high priority. This lead to challenges with the amount of participants, cancellations and time-constraints when conducting the data collection.

2.2 Apparatus

The data collection was conducted in the Navigation Simulator (NavSim) at the Navigation Competence Center at the Royal Norwegian Naval Academy. Earlier work has argued that the Skjold-class simulator at the NavSim provides eye tracking data with quality equal to live data [9].

The navigation bridge of the Skjold-class is shown in Fig. 5, and to better organize the eye tracking data, Areas of Interest (AOIs) of the bridge was defined. AOIs defines important regions in the visual scene, and further allows events such as dwells, transitions and AOI hits to be defined [35]. The AOIs are shown in Fig. 5, and is in accordance with the visual areas most commonly used by the navigator on board a Skjold-class Corvette.

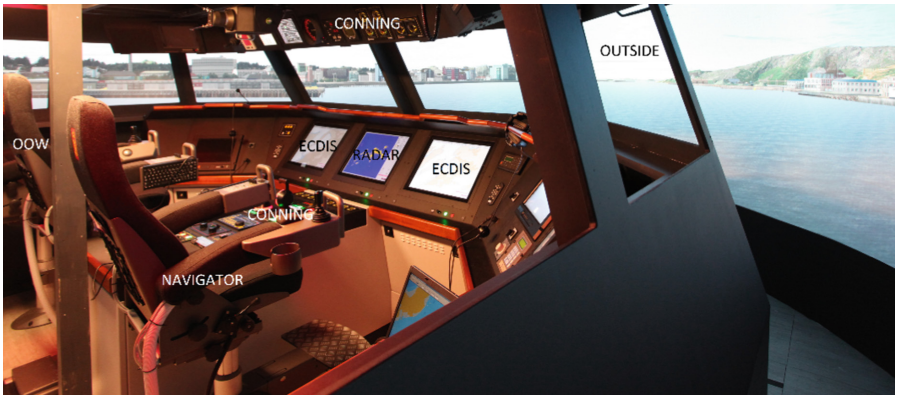

Fig. 5. Skjold-class bridge layout with primary AOIs

The AOIs were defined by using experience from earlier studies, together with a pre-study conducted with three persons in three runs. This resulted in four main AOIs, which are divided into 7 AOIs in total. The AOIs are:

1. Outside (AOI_O): The surroundings of the ships, and are defined by the boundaries of the windows on the ships bridge.
2. ECDIS (AOI_E): The ECDIS information is presented on the MFD in front of the navigator.
   a. AOI_E also consists of the Route Monitor window (AOI_M) as a part of the ECDIS application [15].

3. Radar (AOI_R): The radar information, presented on the centre MFD on the ships bridge in Fig. 5.
   a. AOI_R consist of the heading bearing (AOI_H) in the upper right corner of the radar application [11].
4. Conning (AOI_C): Consisting of information from the displays, consoles and autopilot related to the propulsion and steering of the ship.
   a. AOI_C consist of the consoles for manoeuvring (AOI_{CO}) and the speed log display (AOI_D) [11].
5. White Space (AOI_W): The other areas than those defined by the AOIs [46].
   a. Both data sets white space was marginal, and has been left out of the graphics, which indicates that most fixations were within a defined AOI.
      i. AOI_W pre-study data set: 0.22%
      ii. AOI_W first data set: 0.15%
      iii. AOI_W second data set: 0.26%

The navigations system (Sect. 1.2) consist of AOI ECDIS, Radar and Conning, and the eye tracking data analysis aims to provide a understanding of the use of these AOIs and thus an understanding of the system awareness which contributes to the Navigator's SA (Fig. 3).

2.3 Validation Procedure

The procedure and scenario for the pre- and post- data collection was identical. The scenario was set up in the simulator instructor software Polaris, and used in all the scenarios. The area of data collection is in Norwegian territorial waters, between Bergen and Floroe. The area, traffic, route and environmental conditions are identical in both the data collections throughout the 19 runs. The pre-planned route has a distance of 20.6 NM, and the average sailing time for each participant was 24.8 min (SD: 3.42 min). A total of 6 h and 12.4 min of eye tracking data has been analysed. The experience of the participant averages 1.9 years (SD: 1.75 years). The timeline for the project is shown in Fig. 6.

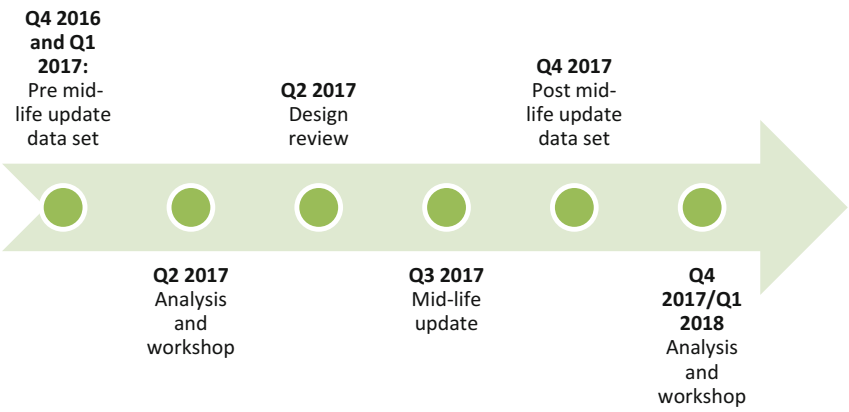

Fig. 6. Timeline process of validating HSC bridge design

The analysis was conducted in the manufacturers software, Tobii Pro Lab. Eye metrics data was captured, and further analysed in Microsoft Excel. In Excel sheets regarding fixations, duration, counts and events was analysed and visualized using diagrams (Figs. 9, 13 and 14). Visualization maps such as heat maps and scan paths were created in Tobii Pro Lab (Figs. 7, 8, 11 and 12). The visualizations maps provide a static overview of the visual attention of the navigator in the given period of time. The process of analysing and interpreting the eye tracking data can be challenging and time consuming, and a rule of thumb is one hour of analysis for every 10 min of eye tracking data.

2.3.1 Statistical Model

The statistical analysis has been conducted in four steps, where the statistical model is established and consist of a normality test, an F-test and a t-test to control if the values disprove the null hypothesis of similarity between the two eye tracking data collections within a significance level of 5%. The F-test is conducted to control the p-value for validation of similarity of the two collected data set. The t-test is conducted to control if the expectations values in the two collected data set are valid.

The generation of the analysis has been conducted in Microsoft Excel, by using the eye metrics data which is generated by the manufacturer software.

2.4 Technical Workshops

To better understand the Eye Tracking data and the analysis of it, workshops with Subject Matter Experts (SMEs) were conducted. This was facilitated through the creation of a Technical Group High Speed Navigation on the manufacturers equipment.

The working group consisted of SMEs, who are active navigators from the high speed navigation community in the RNoN. Representatives from the ECDIS manufacturer contributed together with HCI experts from the RNoN, which is supported by the call for more usability testing in complex systems [25].

The SMEs used the working group as a forum to express their opinions regarding the possibilities and the challenges with the existing navigation system. These opinions were correlated towards the presented eye tracking data and analysis, and discussed in the working group. System Problem Reports (SPR) and Engineering Change Proposals (ECPs) were produced where opinions from the SMEs and eye tracking data correlated. Amongst these were the three design issues described in Sect. 3.3, thus we investigated if eye tracking data collected from ETGs can be used to validate a design-review of a maritime HSC bridge.

The technical group conducted workshops both pre- and post-mid-life upgrade, and the feedback from the post mid-life update was correlated with the eye tracking data. The SMEs response to the revised design of the three main design issues was positive.

3 Results

3.1 Pre Mid-Life Update Data Set

The first data set consists of data from 10 participants, nine males and one female. Average age of participants 29 years (SD: 4 years). Average experience 1.6 years (SD: 1.6 years). The average time for conducting the passage was 24.5 min (SD: 3.9 min.

The first data set identified three main design issues, supported by earlier work [11]:

1. Poor availability of the presentation of heading bearing in radar GUI.
2. Challenges with the HCI with the distance measurement unit (Electromagnetic Log – speed log).
3. Sub-optimal GUI in route monitor window.

It is important to understand where the visual attention of the navigator is allocated during a passage. The visualization maps in the first data set is shown in Figs. 7 and 8.

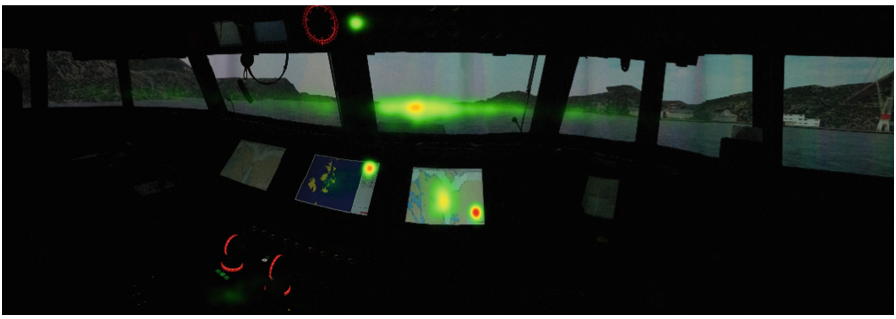

Fig. 7. Heat map from pre mid-life update data set

The heat map identifies the hot spots where the navigator addresses its' attention, and the three design issues is identified. Number 1 in the top right corner of the radar (centre MFD), number 2 in the top centre of the figure, where the speed log is placed. Design issue number 3 is the route monitor window in the lower right corner of the ECDIS GUI on the right side MFD (reference to Fig. 5).

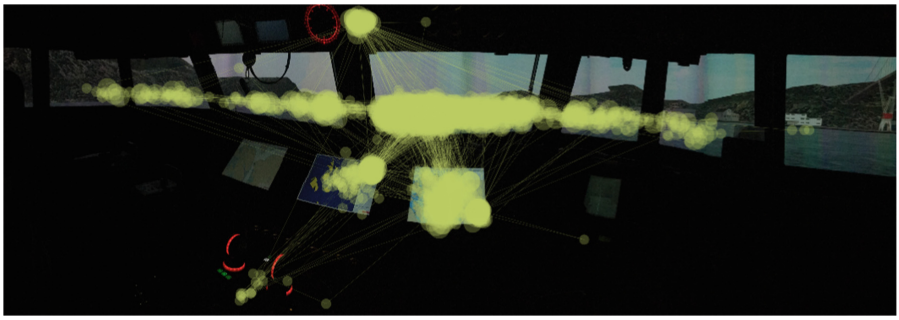

Fig. 8. Scan path from pre mid-life update data set

Analysing the scan path from the first data set, the three design issues are evident. Each fixation is represented by a circle, and the size of the circle represents the fixation time (larger circle, longer fixation).

The total time spent in an AOI can be an indication of the importance of the AOI. It could also indicate a design issue or high mental workload [35], and thus contribute to a decrease in SA for the navigator [40]. The total time spent in the AOI in the first data set is shown in Fig. 9.

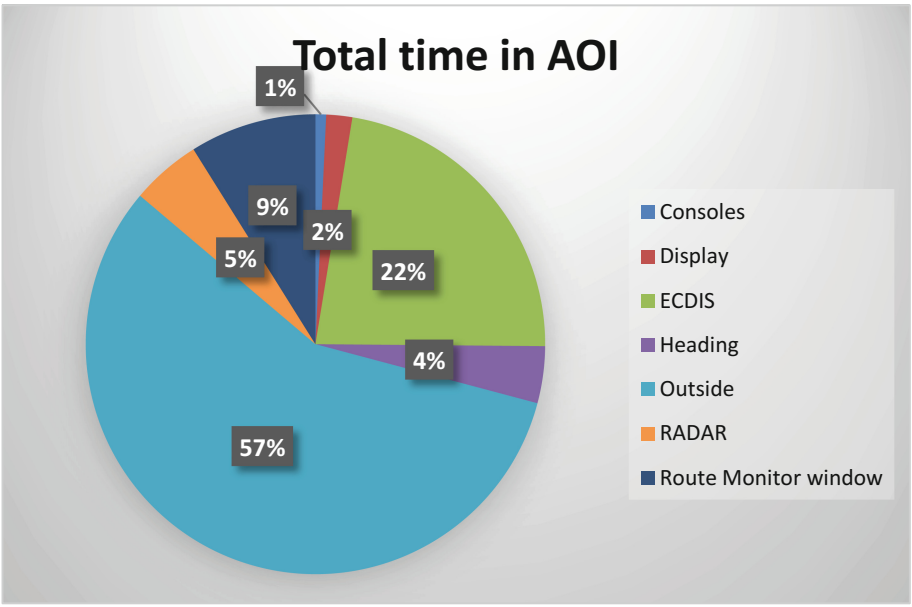

Fig. 9. Total time in AOI from first data set (pre mid-life update)

The pie chart provides valuable insight in the visual attention of the navigator [1], and the main objective is to provide more time for the navigator to control the surroundings to facilitate a higher SA (Fig. 3 – Spatial and Task Awareness). A suggestion of an optimal visual attention to AOI Outside is 80% in good visual condition conducting the passage in visual sailing mode [1], in order to support the navigators SA. The SD in AOI outside in the pre mid-life update data set is 8.3%.

3.2 Mid-Life Update Navigation System Skjold-Class Corvette

The three design issues were addressed during a design-review and mid-life update of the navigation system on board the Skjold-class Corvettes. The SPRs were discussed in the working group, and ECP developed for each of the design issues.

ECP for design issue 1 was moving the presentation from the top right corner of the radar GUI to a larger presentation in a new High Speed Craft Route Monitor (HSCRM) window. The final version of the HSCRM window is shown in Fig. 10, and the

heading bearing is presented with large fonts in the upper left corner of the GUI (#1). The HSCRM window is to be placed in the centre-top left corner of the ECDIS application, this in order to have a short visual passage from the display to the outside (surroundings) of the vessel, and contribute to a higher degree of SA by supporting the spatial, task and system awareness [15].

Fig. 10. HSCRM window from design review [15]

ECP for design issue 2 suggest moving the reset button for the trip meter from the overhead panel of the speed log [14], to the arm rest panel located on the left armrest of the navigator's chair (reference to Fig. 5). This implies the physical movement of the reset button from the speed log panel to within arm's reach of the left hand of the navigator. The display of the trip meter is co-located with other relevant information in the HSCRM window, and is shown on the top second line in Fig. 10 (#2). This makes the speed log display excessive, and the navigator only needs to address the HSCRM window.

ECP for design issue 3, a new route monitor window design, is shown in Fig. 10 and has been elaborated in earlier work [15]. The aim of this change was to sort and present the information needed for the navigator to maintain a high degree of SA. The presentation of this information is in line with the standard operating procedures on HSC in the RNoN [47]. A challenge identified in the workshops is that the HSCRM window will probably lay hold of relative more time from the navigator's visual attention, due to the relative large amount of information co-located in this GUI.

3.3 Post Mid-Life Update Data Set, Validating Design Updates and Measuring Impact

The second data set consists of six participants, all male. Average age of participants 29 years (SD: 4 years). Average experience 2.3 years (SD: 1.8 years). The average time for conducting the passage was 25.3 min (SD: 1.9 min).

The purpose of the design review was to free time for the navigator to control the surroundings of the vessel (AOI Outside), and contribute to a better SA for the HSC navigator.

In order to evaluate the end-state, a final eye tracking data set was collected (Fig. 6). Figures 11 and 12 shows the visualization maps for the validation data set.

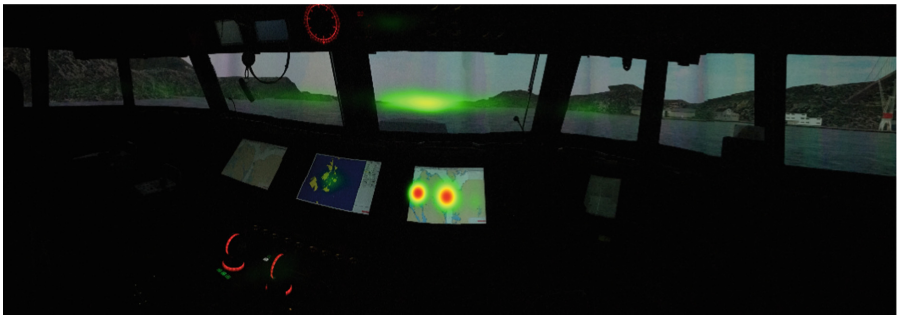

Fig. 11. Heat map second data set (validation)

When comparing the heat maps from the two data sets (Figs. 7 and 11), the heat map clearly identifies the three design flaws in Fig. 7, while these three areas are not present in Fig. 11. According to the heat map, more of the attention has been addressed to the ECDIS, Outside, Route monitor window and to the centre of the MFD with the radar application. There are fewer AOIs for the navigator to direct the visual attention towards, since AOI Heading, AOI Display and AOI Consoles is marginalized. This should in turn contribute to freeing time for the navigator to focus in more important AOIs, and contribute to increase the SA of the navigator. The eye tracking data visualization clearly indicates fewer AOIs in the new bridge design and GUI, more visual attention directed towards operational important information in AOI Outside, ECDIS and radar, which should contribute to safer operation.

Comparing the scan paths from the two data sets (Figs. 8 and 12), the second data set (Fig. 12) indicates a tidier scanning pattern, where fewer AOIs are visited. As shown with the heat map, less important AOIs such as AOI Heading, AOI Display and AOI Consoles are marginalized. This should contribute to a more efficient visual search for the navigator, and thus supporting an increase in the SA of the navigator. This finding supports the suggested Scan Pattern for the Maritime Navigator [1], which aims to streamline and optimize the visual search for the navigator. Note that the heat map in Fig. 11 shows inferior resolution inside the AOIs, compared to the scan path in Fig. 12. As an example the amount and placement of fixations inside AOI ECDIS becomes

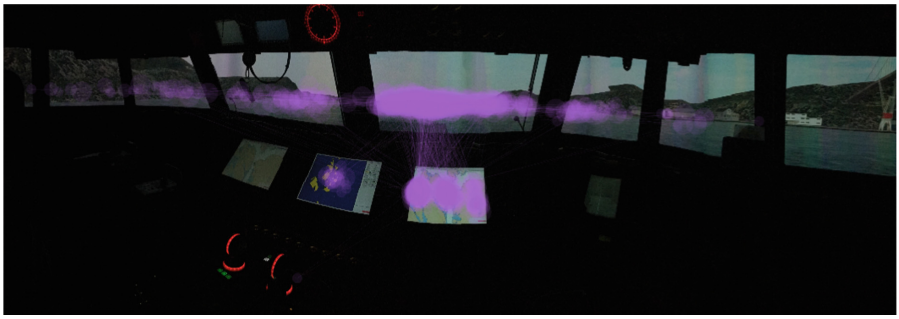

Fig. 12. Scan path second data set (validation)

more distinct in the scan path, than in the heat map. When analysing the heat map, be aware of the strength of the colour coding in the generation of the heat map can be adjusted in the manufacturer software [48], and is not uniquely. The increased resolution of the amount and placement of fixations in the scan path visualization, will support a better understanding of the eye tracking data.

The analysis of the eye tracking data from the post mid-life data set indicates that several of the AOIs have been marginalized in the mid-life update, shown in Fig. 13. AOI Console, Display, Heading and Radar has less than 1.5% of the total time. Since this was a passage conducted in daylight, it would be reasonable to suggest a vigorous increase in the attention to AOI radar during hours with reduced visibility or darkness. The total time in AOI for the second data set indicates an increase in the time spent addressing the ECDIS, and a retrogression in the accumulated visual attention in AOI Outside. One of the main objectives for the design review was to transfer more of the visual attention of the navigator to the actual surroundings of the vessel (AOI Outside).

4 Discussion

By comparing and analysing the visualization maps (Figs. 7, 8, 11 and 12), one could argue that the design changes conducted in the mid-life update has contributed to fewer areas for the navigator to focus on. Comparing the heat map (Figs. 7 and 11) indicates that the overhead displays, consoles and upper right corner in the radar (heading bearing) is removed as areas where the navigator focusses its' visual attention. Attention to these areas were identified as design flaws in the pre mid-life data set. The post mid-life update heat map (Fig. 11) indicates more visual attention to AOI ECDIS, and clearly indicates increased visual attention to the new HSCRM window located in the centre-left part of the AOI ECDIS as expected. The heat map also suggests more visual attention to the centre part of AOI radar, which shows an increased awareness from the navigator towards the operational valuable information provided from the radar (Fig. 3 – System awareness). By addressing attention to the centre of the radar, the navigator interprets the radar picture and evaluates and compares the surroundings of the vessel with a terrestrial mean. This will contribute to a higher degree of SA for the navigator, and thus supporting safe operation of the vessel.

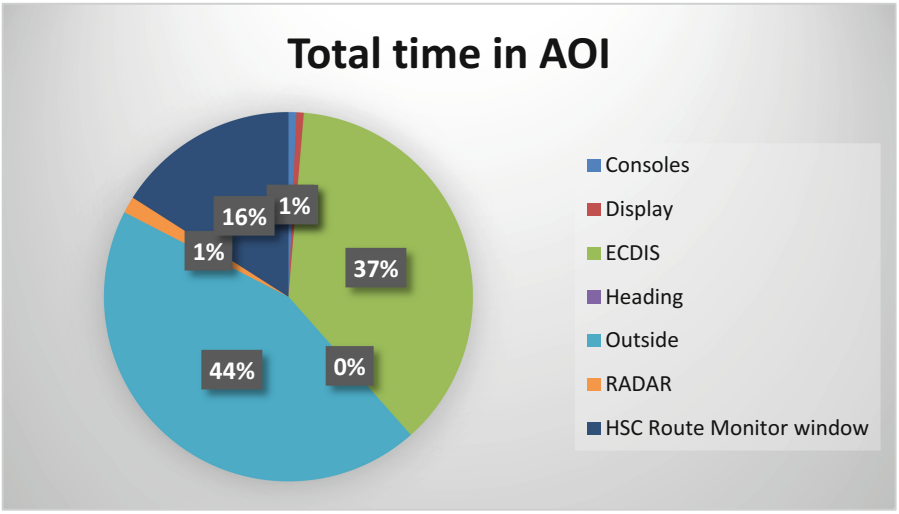

Fig. 13. Total average time in AOI post mid-life update data set

Analysing the scan path (Figs. 7 and 11), indicates a neater scan pattern for the navigator. The post mid-life update data set holds less scanning clutter, and this could contribute to a more efficient and less time consuming visual scan pattern for the navigator [1].

Total time in AOI (Figs. 9 and 13) shows an undesirable increase in the visual attention towards AOI ECDIS, and a decrease in the attention in AOI Outside. One could argue that an increase in attention towards AOI ECDIS will support increased SA (Fig. 3 - Task and System awareness) for the navigator as long as the chart is in focus, but the solution of the eye tracking data is not good enough to support the assumption that the visual attention is allocated to the chart alone. The design revisions aim was to support more attention towards AOI Outside, and the post mid-life update data set indicates the contrary. To better understand this finding, each of the participant's data set has been analysed, and there are discrepancies in the visual attention which are ambiguous and challenging to analyse.

When analysing the data individually, the difference from participant to participants becomes clear. If introducing experience and familiarization with the new design and software as a variable, it is a clear indication that the amount of time spent in AOI Outside is dependent on experience and familiarization. This is shown in Table 1.

When analysing the data from the participants in Table 1, the values gives an indication that the experience and familiarization time with the new SW, installed during the mid-life update, is a variable affecting the visual attention of the navigator. This is shown with participant 5 s time in AOI in Fig. 14, showing a high amount of attention towards AOI Outside, which is opposed to the accumulated visual attention in Fig. 13.

Table 1. Relation between experience and total time in AOI.

Participant	Total time in AOI outside (%)	Years of experience (years)	Time with new SW (months)
#1	43.52	1	2
#2	27.17	0.5	0
#3	46.24	4.5	0
#4	45.24	2.5	1
#5	68.49	5.5	4
#6	39.16	0.5	3

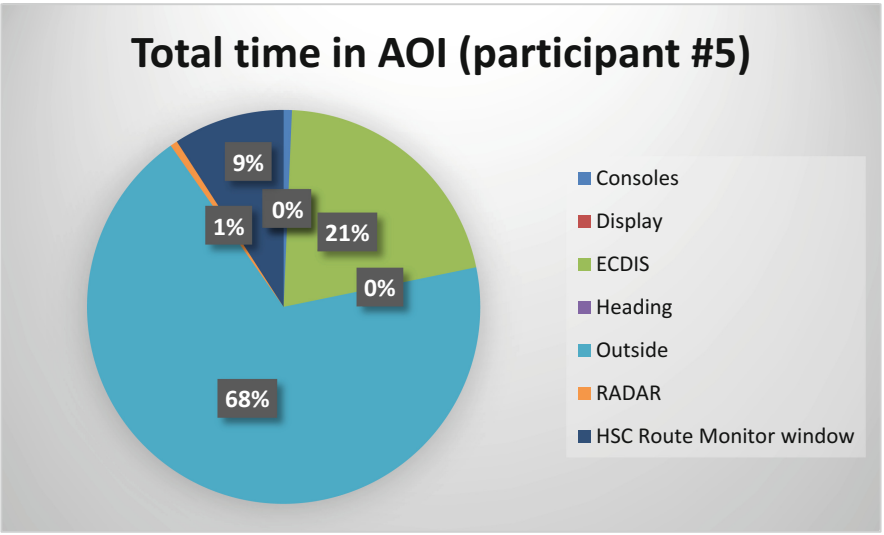

Fig. 14. Total time in AOI post mid-life update data set for participant #5

To better understand and analyse this finding, the most and least experienced participant of the subjects who participated in both data collections where analysed. This was the same persons in both data sets, and is shown in Tables 2 and 3.

Table 2. Comparison of the most experienced participant in the two data sets.

Data set/AOI (%)	Outside	ECDIS	Route monitor	Radar	Conning
Pre mid-life update	70%	14%	8%	3%	5%
Post mid-life update	68%	21%	9%	1%	1%

Table 3. Comparison of the least experienced participant in the two data sets.

Data set/AOI (%)	Outside	ECDIS	Route monitor	Radar	Conning
Pre mid-life update	47%	29%	10%	5%	9%
Post mid-life update	27%	49%	21%	1%	2%

Analysing Tables 2 and 3 can explain the deviation in the expected increase towards AOI Outside. Both experience and familiarization with equipment is known to be important variables when utilizing the navigation system [49]. Table 2 show how an experienced navigator with 4 months of familiarization on the new GUI shows good progression in utilizing the visual scan and direct the attention towards AOI Outside. Tables 3 indicates how an inexperienced navigator has challenges with operating unknown software, and must thus direct more attention towards the new design (the ECDIS and HSCRM window). Glover [50] presents the planning-control theory in visual representation, where he argues that human action is directed by a control system, while the perception is commanded by a planning system. This implies that a human (the navigator) take account for a wide variety of visual and cognitive information when conducting the planning of an action. This information is further integrated with memories of past experience, which could explain why experience is an important factor when using a system. This provides a link to how experience contributes to the navigator's SA.

Table 4 shows the higher SD in the post mid-life update. The SD could be a measure of the familiarity with the software and GUI. This is analysed as an indication of a higher familiarization with the software and GUI used in the pre mid-life update. All participants were familiar with the GUI in the pre mid-life update, since it had been in use for several years.

Table 4. Comparison of standard deviation in the two data sets

Data set/measure	Standard deviation in AOI outside
Pre mid-life update	8.3%
Post mid-life update	12.3%
Increase in % between pre- and post-mid-life update measures	48.2%

The importance of familiarization and experience is supported by earlier studies with eye tracking, and the findings in this study indicates the importance of both experience and familiarity with new software and design as factors [51–53]. It also indicates an important finding concerning operational use after post mid-life updates, which indicates that the low level of experience and low level of familiarization with new software decreases the visual attention towards AOI Outside. This could in turn contribute to a decrease in the SA of the navigator, and thus in the degree of safe operation. The importance of familiarization is thus supported and outlined by the findings in the two data sets [49].

The design of the method will contribute to less uncertainty when analysing pre- and post-mid-life updates of design. The pre mid-life update data set consist of 10 recordings and participants, while the post mid-life update data set consists of six recordings and participants. Five of the participants attended both data collections, and the two data sets where identical in conduct but not with regards to attendance of participants. With an increased amount and same number of participants in both data

sets, the analysis will be less ambiguous. It would strengthen the data set if the same participants took part in both data collections, and the design of the two data sets should be identical to avoid sub-optimal analysis of data sets. The findings in the data set does not support the hypothesis that the two data sets are similar within a statistical significant level of 5%, partly due to the low number of participant (F-value 0.45, p-value = 0.14). To achieve a p-value of less than 5%, with the assumptions of the same values as in the current data set, the amount of participants must be almost four times higher. This would be very difficult and time consuming to achieve in an operational environment with personnel in active duty.

Collecting eye tracking data in an operational environment, such as the bridge simulator, is challenging [54], and the ETGs and the manufacturer software is not mature to meet the demands of the operational environment in this study. It is also evident that data collection with personnel in active duty is challenging and changes in plan on short notice must be expected. Research will not supersede operational demand.

When comparing the analysis of the eye tracking data with the information collected from the SMEs in the working group, there are sufficient indications of an improvement in the mid-life update bridge design to support a higher degree of SA for the navigator. The qualitative measurements from the workshops is emphasised as an important support for the quantitative measurements, due to the ambiguities in the eye tracking data due to immature technology (ETG robustness and manufacturer software) and sub-optimal method design.

5 Conclusion

ETGs has shown a potential to support identifying design and GUI challenges that contributes to a decrease in the SA of the navigator, and in validation of design changes of ship bridges. This study shows that the quantitative data needs support from qualitative data to be unambiguous. The use of eye tracking data such as visualization maps provides a simple and intuitive measure for identifying changes in visual search pattern after a design alteration, but the process of analysing the data is time consuming. The eye tracking data is useful as a basis for the design-review, and as evidence and support for the discussions and conclusions in the technical working group. However, eye tracking technology used to collect data in an operational environment with ETGs, is in this work assessed to be immature.

The collected data set shows the uncertainties related to eye tracking data when the amount of participants is relatively low, and the challenges concerned with few possible participants when conducting studies in a narrow domain.

The importance of experience and familiarization with new design is salient, and this study shows that the participants must be given ample time to familiarize themselves with the new design and software to conduct a better and less unambiguous analysis of the eye tracking data. This finding is also important for the operational domain, concerning familiarisation with new equipment before operational use.

The method and procedure when conducting the data collection are imperative with regards to the quality of the data collected. The cost and effort of collecting an eye

tracking data set in an evaluation of a bridge design or software GUI, must be weighed towards the benefits, and the technology is at this time argued to be immature to collect eye tracking data from an operational environment.

If conducting maritime usability studies with data collected by ETGs, it is recommended to support the quantitative measurements with qualitative data for correlation and less ambiguity.

5.1 Further Work

Collect a new post mid-life update data set with optimal method design, in order to control the main objective of increased time in AOI Outside.

References

1. Hareide, O.S., Ostnes, R.: Scan pattern for the maritime navigator. TransNav **11**(1), 39–47 (2017)
2. IMO. Resolution MSC.252(83): Adoption of the Revised Performance Standard for Integrated Navigation Systems, London, p. 49 (2007). http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Maritime-Safety-Committee-(MSC)/Documents/MSC.252(83).pdf
3. Gould, K., Røed, B.K., Koefoed, V.F., Bridger, R.S., Moen, B.E.: Performance-shaping factors associated with navigation accidents in the Royal Norwegian Navy. Mil. Psychol. **18**, S111–S129 (2006)
4. Gould, K., Røed, B.K., Saus, E.-R., Koefoed, V.F., Bridger, R.S., Moen, B.E.: Effects of navigation method on workload and performance in simulated high-speed ship navigation. Appl. Ergon. **40**(1), 103–114 (2008)
5. Van Orden, K.F., Limbert, W., Makeig, S., Jung, T.-P.: Eye activity correlates of workload during a visuospatial memory task. Hum. Factors: J. Hum. Factors Ergon. Soc. **43**(1), 111–121 (2001)
6. Wickens, C.D.: Situation awareness and workload in aviation. Curr. Dir. Psychol. Sci. **11**(4), 128–133 (2002)
7. Øvergård, K.I., Bjørkli, C.A., Røed, B.K., Hoff, T.: Control strategies used by experienced marine navigators: observation of verbal conversations during navigation training. Cogn. Technol. Work **12**(3), 163–179 (2010)
8. Røed, B.K.: Designing for High-Speed Ships. Norwegian University of Science and Technology, Trondheim (2007)
9. Hareide, O.S., Ostnes, R.: Comparative study of the Skjold-class bridge- and simulator navigation training. Eur. J. Navig. **14**(4), 57 (2016)
10. Forsman, F., Sjörs-Dahlman, A., Dahlman, J., Falkmer, T., Lee, H.C.: Eye tracking during high speed navigation at sea. J. Transp. Technol. **2**, 277–283 (2012)
11. Hareide, O.S., Ostnes, R.: Maritime usability study by analysing eye tracking data. J. Navig. **70**(5), 927–943 (2017)
12. Sanfilippo, F.: A multi-sensor fusion framework for improving situational awareness in demanding maritime training. Reliab. Eng. Syst. Saf. **161**, 12–24 (2017)

13. Bjørneseth, F.B., Renganayagalu, S.K., Dunlop, M.D., Homecker, E., Komandur, S. (eds.): Towards an experimental design framework for evaluation of dynamic workload and situational awareness in safety critical maritime settings. In: Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers. British Computer Society (2012)
14. Hareide, O.S., Ostnes, R., Mjelde, F.V. (eds.): Understanding the eye of the navigator. In: European Navigation Conference. Confedent International, Helsinki (2016)
15. Hareide, O.S., Mjelde, F.V., Glomsvoll, O., Ostnes, R.: Developing a high-speed craft route monitor window. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10285, pp. 461–473. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_33
16. Gould, K.: Faster, better, safer?: studies of safety, workload and performance in naval high-speed ship navigation. Universitetet i Bergen (2009)
17. Bjørkli, C.A., Øvergård, K.I., Røed, B.K., Hoff, T.: Control situations in high-speed craft operation. Cogn. Technol. Work **9**(2), 67–80 (2007)
18. da Conceição, V.P., Dahlman, J., Navarro, A. (eds.): What is maritime navigation? Unfolding the complexity of a sociotechnical system. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, Los Angeles (2017)
19. Dobbins, T., Hill, J., Brand, T., Thompson, T., McCartan, S.: Standardised information architecture to support the Dynamic Navigation (DYNAV) standard operating procedure. In: The Royal Institution of Naval Architects 2016 (Human Factors Conference), p. 7 (2016)
20. Forsman, F., Dahlman, J., Dobbins, T. (eds.): Developing a standard methodology for dynamic navigation in the littoral environment. In: Royal Institute of Naval Architects, International Conference, Human Factors in Ship Design and Operation (2011)
21. Hareide, O.S., Jøsok, Ø., Lund, M.S., Ostnes, R., Heikala, K.: Enhancing navigator competence by demonstrating maritime cyber security. J. Navig. **71**(5) (2018)
22. Norris, A.: ECDIS and Positioning. Nautical Institute, London (2010)
23. Weintrit, A.: The Electronic Chart Display and Information System (ECDIS), An Operational Handbook: A Balkema Book. CRC Press, Taylor & Francis Group, Boca Raton (2009)
24. Thornton, P.: The ECDIS Manual. ECDIS Ltd. Editors. Witherby Publishing Group Ltd., Glasgow, 443 p. (2012)
25. Redish, J.G.: Expanding usability testing to evaluate complex systems. J. Usability Stud. **2** (3), 102–111 (2007)
26. Nyhamn, S.: How are Radar and AIS Utilised in Anti-collision on Modern Integrated Bridge Systems (IBS) in the RNoN, Within Norwegian Littoral Waters?. University of Nottingham, Nottingham (2013)
27. Hareide, O.S.: Control of ECDIS (electronic charts and display information system) on high speed crafts in littoral waters [M.Sc.]. University of Nottingham (2013)
28. Fagerholt, R.A., Kongsvik, T., Moe, H.K., Solem, A.: Broutforming på hurtigbåter. Kartlegging av problemer med utforming og funksjonalitet på teknisk utstyr på hurtigbåt-bro. Rapport. NTNU Samfunnsforskning AS (2014)
29. Last, D., Grant, A., Ward, N. (eds.): Demonstrating the effects of GPS jamming on marine navigation. In: 3rd GNSS Vulnerabilities and Solutions Conference, Croatia (2010)
30. Glomsvoll, O., Bonenberg, L.K.: GNSS jamming resilience for close to shore navigation in the Northern Sea. J. Navig. **70**(1), 33–48 (2017)
31. Grant, A., Williams, P., Ward, N., Basker, S.: GPS jamming and the impact on maritime navigation. J. Navig. **62**(2), 173–187 (2009)

32. Humphreys, T.E., Ledvina, B.M., Psiaki, M.L., O'Hanlon, B.W., Kintner Jr., P.M. (eds.): Assessing the spoofing threat: development of a portable GPS civilian spoofer. In: Proceedings of the ION GNSS International Technical Meeting of the Satellite Division (2008)
33. Bhatti, J., Humphreys, T.E.: Covert control of surface vessels via counterfeit civil GPS signals. University of Texas (2014, unpublished)
34. Duchowski, A.T.: Eye Tracking Methodology. Theory and Practice. Springer, London (2007). https://doi.org/10.1007/978-1-84628-609-4. 328 p.
35. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye Tracking: A Comprehensive Guide to Methods and Measures. OUP Oxford, Oxford (2011)
36. Jacob, R., Karn, K.S.: Eye tracking in human-computer interaction and usability research: ready to deliver the promises. Mind **2**(3), 4 (2003)
37. Bergstrom, J.R., Schall, A.: Eye Tracking in User Experience Design. Elsevier, New York (2014)
38. Nielsen, J., Pernice, K.: Eyetracking Web Usability. New Riders, Indianapolis (2010)
39. Groen, M., Noyes, J.: Using eye tracking to evaluate usability of user interfaces: is it warranted? IFAC Proc. Vol. **43**(13), 489–493 (2010)
40. Pernice, K., Nielsen, J.: How to Conduct Eyetracking Studies. Nielsen Norman Group, Fremont (2009)
41. Renshaw, J., Finlay, J., Tyfa, D., Ward, R.D.: Designing for visual influence: an eye tracking study of the usability of graphical management information. Hum.-Comput. Interact. **1**, 144–151 (2003)
42. Papachristos, D., Koutsabasis, P., Nikitakos, N. (eds.): Usability evaluation at the ship's bridge: a multi-method approach. In: 4th International Symposium on Ship Operations, Management and Economics (2012)
43. Bjørneseth, F.B., Clarke, L., Dunlop, M., Komandur, S. (eds.): Towards an understanding of operator focus using eye-tracking in safety-critical maritime settings. In: International Conference on Human Factors in Ship Design & Operation (2014)
44. Hareide, O.S., Ostnes, R.: Maritime usability study by analysing eye tracking data. In: International Navigation Conference Proceedings, p. 17 (2016)
45. Muczyński, B., Gucma, M., Bilewski, M., Zalewski, P.: Using eye tracking data for evaluation and improvement of training process on ship's navigational bridge simulator. Zeszyty Naukowe/Akademia Morska w Szczecinie. **33**(105), 75–78 (2013)
46. Bojko, A.: Eye Tracking the User Experience: A Practical Guide to Research. Rosenfeld Media, New York (2013)
47. RNoN. SNP 500. In: Centre NC, Editor. Royal Norwegian Naval Academy, Bergen (2018)
48. Tobii. Tobii Pro Lab User Manual. Internet, 1.79 (2017)
49. IMO. ECDIS - guidance for good practice. In: Committee MS, Editor. IMO, London, p. 25 (2017)
50. Glover, S.: Separate visual representations in the planning and control of action. Behav. Brain Sci. **27**(1), 3–24 (2004)
51. Kovesdi, C., Spielman, Z., LeBlanc, K., Rice, B.: Application of eye tracking for measurement and evaluation in human factors studies in control room modernization. Idaho National Laboratory (INL), Idaho Falls, ID (United States) (2017)
52. Reinerman-Jones, L., Matthews, G., Wohleber, R., Ortiz, E. (eds.): Scenarios using situation awareness in a simulation environment for eliciting insider threat behavior. In: 2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA). IEEE (2017)
53. Zheng, S.: Impact of eye-trackers on maritime trainer-trainee experience. Master thesis, NTNU Aalesund (2014). https://brage.bibsys.no/xmlui/handle/11250/274107
54. Lappi, O.: Eye tracking in the wild: the good, the bad and the ugly. J. Eye Mov. Res. **8**(5):1, 1–21 (2015)

The Wide Area Virtual Environment: A New Paradigm for Medical Team Training

Alan Liu(✉), Eric Acosta, Jamie Cope, Valerie Henry, Fernando Reyes, Joseph Bradascio, and Wesley Meek

Val G. Hemming Simulation Center,
Uniformed Services University of the Health Sciences,
Bethesda, MD 20814, USA
alan.liu@simcen.usuhs.edu
https://simcen.usuhs.edu

Abstract. Medical simulation can provide safe, consistent, and repeatable learning opportunities to learners without risk to patient safety. There is an increasing awareness of the value that simulation brings to learning. The Wide Area Virtual Environment (WAVE) is an 8,000 sq. ft. facility designed to provide an immersive virtual environment for medical team training. It combines computer generated 3D stereoscopic images, standardized patients, human patient simulators and theatrical effects to render training scenarios with an unprecedented level of realism. The WAVE represents a novel application of human-computer interaction. It forms the basis for a synergistic amalgamation of live, virtual, and constructive simulation for medical instruction. In this paper, we discuss the design, construction, and use of the WAVE.

Keywords: Immersive virtual environment · Medical team training
Mixed modality

1 Introduction

Medical simulation modalities can be broadly divided into three categories: standardized patients, human patient simulators, and virtual-reality trainers. Standardized patients are individuals trained to play the role of a patient in a medical scenario. Human patient simulators are computerized mannequins capable of modeling and presenting physiology to learners. Virtual reality trainers present computer generated scenarios designed to develop cognitive as well as dexterous skills. Relatively little has been done to combine all three modalities.

The Wide Area Virtual Environment (WAVE) is an 8,000 sq. ft. facility designed to provide an immersive virtual environment for medical team training. It integrates all three modalities. The WAVE is capable of rendering highly realistic training scenarios that challenges a team's ability to provide care under difficult conditions. The WAVE represents a novel application of human-computer

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 293–304, 2018.
https://doi.org/10.1007/978-3-319-91467-1_23

interaction. It forms the basis for a synergistic amalgamation of live, virtual, and constructive simulation for medical instruction.

This paper will briefly explore the three modalities of simulation. It will describe the design and functionality of the WAVE. This paper will also describe how the three modalities are combined in the WAVE environment to present a unique learning environment. Finally, our experience in using this capability, and implications for the future of medical simulation will be discussed.

2 Background

The practice of medicine requires the application of knowledge, dexterous skills, and experience. Knowledge can be acquired through classroom learning. However, skills acquisition and experience require practice. These skills may be acquired through patient interactions. The notion of 'See one, do one, teach one' was widely accepted [1–3]. Learning on patients is not without risk. To ameliorate harm, cadavers and live animals substituted for patients. These methods have their disadvantages. Cadavers do not model a live patient's physiology. Animal anatomy can be poor substitutes for human anatomy. There are also ethical concerns over the use of live animals for teaching. Medical simulation seeks to address these limitations. It provides a consistent, repeatable environment for learning dexterous skills. Medical simulation also helps in developing experience.

Modern medical simulation modalities fall into three categories: standardized patients, part task trainers, and virtual simulations.

Standardized Patients (SPs) are individuals trained to portray patients in a learning scenario [4]. SPs maintain consistent portrayal of an individual with a medical condition. SPs can be used in clinical settings such as a doctor's examination room. They can also be used in unconventional scenarios, such as a mass casualty or natural disaster event. Generally, SPs are used when human interaction is part of the learning objective. They are also used for practicing non-invasive medical skills.

Part task trainers are designed to facilitate the practice of specific medical or surgical skills. Examples include: surgical airway, heart catheterization, and venipuncture. Part task trainers generally involve procedures that are invasive. E.g., they involve the puncturing the skin, cutting, or the insertion of medical devices within the body. Due to the specific focus, part task trainers may not replicate full human anatomy. For example, venipuncture trainers may replicate just the arm. Since they focus only on specific tasks, they are not well suited in scenarios where multiple skills need to be integrated with cognitive assessment to render lifesaving aid to the victim.

The Human Patient Simulator (HPS) addresses this limitation. A HPS is a computer-controlled mannequin. It incorporates mechanical devices within the mannequin to simulate physiological activity. For example, learners can feel a pulse, detect breathing through airflow and chest movement, and can observe pupillary response. Some procedures normally performed on part task trainers can also be performed on an HPS. E.g., chest tube insertion and cricothyroidotomy. HPS also incorporate human physiological models. They can respond to

treatment by altering heart-rate and breathing. Ruggedized HPS can be used in field conditions to simulate casualties.

Despite their utility, HPS have their limitations too. Commercially available HPS are not designed to operate autonomously. Each HPS requires human oversight to ensure specified learning objectives are being addressed. As with part task trainers, HPS require a supply of consumables. Skin patches and body inserts are used up as learners practice invasive procedures. They must be replaced prior to use by another learner.

In contrast to SPs and HPS, virtual reality trainers do not rely on a physical representation of the patient. Instead, a computer generated analogue is employed. Being virtual, they can be deployed on both mobile and desktop environments. They expand learning opportunities available to the student. Virtual reality trainers have capabilities that were traditionally the domain of SPs or HPS. For example, virtual standardized patient simulators [5] have been developed. They incorporate speech recognition and natural language processing capabilities. Virtual SPs have been used in limited settings for medical student education [6].

Virtual simulators have also been used for dexterous skills training. The Haptic Workbench [7] incorporates a 3D stereoscopic display and Phantom haptic interface devices [8] co-located within the same virtual space. Within this space, learners can see and touch computer-generated 3D objects. Medical simulators developed using this approach include cricothyroidotomy [9] and craniotomy [10].

Being fully virtual, these simulators generally do not require consumables. A greater range of invasive procedures can be simulated. Virtual simulators can objectively track learner performance. Despite these advantages, virtual simulators have not replaced the other modalities. They cannot fully replace SPs. Virtual simulators still cannot simulate human interaction at the level SPs can. At the same time, tactile feedback is still limited compared to part task trainers.

Few attempts have been made to combine modalities and overcome their limitations. [11] describes using an HPS within a CAVE environment. The objective is to use the CAVE to simulate an operating room to the learner, who is practicing on the HPS. The virtual environment was passive, and did not interact with the learner. In contrast, [12] incorporated a virtual avatars capable of providing feedback to the learner during the procedure.

These experiments demonstrated the utility of combining disparate modalities. They overcome limitations inherent with one approach by leveraging the strengths of a different approach. While promising, previous attempts have been limited in scope. They focus on a single learner.

In the next section we describe the WAVE, a large scale facility integrating SPs, part task trainers, and interactive virtual reality for medical team training.

3 Methods

The Wide Area Virtual Environment (WAVE) is an immersive virtual reality theater for medical team training. The WAVE seeks to combine modalities. The

objective is to use the strengths of one modality to overcome limitations found in others. The WAVE is designed to support scenarios involving small medical teams over a period of up to four days. In this section, we describe the operation and design of the WAVE.

3.1 Layout

The WAVE is comprised of screens arranged to form two circular pods connected by a corridor. Each pod is approximately 25 ft. in diameter. The corridor is 20 ft. long. Each end of the corridor is 12 ft. wide, tapering to 9 ft. in the middle. At the middle, an electrically operated curtain is positioned. It can be raised or lowered remotely based on scenario requirements. The total useable area of the WAVE is approximately 1,100 sq. ft.

Within the WAVE, learners wear lightweight stereoscopic glasses. They perceive an immersive 3D virtual environment while moving about freely. During a training exercise, this virtual environment is augmented with live actors, human patient simulators and props to simulate an actual environment. Depending on course requirements, this environment may represent combat, humanitarian, or civilian scenarios.

The WAVE is accessed via entrances to each pod. In each pod, two screens are mounted on wheels and hinged so that they can be swung open or closed. The screen doors are balanced to facilitate ease of operation. Learners enter via this entrance. Once inside, the doors can be closed and illuminated to form a seamless environment. Figure 1 illustrates.

3.2 Concept of Operation

The WAVE is designed to support training activities of up to four days in duration. During this time, the nature and scope of the environment can change based as the exercise progresses. The WAVE accomplishes this by changing both the virtual and physical environment during a training event. This is done by using pods alternately. A hypothetical scenario is described.

In this training activity, learners enter Pod A to rescue soldiers wounded by an Improvised Explosive Device (IED) attack. The wounded may be portrayed by SPs or HPS, depending on training objectives. As patients are being treated, the team encounter hostile fire and respond by returning fire while the medic performs first aid. The team succeeds in repelling the attack and calls for air evacuation.

While learners are engaged in the point of injury scenario, Pod B is being prepared for the next step of the exercise. A mobile motion platform is brought into Pod B and a mockup of a UH-60 helicopter assembled. As the medical team makes patients ready for transport, the center curtain lifts, allowing the team to begin moving patients into the UH-60. The scenario continues into the air evacuation phase of the exercise. The medical team provides life support to the patient while inside the UH-60.

During this time, Pod A is reconfigured so it is no longer the scene of an IED attack. Instead, medical equipment found in a forward operating base is brought in. The IED virtual environment is replaced with a virtual operating room. After the UH-60 lands, the patient is wheeled back into Pod A. The scenario then continues into the operating room phase of the exercise.

By alternating pods and configuring them in parallel with an ongoing scenario, training exercises can continue indefinitely. Currently, the WAVE and its auxiliary infrastructure is designed to support a continuous exercise of up to four days (96 h) in length.

3.3 Visual and Audio Rendering

The WAVE uses an array of back-projected screens to generate 3D stereoscopic images. The WAVE's visual rendering components are modular. The basic component is the display module. Each display module consists of a screen, a projector pair, and a pair of image generators. Front coated perspex screens are used as projection surfaces. Each screen is 108 in. tall and 81 in. wide. Each screen is back-projected by two 15,000 ANSI lumen projectors. Light from each projector is polarized before reaching the screen to facilitate 3D viewing. Each projector is driven by one image generator. An image generator comprises of a commercial off-the-shelf computer with a high-end consumer graphics card (nVidia GTX 980 at the time of writing). 24 display modules are used in the WAVE. There are 10 in. each pod and four in the corridor. Screens serving as doors (Fig. 1) are hinged. They also have supporting wheels to facilitate operation. Wheel positions are indexed relative to the floor to ensure consistent screen positioning. Figure 2 illustrates display modules.

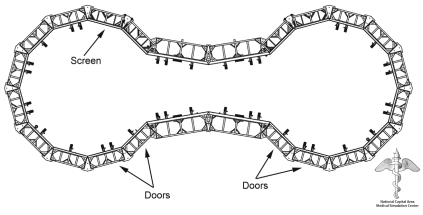

Fig. 1. WAVE layout

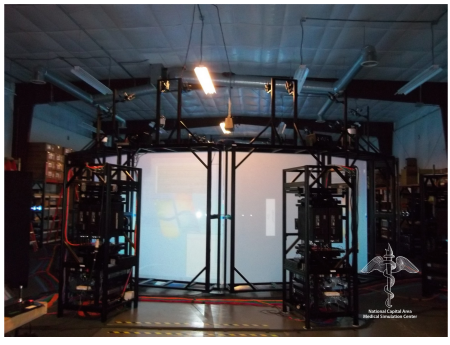

Fig. 2. Back of WAVE showing two display modules

Audio rendering complements the realism of the visual environment. Audio consistent with the scenario is generated during a training exercise. Audio

rendering in each pod is accomplished by a seven speaker system arranged in a ring above the display modules. Both ambient audio and directional sounds can be rendered. A 3 kW sub-woofer is positioned directly over each pod to provide high-intensity low frequency acoustic effects. E.g. explosions. A separate four speaker, one sub-woofer audio system is used in the corridor.

3.4 Tracking and Monitoring

An array of 24 Vicon motion tracking cameras is mounted just above the screens. This system is used to track specially marked individuals and equipment during training exercises. This system facilitates an interactive experience in the WAVE. For example, learners in a hazardous material training exercise can use tracked radiation detectors in the WAVE to receive real-time feedback on simulated hazards.

An array of 20 video cameras allow training exercises to be recorded. The cameras have pan-tilt-zoom capability, allowing any point within the WAVE to be examined. These cameras are sensitive to IR illumination. As such, they are capable of operating under low light conditions, or in total darkness.

3.5 Theatrical Effects

Suspension of disbelief is necessary for learners to remain engaged within the WAVE. In addition to 3D graphics and directional audio, the WAVE employs a series of theatrical effects. They are synchronized to the virtual environment. The virtual environment and theatrical effects work collectively to extend suspension of disbelief into the space physically occupied by learners. Currently, theatrical elements include: scent generators, smoke generators, air cannons, computer-controlled lighting, and a portable motion platform. We discuss each in turn.

Scent Generators. Olfaction is a primal sense [13]. The WAVE engages the olfactory sense with an array of six scent generators. A scent generator is a compact device consisting of a network interface, a small air compressor, bottles of scent liquid, and an exhaust fan. During operation, compressed air is used to force a fine spray of scent liquid from the bottle into the airflow of the exhaust fan. Six scent generators are available for training exercises. Two are mounted directly over each pod. The remaining four are self-contained portable units to be positioned as desired. Each scent generator can release up to six distinct odors. A range of different scents can be used based on the scenario. They include: burnt flesh, diesel fuel, burnt wood, urine, gunpowder, and gangrene. Figure 3 illustrates.

Smoke Generators. Smoke generators can simulate the effect of fire or explosions in a training scenario. WAVE smoke machines are self-contained, portable

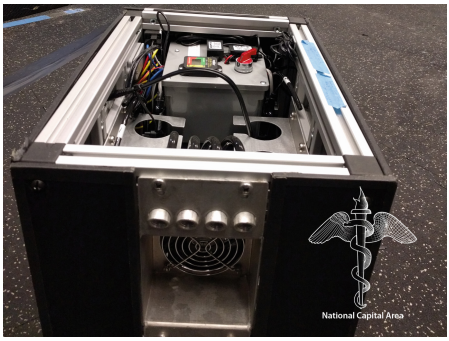

Fig. 3. Scent generator

Fig. 4. Smoke generator

units with a network (WiFi) interface. They are automatically controlled by the WAVE scenario. Smoke generators are synchronized with on-screen explosions, fires, or other virtual environmental features. Figure 4 illustrates.

Air Cannons. Air cannons add a kinetic element to the range of theatrical effects. Each air cannon is a self-contained unit. It consists of a compressed air tank, an electrically operated air valve, a WiFi trigger and batteries. A large directional barrel is fitted over the air release. Figure 5 illustrates. The barrel can contain harmless lightweight material to be launched toward learners. Typical materials include: textured cork (resembling dirt and asphalt), and clear silicone caulk pieces (resembling broken glass).

Computer Controlled Stage Lighting. Full spectrum LED stage lighting is used to illuminate the interior of the WAVE. Ambient lighting is matched with displayed scenes. They change based on the environment displayed on screen. In addition to ambient illumination, these lights can also be programmed to match training activities. For example, they can strobe in a manner consistent with a fire emergency. Lights close to emergency vehicles can similarly strobe using a color consistent with emergency vehicle lights.

Motion Platform. Some simulation exercises involve an air or ground evacuation phase. In these scenarios, the WAVE incorporates a portable motion platform. This device is wheeled to facilitate rapid deployment and movement in and out of WAVE pods as required. The motion platform provides three degrees of freedom: heave, pitch, and roll. Coupled with a mockup of the transport vehicle, this device generates motion consistent with a helicopter in flight, or ground transport over varying terrain. Vehicle movement can make some treatment procedures difficult. Learners gain experience tending to patients in these environments with the motion platform. Figure 6 illustrates.

Fig. 5. Air cannon

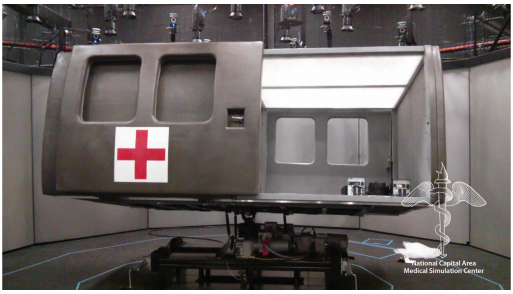

Fig. 6. Motion platform with helicopter mockup

3.6 Command and Control

Sections 3.3 through 3.5 described various elements that comprise the WAVE. Coordinating these elements in a synchronized fashion is done by the WAVE software. The command and control system of the WAVE is based on standard networking protocols (i.e., TCP/IP). Web-enabled relays are used for devices that require triggers (e.g., air cannons, smoke and scent generators). Hardwired Ethernet connections are used for fixed items. Portable units are WiFi connected. Protocol bridges are used for devices whose native control standard is not IP based. E.g., a DMX [14] to IP bridge is used to control the stage lighting.

4 Results

The WAVE has been in continuous operation since 2012. During this time, the WAVE has supported the learning objectives of the Uniformed Services University of the Health Sciences as well as regional military and federal emergency response teams. Smaller systems, termed WAVElets, have been deployed within the Air Force, Army, and Navy. Both military and civilian emergency response scenarios have been developed. Medical scenarios covering the Continuum of Care [15] Role 1 (point of injury care) through Role 4 (definitive care) are available. Joint en-route care scenarios are currently being developed. Scenarios supporting chemical casualty care are also available. Civilian emergency response scenarios include: civil disturbance, active shooter, improvised explosive detonation in enclosed (subway) and outdoor venues. We describe some of these scenarios.

Figure 7 illustrates a point of injury scenario. This scenario combines 3D computer generated audio-visual rendering with SPs, and multiple theatrical effects. Combat medics enter the WAVE to treat an SP playing the role of a wounded soldier. The instructor is at the extreme right. Learners must provide immediate aid to the wounded soldier. At the same time, awareness of their surroundings must be maintained. In this scenario, the WAVE provides context to the training exercise. Enemy combatants can also be activated to test situational awareness. An air cannon is hidden behind the sandbag barriers. Unchallenged enemy fighters will throw grenades. This causes smoke and air cannons to be discharged in close proximity. The multi-sensory response reinforces the importance of situational awareness without risk to learners.

Figure 8 illustrates a civilian mass casualty pipe bomb scenario. Multiple SPs play the role of bombing victims. Additional casualties are depicted virtually in the background. Learners must quickly identify casualties with varying severity of injuries. They must decide how to deploy limited resources to save as many individuals as possible. Background audio effects include: emergency vehicle sirens, screams from victims, and 2-way radio chatter. SPs playing the role of traumatized victims increase the level of chaos. Both the virtual and physical environment work to portray a realistic mass casualty scene. Students learn to perform the correct lifesaving measures without being overwhelmed by the situation.

Fig. 7. Point of injury–military scenario

Fig. 8. Mass casualty–civilian scenario

Figure 9 shows a wounded soldier being removed from a UH-60 helicopter after air evacuation. The UH-60 is mounted on the motion platform. During the scenario, the WAVE display simulates flight over varying terrain. The UH-60 sways in synchronization with the displayed terrain, simulating rotary wing flight. The WAVE's audio system generates background engine noise. Sound levels are set high enough to interfere with speech, simulating actual flight conditions. Learners practice en-route care in this environment. They learn how to provide care under conditions where sound, movement and vibrations can interfere with patient monitoring. Both SPs and HPS have been used to play the role of patient.

Figure 10 depicts an emergency room in a field hospital. The WAVE simulates the interior of a busy emergency room environment. HPS are used to simulate patients. Actual emergency room equipment, such as EEG monitors, are attached to the HPS. They provide physiological feedback on the patient's condition. Auditory cues include medical instrument signals, patient screams, and urgent conversation from other medical teams in the background. Additional stressors such as an air raid siren can be activated as required. Simulated incoming mortar explosions can also be triggered. When a mortar round lands close to the tent, WAVE projectors flicker in synchronization with the stage lights to simulate a momentary power outage.

Fig. 9. Aeromedical evacuation

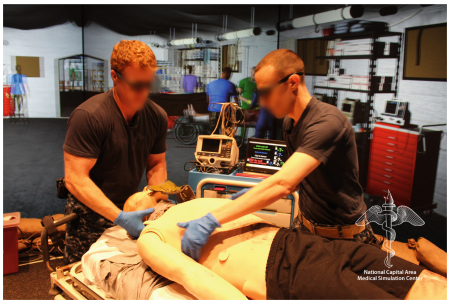

Fig. 10. Field hospital

5 Discussion

Medical education and training have generally used simulation modalities in isolation. In contrast, the WAVE adopts a tightly integrated approach to combining modalities. The WAVE is a unique environment for conducting capstone exercises. i.e., training events requiring learners to apply a combination of skills to address complex problems under stressful conditions. The WAVE fills a gap between classroom learning, and actual deployment. The WAVE delivers training exercises that are more flexible and at a lower cost than field exercises. Since all major elements are computer-controlled, exercises can be repeated, or stopped and restarted at a specific point. This ability reinforces learning, and allows the team to focus on deficiencies. In contrast, field exercises involve coordination between many disparate components. Repeating or restarting an exercise is not always practical. The WAVE is also well suited for mission specific training. E.g., medical support for VIPs. The virtual environment can be configured to exactly match event venues. This allows first responders to become familiar with evacuation routes and safe zones without physically practicing in that venue.

As hardware costs decrease, we anticipate increased application of fused modality training environments for medical instruction. Since individuals, equipment, and patients can be tracked within the WAVE, a distributed training capability involving multiple networked WAVE environments is possible. Teams can

practice in real-time on the same mission, even though they are geographically separated.

The WAVE is best suited to supporting small, specialized medical teams. It is best suited to training scenarios that cannot be readily replicated, e.g., chemical/biological attacks, or suicide bombings. The WAVE is also well suited for capstone exercises. It is not suited for learning individual procedures or skills where a stressful environment detracts from instruction.

6 Conclusion

The WAVE is a novel simulation platform. The objective is to deliver training scenarios with an unprecedented level of realism. It combines the three primary modalities of medical simulation: live actors, human patient simulators, and virtual reality. The WAVE does so with a high level of integration to deliver a seamless learning experience. By using a dual-pod configuration, the WAVE is capable of supporting training exercises of an indefinite duration. The WAVE is best suited for training specialized medical teams in difficult to replicate scenarios. Blended modality environments such as the WAVE hold promise in distributed training applications.

References

1. Tuthill, J.: See one, do one, teach one. Lancet **371**(9628), 1906 (2008)
2. Vargo, J.J.: See one, do one, teach one. Gastrointest. Endosc. **67**(3), 419–421 (2008)
3. Williams, R.G., Klamen, D.L.: See one, do one, teach one—exploring the core teaching beliefs of medical school faculty. Med. Teach. **28**(5), 418–424 (2006)
4. Barrows, H.S.: An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad. Med. Phila. **68**, 443 (1993)
5. Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., West, S.L.: The virtual standardized patient. In: Medicine Meets Virtual Reality, pp. 113–138 (2000)
6. Huang, G., Reynolds, R., Candler, C.: Virtual patient simulation at US and Canadian medical schools. Acad. Med. **82**(5), 446–451 (2007)
7. Stevenson, D.R., Smith, K.A., McLaughlin, J.P., Gunn, C.J., Veldkamp, J.P., Dixon, M.J.: Haptic workbench: a multisensory virtual environment. In: Stereoscopic Displays and Virtual Reality Systems VI, vol. 3639, pp. 356–367 (1999)
8. Massie, T.H., Salisbury, J.K., et al.: The PHANToM haptic interface: a device for probing virtual objects. In: Proceedings of the ASME Winter Annual Meeting, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, vol. 55, no. 1, pp. 295–300 (1994)
9. Liu, A., Bhasin, Y., Bowyer, M.: A haptic-enabled simulator for cricothyroidotomy. Stud. Health Technol. Inform. **111**, 308–313 (2005)
10. Acosta, E., Liu, A.: Real-time volumetric haptic and visual burrhole simulation. In: Virtual Reality Conference, VR 2007, pp. 247–250. IEEE (2007)
11. Wilkerson, W., Avstreih, D., Gruppen, L., Beier, K.-P., Woolliscroft, J.: Using immersive simulation for training first responders for mass casualty incidents. Acad. Emerg. Med. **15**(11), 1152–1159 (2008)

12. Scerbo, M.W., Belfore, L.A., Garcia, H.M., Weireter, L.J., Jackson, M.W., Nalu, A., Baydogan, E., Bliss, J.P., Seevinck, J.: A virtual operating room for context-relevant training. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 51, no. 6, pp. 507–511 (2007)
13. Wier, K.: Scents and sensibility. Monit. Psychol. **42**(2), 40 (2011)
14. https://en.wikipedia.org/wiki/DMX512
15. http://www.cs.amedd.army.mil/FileDownloadpublic.aspx?docid=1a73495d-1176-4638-9011-9e7f3c6017d8

Using Bots in Strategizing Group Compositions to Improve Decision–Making Processes

Shai Neumann[1(✉)], Suraj Sood[2], Markus Hollander[3], Freda Wan[4], Alexis-Walid Ahmed[4], and Monte Hancock[4]

[1] Eastern Florida State College, Melbourne, FL, USA
neumanns@easternflorida.edu
[2] Department of Psychology, University of West Georgia, Carrollton, GA, USA
[3] Center for Bioinformatics, Saarland University, Saarbrücken, Germany
[4] Sirius18, Reston, USA

Abstract. This paper explores utilization of bots created to play as participants in two games that have relevance in the field of economics. The first game is an iterated version of the prisoner's dilemma that has relevance to decisions requiring trust while conducting business and involves a binary decision. The simulation examines one dimension of the Myers–Briggs personality type and its potential association with strategies that may be employed during an iterated version of the game, leading to different levels of performance. A web–based bot is used for such simulation. Results appear to support the possibility of exploiting personality type in this dimension in the context of similar circumstances.

The second game is an iterated power to take game that simulates interactions between taxing authorities (takers) and tax–payers (responders), and involves a continuous set of possible decisions for each stage of the game. A specific Myers–Briggs personality type of specified extremity is embedded into each player bot by randomly generating answers to a version of the Myers–Briggs Type Indicator. These answers are individually associated with preferences and strategies that allow the bots to react to the changing game state according to their personality type.

This paper explores the combination of, among other things, team sizes, initial conditions of income tax rates in some of the major economies on different continents, adjudication methods for group decisions, as well as personality type extremity. In a large number of simulations, various group compositions in terms of MBTI personality type are matched up against each other under these conditions, and their performance is ranked for each role.

Results appear to suggest that selective recruitment practices based on personality type can enhance performance of both takers and responders and may lead to improvements of certain aspects of economic conditions.

Keywords: Iterated prisoner's dilemma · Bots
Iterated power to take · Personality type

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 305–325, 2018.
https://doi.org/10.1007/978-3-319-91467-1_24

1 Introduction

Experimental economics is a branch of economics that deals with experiments in lab settings that are designed to simulate individual or group behavior in order to better understand the functioning of economic systems. In many cases, human subjects play games and their decisions during games are studied. Such settings have their limitations in terms of scope of participants, number of iterations of a game they can play, and validity outside the particular lab. Many experiments take place on university campuses and samples are drawn from a limited student population in a specific region of a specific country. The global economy involves a highly diverse population living under different cultural, political, and economic conditions.

Consideration of the increasing role bots play in social media, reaching high levels of sophistication that makes them sometimes difficult to distinguish from human behavior, raises the possibility that bots can be created to play games of experimental economics, with the advantage that settings will reflect realities in different countries, and different economic and political systems. Using psychological instruments, human diversity can be infused into bot behavioral preferences, groups sizes can vary a lot more, and number of iterations can be expanded significantly.

We explore two iterated games that may be associated with economic related decision making, namely the prisoner's dilemma and the power to take game. The first requires a binary decision and the second requires a decision in a continuous range of values in some interval. In both cases, the iterated nature suggests the possibility of bots producing insights that allow for enhancements of some human decision making processes. Additionally, using personality classification may improve predictability of particular strategies employed by individuals or groups.

As a personality model, we employed the widely used Myers–Briggs–Type–Indicator (MBTI) that divides the general population based on four dichotomous dimensions into 16 mutually exclusive subpopulations with distinct personalities and distinct expected preferences and behaviors under various circumstances [12]. The MBTI dimensions model how different personality types perceive information and how they form judgments about said information. The first dimension is "Introversion" (I) versus "Extroversion" (E) describes how different personalities relate to other people and the inner versus the outer world. The "Intuition" (N) versus "Sensing" dichotomy constitutes the second dimension that focuses on preferring abstract over concrete information. The third dimension, "Thinking" (T) versus "Feeling" (F), is about judging based on objective, logical judgments versus intra– or interpersonal values. "Perceiving" (P) versus "Judging" (J) is the last dimension that is about spontaneous adaptation versus adherence to plans.

2 Background

2.1 Prisoner's Dilemma

The prisoner's dilemma is a game in which two individuals are in jail and are separately interrogated for a suspected crime [14]. Each person may either "cooperate" with the other – saying nothing about any potential crime while being questioned – or "defect" against them by confirming their complicit involvement. Dawkins in [5] offered the following payoff matrix to describe the possible combinations of choices and corresponding outcomes for oneself and another suspect when the prisoner's dilemma is "iterated", i.e. repeated several times (Fig. 1).

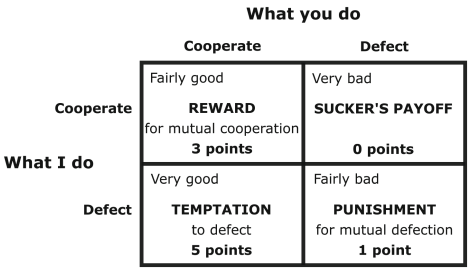

Fig. 1. Dawkins' iterated prisoner's dilemma payoff matrix.

The present study sought to add to the research on the iterated prisoner's dilemma by adding personality factors as influencing variables. Specifically, two players were created with independently–determined Myers–Briggs personality types, one representing a human user and the other a bot. It was assumed that Myers–Briggs Thinking and Feeling types each have distinctly–preferred strategies while playing the prisoner's dilemma. Of primary interest was how each type would perform relative to the other given this personality difference between them. Strategies were matched to Thinking and Feeling types based on the answers to questions such as:

1. Which personality preference is more likely to indicate one's tendency to cooperate with others?
2. Which personality preference is more likely to indicate one's tendency to act in their self-interest?
3. Which personality preference is more likely to indicate one's tendency to make peace offerings following mutual conflict?
4. Which personality preference is more likely to indicate one's tendency to be wary regarding the intentions of others?
5. Which personality preference, if any, is more likely to indicate a desire for all parties involved in a situation to benefit (i.e., to strive toward the "common good")?

Strategies. "Nice" strategies were defined in Dawkins' [5] sense as consisting of its user "never [being] the first to defect" (p. 12). Opposed to nice strategies are "nasty" strategies, the more regular use of which were further hypothesized to correlate negatively with their player's score. Dawkins defined nasty strategies as being those with which a player would "sometimes defect, however rarely, when not provoked".

Since users of nice strategies prefer to cooperate, it is arguable that they primarily act in both their and the other player's interests. This does not hold true for users of nasty strategies, who are comparatively more likely to defect and thus act in their own interest at the expense of the other player's.

Hypothesis. The hypothesis for this study was that a direct positive correlation would exist between the score of a player (whether automated or non–automated) and the extent to which they employ nice prisoner's dilemma strategies. This hypothesis reflects Dawkins' own conclusion, which was that "nice guys do well in this game" (p. 212).

2.2 Power to Take Game

The power to take is a game played by two opposing sides, with one side playing the role of a take authority that claims a percentage of the other sides's income, and the responding side deciding what percentage of their own wealth to destroy based on the take rate. The game was used in a laboratory experiment by Bosman and van Winken in [1] with students as players to investigate the impact of emotions on behavior. Initially, it was played by individuals, however, Bosman et al. expanded the experiment in [2] to be played in teams in order to explore group decision making.

Previous Findings in Power to Take Experiments. In [1], Bosman and van Winken found that responders typically destroy nothing or everything. The latter typically happened when the responders experienced negative emotions in response to a take rate that was subjectively too high. In line with this finding, responders' expectations regarding an acceptable take rate was a significant factor for the probability of destroying income.

In [2], Bosman et al. found that the individual decisions in a group follow the same trend observed in the game played by single players. Additionally, most takers thought a 50% take rate was fair, whereas responders tended to find a take rate of 0% as fair. However, only a small minority of the players were concerned with fairness in the group decision making process. Furthermore, both sides ignored the group decision making process of the other side and viewed the other team as a single entity. Lastly, they observed that group decisions were largely in line with a simple majority rule as well as the average of individual take inputs.

Overall, they came to the conclusion that emotional reactions, which they termed "emotional hazard", result in decreased efficiency both in the two–player

and group power to take game. For group decision making processes it seems to be important to consider both the decision making rule as well as the impact of individual decisions. They pose the question of why there are large differences between individual players and how these differences relate to the group decision making process.

Possible Impact of Personality Type on Group Decision Making. Bosman et al. suggested in [2] that expectations–based classification is an important explanatory variable for destruction rate as well as take rate. This fact may suggest that different personality types will have different expectations and will react differently to different realities. Differences in personality type might partially account for the different importance placed on fairness and other factors during the decision making process of both individuals and groups. The personality type composition of teams could have an impact on intra–group dynamics and could thus result in different group behavior.

Relevance of the Power to Take Game for Economics. The take authorities and take rate could be interpreted as the government imposing tax rates on the populations, whereby the destroy rate of the responders could represent the amount of effort people will make to reach a certain level of wealth given a tax rate. It is reasonable to assume that under extremely high tax rates a large segment of a highly talented workforce would prefer to reduce workload to avoid the unattractive prospect of seeing very little reward for maximum effort. When that is the case, the productivity of an economy suffers, and tax collection is reduced. If that does not happen in a certain society, then 100% tax rate would become a viable option and a taxing authority will likely be happy to impose it. However, in situations where the tax payers respond by reducing their productivity to nearly 0%, such an economy could be destroyed.

In an iterated version of the game, the players do not only respond to the immediate decisions of the opposing side, but also to the game history up to the current game round. This might be similar to tax authorities considering the effectiveness of prior tax rates when deciding to increase or decrease the tax rate. Similarly, tax payers might have expectations regarding an acceptable tax rate based on prior tax rates, as well as based on the reactions of the tax authority to signals of dissatisfaction.

Research Questions. Our interest in this context is to see what kind of personalities, individually, and in groups, would potentially perform either one of the two roles in the game particularly well, and which personalities would underperform in either role. A related question of interest is whether certain compositions of personality types can be linked to increased or decreased performance in the game, and, as an extension, which personalities or composition of personalities would be suitable as tax "collectors" and which would be successful as tax payers. Furthermore, we wish to investigate if there are take rates that overall result

in the best (or worst) performance regarding overall economy, tax collection or net earnings. Finally, it is of interest to examine the contribution of adjudication methods, team size, and prior, possibly cultural, history of tax rates to the group decision making process.

3 Methodology of the Prisoner's Dilemma

3.1 Type Designation and Personality–Based Strategy Assignment

MBTI F types were assigned three nice strategies out of a total of five, meaning that they usually did not defect unless the other player did first. In contrast, T types were assigned three nasty strategies (out of five strategies, total). These decisions were made because "individuals who score the highest on scales of disagreeableness [meaning they are more likely to also score as T rather than F types [10]] appear to others as being conceited, egocentric, competitive, antagonistic, skeptical, overcritical, or distrustful toward rivals" ([11], p. 120). Those who are more likely to score as F types were thus assumed to be more cooperative and trustful toward other players.

Before the games started, the human user and bot (named "Sirius" after this project's group) were typed independently. The human user was prespecified to always be an F, and Sirius a T. However, along each of the I/E, N/S, and J/P dimensions, players were typed randomly. After the program generated the user and Sirius' types, they were assigned distinct strategies for the coming set of prisoner's dilemma rounds based primarily on whether they were Thinking or Feeling types.

Some strategies (i.e., "Tit for Tat" and "Random") were possible for both players regardless of their types, whereas others were only possible for either T or F types. For instance, "Always Defect" was only possible for the former and "Remorseful Prober" was only possible for the latter. F strategies included Tit for Tat, "Naive Peacemaker", Remorseful Prober, "Always Cooperate", and Random; T strategies included Tit for Tat, "Tit for Two Tats", "Naive Prober", Always Defect, and Random. Descriptions of each of these strategies were borrowed from existing literature and work surrounding the iterated prisoner's dilemma [4,7].

3.2 Simulation

Following execution of the designation program, strategies were inputted into a minimally–adapted version of a Web–based prisoner's dilemma simulator [4] (Fig. 2). Since multiple strategies were assigned uniquely to both T and F players and the comparative performance of these two types was of primary interest, 50 iterations per type and strategy designation were run.

Each simulation was followed by the appearance of a dialog box which showed the game's final outcome.

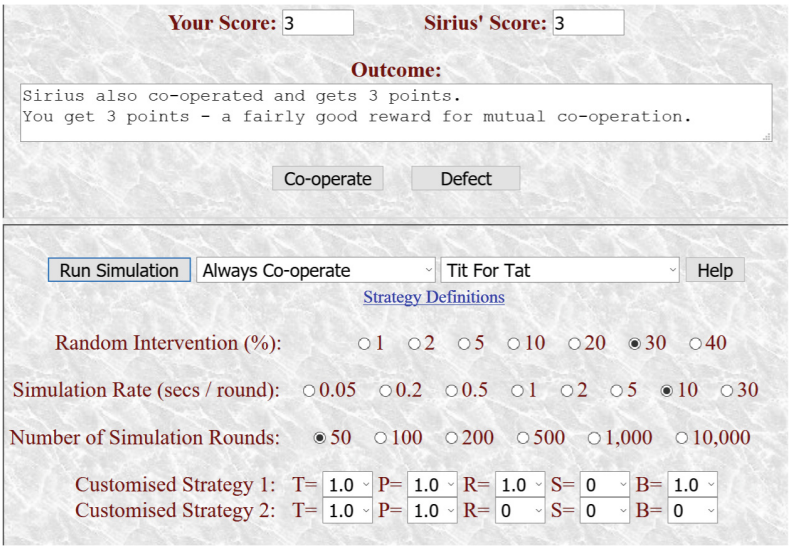

Fig. 2. First round of the iterated prisoner's dilemma (human strategy on the left; bot on the right).

4 Methodology of the Power to Take Game

In this study we simulated iterated power to take games with bots as players in order to examine factors and behaviors that might be relevant for group decision making processes and reactions to taxation circumstances over time. The investigated factors include personality type, team size, cultural background, adjudication method and team composition. Differences in behavior can be caused by, among other things, different personality traits that have been measured in models such as the Five Factor Model (FFM) [10]. For this study, a similar instrument, the Myers–Briggs Type Indicator (MBTI) [9], that nevertheless taps into personality traits described by the FFM [10] was employed to model automated bot behavior after human behavior in the power to take game.

4.1 Bias–Based Reasoning

In this study, we decided to implement a bias–based Knowledge Based Expert System (KBES) as described by Hancock [6]. The expert system allowed us to embed several heuristics for strategies into the player bots that might be employed by humans in the power to take game. Just as experts are biased by *a priori* beliefs, bias–based reasoning was used to determine the bots' *a priori* biases in making game–related decisions. This enabled the bots to react dynamically to changing game situations based on their personality–based preferences and disinclinations towards different factors and strategies.

A collection of preferences $p_i \in [0,1]$ for a certain factor or strategy and a collection of disinclinations $d_i \in [0,1]$ for that same factor or strategy are accumulated into the overall preference p and disinclination d as follows:

$$\begin{aligned} p &= p + p_i\ (1-p) \\ d &= d + d_i\ (1-d) \end{aligned} \tag{1}$$

with $p := 0$ and $d := 0$ at the beginning. This assumes values in $\in [0,1]$.

This method was used to compute the strength of the bots' reactions to changing circumstances. Additionally, it was employed to determine their preference for certain personality–based factor we deemed relevant for behavior during the power to take game.

Given the preference p and disinclination d of a bot towards a certain factor or situation, the attitude Δ of a bot was calculated as follows:

$$\Delta := p - d \tag{2}$$

Δ assumes values between -1 and 1, whereby values >0 imply a preference for that factor in questions, whereas values <0 imply a preference for the opposite. Since we were mainly interested in the positive preference towards a specific factor and not its opposite, we treated values <0 as 0.

4.2 Game Basics

In the first step, the game loads a configuration file that specifies: the number of rounds, the amount of money each round, adjudication method, prior take rate history, exact team composition in terms of MBTI type, the personality extremity, and the minimum variability that each bot exhibits. Based on these specifications, the bots are initialized for each team and play the game as described below.

Bot Creation. Given the MBTI four letter code, as well as the two parameters lower bound $l \in [0,1]$ and upper bound $u \in [0,1]$ that describe how extreme the personality is supposed to be:

1. For each dimension with n questions, randomly generate between $l \cdot n$ and $u \cdot n$ positive answers to the corresponding questions in the MBTI questionnaire. The random generation is uniform.
   For example, for the I–E–dimension with $n = 10$ questions of an introvert, $l = 0.8$ and $u = 0.9$, randomly generate 8 or 9 introvert answers plus, correspondingly, 2 or 1 extrovert answers.
2. Based on the bot's specific answers to the MBTI questionnaire, compute various aspects of personality that are relevant to determine employed strategies.

Game Loop. In each game round:

1. Each responder bot earns the specified amount of money.
2. Taker bot decision:
   (a) The taker bots decide individually what percentage of the responder bots' income they want to claim.
   (b) The game collects these individual taker decisions.
   (c) Each taker bot decides if it wants to adjust its decision based on the opinions of its team members.
3. The game adjudicates the final take rate.
4. Responder bot decision:
   (a) The responder bots decide individually how much of their own income they want to destroy.
   (b) The game collects these individual responder decisions.
   (c) Each responder bot decides if it wants to adjust its decision based on the opinions of its team members.
5. Remove the destroyed income from the responder bots.
6. Take the taker rate from the remaining responder bot income and add it to the collective money pool of the taker bots.

4.3 Relevant Personality Type–Based Factors

In similar studies involving MBTI and bots, strategies were assigned to the bots based on dimensions alone, sometimes taking into consideration the strength of the preference for that dimension [16]. However, after looking at a version of the MBTI test, different question in the same MBTI dimension appeared to imply different preferences for strategies in the power to take game. Thus, going by dimensions alone seemed insufficient for this context. On the other hand, neither did it seem to be a good idea to look at each test question individually in order to extrapolate entire strategies from that single question alone.

Therefore, we focused on several factors related to the information that was available to the bots during the game, or that were identified in previous studies as potentially relevant for the power to take game [1,2]. We then looked at each question in the MBTI test and assigned values representing if a particular answer implied a preference or disinclination towards that factor, and how pronounced that preference or disinclination was. Bias–based reasoning was then used to compute the values of these factors for each bot during bot creation based on the specific answers it selected on the MBTI test. Some factors required additional computation that are explained below. An overview of the factors and their value ranges in extreme personalities can be found in Table 1.

Consensus. In group setting like our version of the power to take game, decisions are not made by individuals in a vacuum. Instead, they are exposed to the opinions of their team members. To what degree they are swayed by their team's opinions partly depends on their personality type. The Thinking versus Feeling MBTI dimensions seemed to be especially relevant for this factor.

Table 1. Factors relevant for the power to take game in MBTI types with 100% in each dimension. The value range of each factor is $\in [0, 1]$.

Type	Consensus	Experiment	Reactivity	Fairness	Greed	Information	Tradition
ENFJ	0.4391	0.5498	0.2166	0.5778	0.4222	0.0734	0
ENFP	0.6654	0.9813	0.4063	0.6322	0.3678	0	0
ENTJ	0	0.4151	0	0	1	0.8958	0
ENTP	0	0.7551	0	0	1	0.7189	0
ESFJ	0.5564	0	0.2192	0.2623	0.7377	0.0851	0.8673
ESFP	0.8351	0.4054	0.4085	0.3094	0.6906	0	0.1244
ESTJ	0	0	0	0	1	0.8967	0.5755
ESTP	0	0.2794	0	0	1	0.7206	0
INFJ	0.3656	0.3466	0.2003	0.5778	0.4222	0.0609	0
INFP	0.5801	0.6962	0.3815	0.6322	0.3678	0	0
INTJ	0	0.2478	0	0	1	0.8503	0
INTP	0	0.5304	0	0	1	0.6818	0
ISFJ	0.4731	0	0.2031	0.2623	0.7377	0.0733	0.8691
ISFP	0.7343	0.1096	0.3834	0.3094	0.6906	0	0.1283
ISTJ	0	0	0	0	1	0.8513	0.5792
ISTP	0	0.0172	0	0	1	0.6835	0
Min	0	0	0	0	0.3678	0	0
Mean	0.2906	0.3334	0.1512	0.2227	0.7773	0.4119	0.1965
Max	0.8351	0.9813	0.4085	0.6322	1	0.8967	0.8691

The consensus factor ranges in value from 0 to 1, whereby 0 implies that the bot is not influenced by its team's opinions at all, whereas 1 implies that it wants to go entirely with the group consensus.

Variability. Human beings are not perfectly consistent in their behavior, and exhibit some degree of variability in their decisions. This is partly caused by factors unknown to the observer, but seems to be influenced by personality type as well. This factor seemed to be largely influenced by the MBTI Judging versus Perceiving dimension.

Given a base variability value $b \in [0, 0.5]$ and the bot personality's difference between preference and disinclination $\Delta_{p,d} \in [0, 1]$ for variability, the variability v of a bot was computed as

$$v := (1 + \Delta_{p,d})\ b \tag{3}$$

The variability factor thus ranges in value from b to $2b \in [0, 1]$, with higher values implying higher variability.

Experiment. This factor is similar to variability but taps into other aspects of personality as well. Namely, it describes how much a bot prefers to stick to

one strategy or experiment with different approaches. The mainly relevant MBTI dimensions seemed to be Judging versus Perceiving and Sensing versus Intuition.

The experiment factor ranges in value from 0 to 1, whereby 0 implies a complete disinclination for experimentation, whereas 1 implies that experimentation is extremely important to the bot.

Fairness and Greed. Bosman et al. identified fairness as a potential factor in the power to take game that was considered by a minority of the players [2]. This factor describes how much the bot values fairness versus maximizing its own gain during the decision making process. Thinking versus Feeling seemed to be the most relevant MBTI dimension.

The fairness factor ranges in value from 0 to 1, where 0 implies a greedy personality with complete disregard for fairness, whereas 1 implies a very fair bot that cares little about maximizing its own gain at the expense of others.

Tradition. Bosman and van Winken found that prior expectations were a significant factor for the probability of destroying income in the power to take game [1]. In our implementation, we added a prior take rate history to the game in order to give the bots expectations at the beginning of the game (Table 2). This factor describes how important these expectations are for the bots. This seemed to be mainly influenced by the Intuition versus Sensing and Judging versus Perceiving MBTI dimensions.

However, the relevance of this prior take history likely declines during the course of the game compared to other factors. Given the bot personality's difference between preference and disinclination $\Delta_{p,d} \in [0, 1]$ for tradition, the tradition factor t of a bot in round i of a total of N rounds was computed as

$$t := \frac{1}{2}\left(t + t\,\frac{N-i}{N}\right) \tag{4}$$

The tradition factor ranges in value from 0 to 1, whereby 0 suggests that the bot does not care at all about conditions before the game start, whereas 1 suggests a strong expectation that these initial conditions are maintained.

Table 2. Power to take: Income tax rates in percent in the United States of America, Germany, United Kingdom, Japan and Russia (Source: KPMG [8]).

Country	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
USA	35	35	35	35	35	35	35	35	35	35	39.6	39.6	39.6	39.6	39.6
Germany	48.5	45	42	42	45	45	45	45	45	45	45	45	45	45	45
UK	40	40	40	40	40	40	40	50	50	50	45	45	45	45	45
Japan	50	50	50	50	50	50	50	50	50	50	50.84	50.84	50.84	55.95	55.95
Russia	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13

Information. The game is potentially played with a larger number of rounds. It seemed unlikely to us that human players would remember the entire game history perfectly, or care for it. Instead, we assumed that recall would be better for the last rounds and that some personality types would consider more information than others. This seemed to be mainly influenced by the Intuition versus Sensing, Perceiving versus Judging and Thinking versus Feeling MBTI dimension.

The information factor ranges in value from 0 to 1, where 0 implies that the bot only looks at immediate information, whereas 1 implies consideration of as much information as possible.

Reactivity. "Emotional hazard" was identified by Bosman and van Winken as a factor that decreased efficiency in power to take games with human players, since responders chose to destroy everything when experiencing negative emotions [1]. Therefore, we decided to introduce the reactivity factor that describes how strongly a bot is going to react if something goes against its wishes or expectations, or in other words, how volatile it is. The main relevant MBTI dimensions seemed to be Judging versus Perceiving and Thinking versus Feeling.

The reactivity factor ranges in value from 0 to 1, whereby 0 suggests no reaction in response to seemingly upsetting situations, whereas 1 suggests extreme reactions.

4.4 Decision Making

Individual Decision Making. Each rule $i \in \{1, \ldots, n\}$ generates a proposed take or destroy rate x_i and corresponding weight w_i. The bot's preliminary decision $\bar{x}$ is the weighted mean

$$\bar{x} := \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{5}$$

Since the weights depend on the personality factors, each proposed take or destroy rate only features as much into preliminary decision as it is of interest for the bot, and it is possible that the output of some rules is completely irrelevant for a bot.

Taker Strategies. A simple rule that fires each round is the "tradition rule" that casts a vote for the mean of the cultural background take rate. The importance of this rule depends entirely on how important the tradition factor is for the bot.

The "greed rule" is another simple rule that is about maximizing the collection by maximizing the take rate. The weight of this rule depends completely on how greedy the bot is.

The "fairness rule" is slightly more complex and attempts to propose a fair take rate that takes into consideration the bot's own expectations as well as

the reactions of the responder team. The importance of this rule hinges on the fairness factor.

The "upset rule" is designed to add potentially irrational behavior in reaction to negative emotions caused by unmet expectations such as high destroy rates in response to culturally acceptable take rates, perceived fair take rates and exploitation of good will. Upset is accumulated using bias–based reasoning and depends on how important these factors are to the bot. How much the taker bot wants to increase the take rate depends on how upset it is, and the weight of this increase depends on the reactivity factor.

The most complex rule is the "strategy rule". It contains common sense reasoning such as not increasing the take rate once the goal take rate is achieved, detecting and countering manipulation by the responder team as well as the employment of one of two different strategies that are selected based on the bot's personality. Bots that are below average in the experiment factor try to introduce higher take rates with small, incremental increases, whereas bots above average in the experiment factor try to achieve their goal take rate by experimenting with larger, but targeted increases in take rate. The weight of these strategies depends on the greed, experiment and information factors as these seemed to be the factors most relevant for opportunistic and strategic thinking.

Responder Strategies. The responder bots possess a "greed rule" as well that attempts to maximize net earnings by minimizing the destruction. The degree by which the destroy rate is lowered and the weight of the proposed destroy rate depends on the greed factor.

The responders' "upset rule" is similar to that of the taker bots and leads to high destroy rates when the bot exhibits a high reactivity and the take rate deviates too much from the cultural background, perceived fair take rate or when lowering of the destroy rate was exploited by the taker team.

The "strategy rule" implements more sophisticated strategic thinking that uses the destroy rate as a signal of what the responders consider to be an acceptable take rate. Additionally, it attempts to detect and circumvent taker strategies aiming to increase the take rate. The weight of this rule depends on the greed, experiment and information factors.

Adjustment of Individual Decisions Based on Group Opinion. After the bots made their individual decisions, they get the opportunity to adjust their opinion based on the mean of the individual decision in the entire team. In our model, the inclination of a bot to reconsider its opinion depended on the consensus factor as well as the (un–)certainty regarding its own opinion. The closer the take or destroy rates proposed by the different rules, the more certain the bot is that it made the right decision. Whereas a wide spread in possible rates leaves the bot conflicted and uncertain.

Given the proposed rates x_i and corresponding weights w_i of rules $i \in \{1, \ldots, n\}$, as well as the number of non–zero weights m and the weighted mean $\bar{x}$, the uncertainty u is the weighted sample standard deviation:

$$u := \min\left(0.5, \sqrt{\frac{\sum_{i=1}^{n} w_i \cdot (x_i - \bar{x})^2}{\frac{m-1}{m}\sum_{i=1}^{n} w_i}}\right) \tag{6}$$

With the consensus factor c, the adjustment rate a is computed as follows

$$a := \frac{1}{2}c + u \tag{7}$$

Given the mean of the individual decisions in the group g and the bot's own, preliminary decision $\bar{x}$, the final decision d is then

$$d := ag + (1 - a)\ \bar{x} \tag{8}$$

Lastly, the variability v is applied to the reconsidered decision by choosing a uniformly random number from the interval $[\max(0, d - v), \min(1, d + v)]$.

4.5 Experiment Design

In a first experiment, we investigated the performance of teams consisting of only one MBTI type with extreme personalities (0.8–0.9 in each dimension) playing other such teams, with different combinations of team size and cultural background.

In a second experiment, we introduced two additional team compositions, namely a team consisting of the real MBTI type distribution in the general U.S. population [13] and a team consisting of a random sample of that real distribution. We constructed all of these teams both with a personality extremity of 0.8–0.9 and a more realistic 0.55–1.0 as control, while keeping the team size constant.

In a third experiment, mixed teams of the best and worst takers and best and worst responders played against each other as well as against a sample of the real distribution. See Table 3 for the precise experiment design.

4.6 Implementation

Game configuration files were generated in YAML format. Configuration file generation, bot creation, the power to take game and experiment analyses were all implemented in Python 3 [15]. The simulations were run and tested on Windows 10.

5 Results

5.1 Prisoner's Dilemma

Mean score for Feeling types was 110.8, standard deviation 43.4, for the Thinking types 130.8, 40.6, respectively. A one–way t–test is significant at $p = 0.037$, suggesting improved expected performance of T type personality over the F type (Table 4).

Table 3. Experiment design for the power to take game.

	Experiment 1	Experiment 2	Experiment 3
# Rounds	100	100	100
Money/round	100	100	100
Adjudication	Mean	Mean, median	Mean, median
Countries	USA, Germany, UK, Japan, Russia	USA, Germany, UK, Japan, Russia	USA, Germany, UK, Japan, Russia
Variability	0.025	0.025, 0.05	0.01, 0.025, 0.05
Type extremity	0.8–0.9	0.8–0.9, 0.55–1.0	0.8–0.9, 0.55–1.0
Team sizes	1, 2, 3, 4, 100	100	2, 4, 10
Composition	Pure types	Pure types, real type distribution, random sample of real type distribution	Team of 2 best takers, team of 2 best responders, team of 2 worst takers, team of 2 worst responders, random sample of real type distribution
# Configurations	6,400	25,920	9,000

5.2 Power to Take

Examination of the factor value ranges of extreme personalities in Table 1 shows that all T types have no interest in consensus or fairness, are not reactive and prioritize maximizing their own gain. In addition, N types are not interested in tradition, which is also the case for STPs. SJs exhibited the highest interest in tradition and SFPs exhibited a small amount. It is notable that all types are greedy to some degree and no type is perfectly fair, with the NFPs being the most concerned with fairness followed by the SFPs.

Experiment 1. Examination of experiment 1 results identifies ENTPs and INTPs as the best taker personalities for highest tax collection rates and ISFJs and ESFJs are the worst personalities. Regarding net earnings, the best responder personalities are INTJs and ESTPs, and the worst personalities are ENFPs and INFPs. These trends hold across different team sizes and cultural backgrounds.

The best matchup for takers in terms of highest tax collection rate were ISTJs as responders, and an examination of matchups between ENTPs and ISTJs clearly demonstrates the ability of the ENTP taker bots to impose high take rates on the ISTJ responder bots and without triggering high destroy rates. In contrast, the worst matchup for taker bots are ENFPs as responders, since

Table 4. Results of 30 iterated Prisoner's Dilemma simulations.

Round	User MBTI	User strategy	User points	Bot MBTI	Bot strategy	Bot points
1	ESFJ	Always Cooperate	150	INTJ	Tit for Tat	150
2	ENFJ	Random	117	INTP	Naïve Prober	112
3	ESFJ	Naïve Peacemaker	46	ESTP	Always Defect	66
4	ISFJ	Naïve Peacemaker	150	INTP	Tit for Two Tats	150
5	ESFJ	Always Cooperate	75	INTP	Random	200
6	ENFP	Always Cooperate	0	ENTJ	Always Defect	250
7	ENFP	Random	112	ISTP	Tit for Two Tats	112
8	INFP	Remorseful Prober	124	ISTJ	Naïve Prober	129
9	INFJ	Tit for Tat	110	INTP	Random	115
10	INFJ	Remorseful Prober	116	ISTJ	Random	111
11	ESFJ	Random	96	INTJ	Random	121
12	INFJ	AlwaysCooperate	147	ESTJ	NaïveProber	152
13	INFP	Random	126	ENTJ	Tit for Two Tats	126
14	ESFP	Naïve Peacemaker	150	INTP	Tit for Tat	150
15	ENFJ	Always Cooperate	150	INTP	Tit for Two Tats	150
16	ENFP	Random	122	ISTJ	Naïve Prober	122
17	ESFJ	Random	116	ISTJ	Tit for Two Tats	111
18	ENFP	Random	23	ISTJ	Always Defect	158
19	ENFJ	Tit for Tat	49	ENTP	Tit for Two Tats	54
20	ESFP	Always Cooperate	150	ENTP	Tit for Two Tats	150
21	ENFP	Remorseful Proper	149	ENTJ	Tit for Tat	149
22	INFJ	Always Cooperate	147	ISTJ	Naïve Prober	152
23	ISFJ	Tit for Tat	125	ISTP	Random	125
24	ISFJ	Random	102	INTJ	Tit for Two Tats	122
25	INFP	Remorseful Prober	150	ESTP	Tit for Tat	150
26	ENFJ	Tit for Tat	49	INTJ	Always Defect	54
27	INFJ	Remorseful Prober	49	INTJ	Always Defect	54
28	ISFJ	Naïve Peacemaker	125	ENTP	Naïve Prober	130
29	ESFJ	Tit for Tat	150	INTP	Tit for Tat	150
30	ISFJ	Naïve Peacemaker	150	ENTP	Naïve Prober	150
Total points			3,325			3,925

they resort to high destroy rates that lead to lowering of the economy size, tax collection for the takers and also lowering of the net earnings of the ENFP responders; a clear loss for all parties involved. The best overall matchup for responders in terms of highest net earnings were ISFJs as takers.

Experiment 2. Matchups of extreme personality types (0.8–0.9 in each dimension) versus the real type distribution in the general, human population with actual type extremity (0.55–1.0 in each dimension) exhibited the same patterns observed in experiment 1 (Table 5). Again, the ENTPs and INTPs were the best tax collectors, while ESFJ and ISFJ taker bots achieved the highest economy rate at the expense of their tax collection. Similar to the observations of experiment 1, the ENTP taker bots managed to establish extremely high take rates, resulting in high tax collection rates, while maintaining a moderate level of resistance by the general population.

Table 5. Power to take, experiment 2: Overall ranking of takers and responders with type extremity of 0.8–0.9 versus the real type distribution with type extremity of 0.55–1.0. The rates are the mean of what the teams achieved compared to what was possible in the corresponding games. The highest rates are marked in light gray, whereas the lowest rates are marked in middle gray.

Team	Takers				Responders			
	Economy Rate		Collection Rate		Net Earnings Rate		Destruction Rate	
ENFJ	88.35%	17	76.11%	8	33.50%	12	32.45%	4
ENFP	88.28%	18	80.13%	5	17.37%	18	67.54%	1
ENTJ	88.46%	13	83.62%	3	36.90%	1	14.96%	8
ENTP	88.73%	11	84.31%	1	30.54%	15	35.56%	3
ESFJ	91.59%	2	37.13%	17	33.77%	11	4.89%	14
ESFP	89.58%	8	65.71%	10	36.45%	4	14.98%	7
ESTJ	91.19%	3	42.18%	15	32.45%	13	3.22%	17
ESTP	88.66%	12	80.72%	4	36.86%	2	9.70%	9
INFJ	89.24%	10	68.08%	9	36.21%	6	16.28%	6
INFP	88.45%	15	77.91%	7	27.14%	17	47.30%	2
INTJ	88.40%	16	78.17%	6	36.80%	3	9.05%	11
INTP	88.45%	14	83.71%	2	36.36%	5	19.35%	5
ISFJ	91.69%	1	37.09%	18	32.06%	14	3.64%	16
ISFP	91.03%	5	44.59%	14	35.34%	9	6.71%	13
ISTJ	91.12%	4	42.05%	16	29.91%	16	2.65%	18
ISTP	89.46%	9	62.78%	11	34.10%	10	4.56%	15
real	90.70%	6	50.49%	12	36.01%	7	9.50%	10
sample	90.68%	7	46.72%	13	35.93%	8	8.84%	12

ENFP and INFP responder bots destroyed the most income against a taking authority from the general population, whereas ISTJ and ESTJ responder bots destroyed the least income. Unlike in experiment 1, ENTJ and ESTP responder bots performed the best in terms of maximizing net earnings for themselves, while ENFP and INFP responder bots performed the worst, which is likely caused by resorting to high destroy rates. These trends held across different matchups and other conditions under investigation.

It is important to note that both the real distribution and sample real distribution responders ranked in the middle tertile in terms of net earnings and destruction rate. This was also the case for such takers regarding economy rate and collection rate.

Table 6 presents an interesting result concerning the relationship between take rate and economy rate. The significant Pearson correlation of $r = 0.4199$ ($p < 0.001$) indicates a general trend of higher tax rates being associated with larger economies. However, this trend does not extend to the full range of take rates. In fact, the worst take rates for the economy rate are the extremely low and extremely high tax rates, a fact that should both be expected and desired.

Table 6. Power to take, experiment 2: Ranking of take rates regarding economy rate, collection rate and net earning rate with the real type distribution with type extremity of 0.55–1.0 as responders. The economy, collection and net earning rates are the mean of the rates in reaction to the respective take rate. Given are the Pearson's correlation coefficient r and the p–value of the correlation based on a two–sided t–test.

	Take rate	Economy rate	Take rate	Collection rate	Take rate	Net earnings rate
Best						
1	28%	91.80%	97%	87.26%	15%	77.21%
2	39%	91.75%	98%	87.05%	14%	76.97%
3	26%	91.59%	99%	85.60%	16%	76.48%
4	45%	91.46%	96%	85.01%	12%	76.11%
5	22%	91.39%	95%	84.92%	13%	75.87%
Worst						
1	12%	86.72%	12%	10.61%	95%	1.25%
2	99%	87.21%	13%	11.49%	96%	1.98%
3	13%	87.36%	14%	12.60%	97%	2.59%
4	90%	87.91%	15%	13.58%	98%	3.62%
5	89%	88.10%	16%	14.55%	99%	4.41%
r	0.419929		0.996563		−0.99833	
p–value	<0.001		<0.001		<0.001	

Experiment 3. When teams comprised of different personalities are matched against each other, we can observe a moderating impact of the adjudication method. This is largely because compositions of different personalities will tend to exhibit higher intra–group variability, so the adjudication method makes a difference as the group size increases.

6 Discussion

6.1 Prisoner's Dilemma

Results of simulations appear to suggest an association between one dimension of MBTI and expected reward in an iterated prisoners dilemma situation. In particular, it appears that a T type personality would be expected to outperform a F type personality. If true, that would be inconsistent with Dawkins' findings suggesting nice players win. As in many cases, conditioning on different assumptions and considering a different set of strategies for the PD game for the T type and F type, with somewhat different overlap, may lead to different results. So in reality, the practical implication of this part of the study is that such methodology of bot utilization for personality related binary decision–making

processes can be of value. Enhancements of this kind to binary decision–making processes may have limited power and may be realized only after a large number of iterations.

6.2 Power to Take

Results based on 41,320 configurations, each played for 100 rounds, appear to suggest that personality type matters for this game, and by extension, matters in the context of taxing authorities interacting with tax payers. This means that choices of particular personalities could lead to improved or reduced performance in any of the measures that one may be interested in, i.e. total tax collection, size of economy and net earnings, compared to other personality choices, or compared to the general population with its mixed personality type composition.

It may be worth noting that prior studies primarily involving students playing the game individually or in groups, obtained results with somewhat undesirable characteristics, namely destruction rates that tended to be either 0% or 100%, which lead to the responders decisions appearing to be essentially binary in those studies. This binary approach by a responder (tax payer) may be rational if the game is played once, since anytime the tax rate is under 100%, responders would maximize net earnings for the game by destroying 0%.

However, taxing circumstances repeat, and the reality is that tax collection as a function tax rates is a continuous function, so an iterated approach may be more appropriate in order to understand the evolution of the tax system with bot–based systems offering an advantage in the examined scope. In a large number of configurations played, the prevailing trends were of increasing tax rates over time, decreasing destruction rates over time, and decreasing net earnings over time. Such circumstances are not unrealistic in the real world, so rules used for bot behaviors, and the aggregate belief and disbelief method may be fairly effective for this problem.

Nevertheless, a potential problem is that we observed a deviation in net earning ranking from the average household income ranking in the general population [3]. This could be an artifact of using extreme personality types versus a real type distribution, but could also be an indication that additional or different factors and strategies might need to be considered, or that our interpretation of the relationship between MBTI test answers and bot behavior does not reflect reality.

In addition, the association of results of such games using our approach to economics may be limited in its validity. An economy is dynamic and iterating an identical game a large number of times may not always be a good model. There were elements of variability infused into behaviors and rules, but the essential rules of the games remained the same in terms of matrix of payoffs for the prisoners dilemma and per game wealth initially awarded to each responder each round in the power to take game.

7 Conclusion

This study suggests that personality may play a role in both binary and continuous decision–making processes as well as both single–player and group settings, and an iterated approach using bots playing humans may enhance our understanding of human decisions, potentially leading to better processes and better outcomes.

Future research may elect to examine different personality instruments, perhaps based on the Five Factor Model and infuse dynamics to the economy. Furthermore, it might be worth examining a variant of the iterated power to take game where the take authority has an income that is subject to the take rate, just like the responders. That way, the take authority would not just be concerned with maximizing the take rate but also take into consideration net earnings and the overall size of the economy, similar to taxing authorities in an economy.

References

1. Bosman, R., van Winken, F.: Emotional hazard in a power-to-take experiment. Econ. J. **112**(476), 147–169 (2002)
2. Bosman, R., Hennig-Schmidt, H., van Winden, F.: Exploring group decision making in a power-to-take experiment. Exp. Econ. **9**(1), 35–51 (2006)
3. Career Assessment Site: MBTI Socioeconomic Infographic. http://careerassessmentsite.com/mbti-personality-types-socioeconomic-infographic/. Accessed 13 Jan 2018
4. Davis, W.: Iterated Prisoners Dilemma Online Game and Simulation (1997). http://www.iterated-prisoners-dilemma.net/. Accessed 21 Jan 2018
5. Dawkins, R.: The Selfish Gene, 30th Anniversary edn. Oxford University Press, Oxford (2006). (Original Work Published in 1967)
6. Hancock, M.: Practical Data Mining. CRC Press, Boca Raton (2012)
7. Knight, V.: Strategies Index (2015). http://axelrod.readthedocs.io/en/stable/reference/all_strategies.html. Accessed 31 Jan 2018
8. KPMG: Individual income tax rates table. https://home.kpmg.com/xx/en/home/services/tax/tax-tools-and-resources/tax-rates-online/individual-income-tax-rates-table.html. Accessed 21 Jan 2018
9. MBTI Personality Test. http://www.maximusveritas.com/wp-content/uploads/2017/07/MBTI-Personality-Type-Test.pdf. Accessed 11 Dec 2017
10. McCrae, R.R., Costa, P.T.: Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. J. Pers. **57**(1), 17–40 (1989)
11. Mortensen, C.D.: Human Conflict: Disagreement, Misunderstanding, and Problematic Talk. Rowman & Littlefield Publishers, Lanham (2005)
12. Myers, I.B.: The Myers-Briggs Type Indicator (Manual). Consulting Psychologists Press, Inc., Palo Alto (1962)
13. The Myers & Briggs Foundation: How Frequent is My Type. http://www.myersbriggs.org/my-mbti-personality-type/my-mbti-results/how-frequent-is-my-type.htm?bhcp=1. Accessed 17 Jan 2017
14. Poundstone, W.: Prisoner's Dilemma: John Von Neumann, Game Theory and the Puzzle of the Bomb. Doubleday, New York, USA (1992)

15. Python Software Foundation: Python Language Reference, Version 3.6. http://www.python.org
16. Salvit, J., Sklar, E.: Toward a myers-briggs type indicator model of agent behavior in multiagent teams. In: Bosse, T., Geller, A., Jonker, C.M. (eds.) MABS 2010. LNCS (LNAI), vol. 6532, pp. 28–43. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-18345-4_3

Augmenting Clinical Performance in Combat Casualty Care: Telemedicine to Automation

Jeremy C. Pamplin[1,2(✉)], Ronald Yeaw[3], Gary R. Gilbert[3], Konrad L. Davis[4], Elizabeth Mann-Salinas[5], Jose Salinas[5], Daniel Kral[3], and Loretta Schlachta-Fairchild[6]

[1] Madigan Army Medical Center, Tacoma, WA, USA
jeremy.c.pamplin.mil@mail.mil
[2] Uniformed Services University of the Health Sciences, Bethesda, MD, USA
[3] Telemedicine and Advanced Technology Research Center, Fort Detrick, MD, USA
[4] Naval Medical Center San Diego, San Diego, CA, USA
[5] U.S. Army Institute of Surgical Research, San Antonio, TX, USA
[6] U.S. Army Medical Research and Material Command, Fort Detrick, MD, USA

Abstract. Emerging efforts in information science offer the possibility for clinicians to better utilize computer technology to decrease cognitive load, enhance decision making, and, improve patient outcomes. Recent natural disasters and mass casualty events across the United States and abroad spotlight the challenges of delivering healthcare in austere contexts. Austerity is a situation defined by limited resources of some or all of the following: equipment, medicines, diagnostics, personnel, knowledge, training, skills, and expertise. It is in this context that the military is focusing efforts to develop new telemedical, autonomous, and robotic systems to support local caregivers. Military human-computer models that support telemedicine and autonomous care in austere environments may help shape similar civilian healthcare solutions in similarly austere contexts of remoteness, natural disaster, and mass casualty. This paper will discuss the clinical challenges and capability gaps of providing comprehensive medical support in this context and some of the tools the military is developing to address them.

Keywords: Military medicine · Telemedicine · Automation
Clinical decision support

1 Background

Care for patients is complex and requires a multidisciplinary team of healthcare professionals to collaborate in order to optimize patient outcomes and avoid potential errors. In the best of circumstances and despite the wide availability of advanced

Disclaimer: The views expressed are those of the author(s) and do no reflect the official policy or position of the US Army Medical Department, Department of the Army, Department of Defense, or the U.S. Government. The investigators have adhered to the policies for protection of human subjects as prescribed in CFR 46.

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 326–338, 2018.
https://doi.org/10.1007/978-3-319-91467-1_25

monitoring and life support capabilities, cognitive lapses of the clinical team, which constitutes a joint cognitive system within a natural work domain, do occur. Such lapses are more likely in resource limited or "austere" circumstances – as in lower income countries or high-income countries during mass casualty events. These lapses result in less than optimal patient care, ineffective utilization of resources, and, in some circumstances, patient harm. While the field of telemedicine is over 50 years old [1], new and emerging technologies offer previously unavailable capabilities for clinicians to better utilize computer systems to decrease cognitive load, enhance decision making, and, hopefully, improve patient outcome. The military has invested significant research funding in developing clinical decision support (CDS)/artificial intelligence (AI), telemedicine/virtual health, and autonomous technologies to enhance combat casualty care throughout the evacuation continuum (Fig. 1) [2, 3]. For the last 20 years of conflict, the US Military has experienced air superiority and freedom of movement across the active battlespace, allowing for quick evacuation casualties to advanced medical support. However, the military anticipates that in future conflicts, access to advanced medical decision-making and surgical stabilization may be restricted or significantly delayed due to distance or adversary denial to freedom of movement. Consequently, there will be increased need to access the capabilities, particularly the expertise, of the multidisciplinary medical team in more austere, severely resource limited, pre-hospital environments at or near the point of injury [4–6].

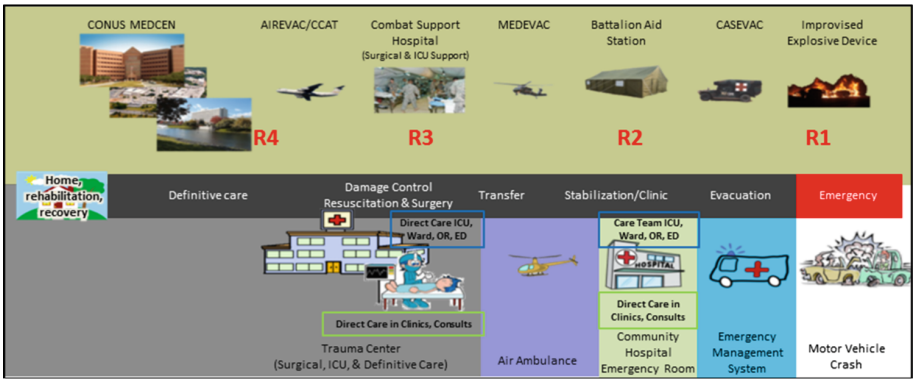

Fig. 1. Military vs. Civilian Care continuum. AIREVAC, Air evacuation; CASEVAC, Casualty Evacuation; CONUS, Continental United States; CCAT, Critical Care Air Transport; ED, emergency department; ICU, Intensive care unit; MEDCEN, Medical Center; MEDEVAC, medical evacuation; OR, Operating room; R1–4, Role 1–4.

Recent natural disasters and mass casualty events across the United States and abroad spotlight the challenges of resource limitations associated with austere medicine. Austerity is not a location, but rather a clinical context in which resources or expertise are severely limited or, in some cases, absent. It is in this context that the military is focusing efforts to develop telemedical, autonomous, and robotic systems to support local caregivers by providing enhanced medical capabilities, confidence,

knowledge, expertise, and, when possible, new technical skills. Consequently, these technologies have potential importance even in well-developed healthcare systems when certain resources, like experience with a new surgical approach or a new medication, are limited.

Examples of such innovation are already apparent in the civilian market following natural disasters [7–10]. In the wake of Hurricane Harvey, Dallas-based Children's Health set up pediatric telemedicine consults for displaced patients in shelters and companies like American Well, Doctor on Demand, and MDLive offered free remote consults to patients in affected areas [11]. After Hurricane Maria's landfall in Puerto Rico, the US Army's 47th Combat Support Hospital enabled telemedical consultation using satellite signal to areas with limited power, water, and even telephone lines (personal communications, authors JCP, KLD, DK).

Through these examples, it is possible to envision a future where military and civilian partnerships might enable medical aid in areas devastated by nature, disease, or even war. Virtual access to the military's robust and experienced trauma centers during peacetime also has the potential to provide enhanced clinical care to places without access to their own resources, providing the possibility for better clinical outcomes and reducing the costs of establishing such trauma/inter-disciplinary teams in every civilian facility.

CDS/AI systems, telemedicine, autonomous medical devices, and robotics will contribute capabilities to the current and future care of casualties in austere environments. These solutions provide comprehensive medical care in austere environments by bringing medical expertise and new capabilities to the point of need. Experts require information and particular resources to deliver their care, but not necessarily physical presence. Clinical decision support and telemedicine are current means to deliver expertise to the point of need; however, the local caregiver who physically delivers it requires specific training and equipment for it to be effective. Automation and robotics may alleviate some of the physical demands of the local caregiver and AI/deep machine learning are expected to advance CDS to the point where remote consultation with a telemedicine provider may become unnecessary for at least some clinical scenarios. In light of these anticipated needs, the Army has created three new Science & Technology research task areas: (1) Virtual Health, (2) Medical Robotics, and (3) Medical Autonomous & Unmanned Capabilities. This paper discusses the clinical challenges and capability gaps of providing comprehensive medical support in these environments, identifies some of the tools the military is developing to address them, and outline their potential benefits for the civilian healthcare.

2 Solutions

2.1 Telemedicine

Telemedicine, also referred to as virtual health, involves the use of telecommunication and information technologies to provide health assessments, treatments, consultation, and other services across distances [12]. There are four well-recognized types of telemedicine: synchronous, asynchronous, remote patient monitoring, and mobile health.

In synchronous telemedicine, the remote expert uses bidirectional communications technology to conduct or direct medical care in real time. This form of telemedicine requires the largest bandwidth, however offers the highest fidelity of information. Asynchronous telemedicine, also called "store and forward" telemedicine, involves the transmission of recorded health information to a remote specialist who then renders care outside of real-time interactions. In remote patient monitoring, personal health and medical data is collected via electronic communication technology, and then transmitted back to a provider (in a different location) to aid in medical decision making. Lastly, mobile health involves the use mobile communication devices, such as cell phones, tablets, or computers, typically in order to conduct public health practice and education.

By leveraging communication technology, telemedicine has the ability to connect less-experienced and sometimes untrained, medical care providers with medical specialists in order to enhance point of care clinical decision making and medical interventions. Optimal telemedical care, especially for complicated patient conditions, requires access to real-time bio-physiologic patient data (i.e. monitors, ventilator, infusion pumps, etc.). Unfortunately, transmission of this information across secure military networks in a standardized fashion for review by the remote expert is not always possible. Telemedicine also requires connectivity between the remote expert and local care-giver or patient. The amount of bandwidth required will vary based on the type of telemedicine encounter being performed, number of casualties, and the software that is being used to conduct it (Fig. 2). Given the bandwidth-constrained nature of most military operational environments, and likely any civilian natural disaster at least in its early stages, this requirement for connectivity is the principle limitation for current forms of telemedicine [7, 11, 13]. While the military has a long history of utilizing and advancing telemedicine and telemedical technologies [13–17], recent language in the National Defense Authorization Act of 2017 [18] will promote and advance this technology solution even further.

One current project, the **AD**vanced **VI**rtual **S**upport for **O**pe**R**ational forces (ADVISOR) is worth highlighting. ADVISOR is a low cost, low bandwidth, highly reliable pilot program funded by the Telemedicine and Advanced Technologies Research Center, Fort Detrick, Maryland. ADVISOR facilitates telephone communications using an automatic call distribution system to connect local caregivers with remote experts [4]. Phone calls are augmented by e-mail messages that include background casualty information and photographs to provide remote consultant consultation. If and when network capabilities allow, and local hardware is available, phone calls may be "escalated" to real-time video teleconferencing (VTC) with or without use of peripheral, video assisted physical exam equipment. ADVISOR brings medical expertise to a patient's side anywhere and at any time to optimize outcomes and reduce costs by avoiding unnecessary deployments of providers and evacuations of patients. Lessons learned strongly suggest that training local caregivers and remote experts with the new technologies is as essential for their adoption in the austere care setting as is the native usability of the technology itself [19]. This low-cost system also seeks to integrate and evolve technologies to optimize combat casualty care across the evacuation spectrum.

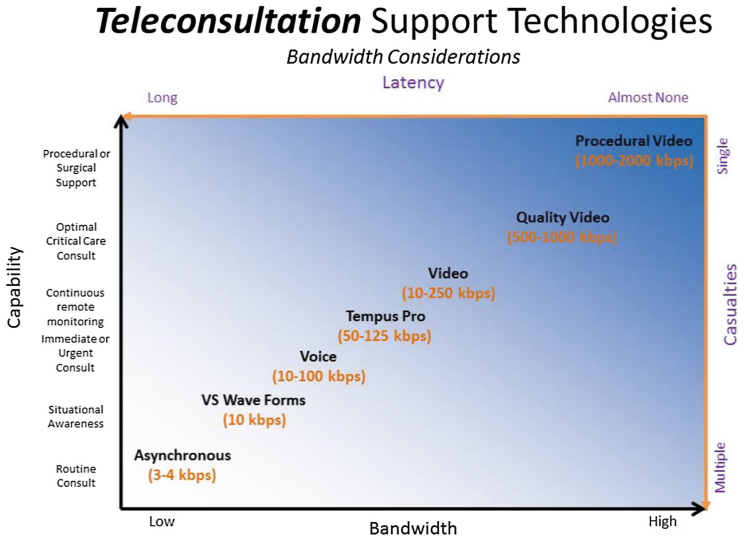

Fig. 2. Telemedical capabilities, number of casualties, and relative network requirements. Not shown is jitter or packet errors, both of which need to be low for higher capabilities.

As of January 2018, the ADVISOR system has supported over 70 real and simulated patient care encounters, primarily with virtual critical care support, but also with general surgery, orthopedic surgery, toxicology, and infectious disease. In both real world and training scenarios, the vas majority clinical support was provided with telephone calls only or telephone calls combined with e-mails (>60%). Video tele-consultation was use in <5% of cases, primarily due to poor network capabilities and low bandwidth. Real world cases have sought urgent recommendations about management for ocular and periocular infection, possible weaponized chemical/biologic exposure, fracture management, sepsis evaluation and management, wound care, surgical decision making, and evacuation guidance. Training cases primarily involve burn care, trauma, hemorrhage, shock, and respiratory failure management.

2.2 Clinical Decision Support Systems

One approach to mitigate the need to project clinical expertise to the point of need is the use of computer and information technologies to assist providers in delivering expert patient care when there are no experts available. This includes situations ranging from delivering care in a facility when there is no local expert available to assist in treating patients to delivering care by non-expert providers in geographically remote environments. Software or information technology-based solutions used to assist in the delivery of care are considered Computer Decision Support (CDS) Systems. In very resource constrained environments, the use of CDS could augment local caregiver capabilities – which in some cases may be no more than lay person rescue – by providing simple tools to allow providers to easily navigate and document patient care

based on a clinical practice guideline to more complex systems that can monitor, diagnose, and make treatment recommendations for the optimal care of the patient [20–23].

One example of a military CDS system developed by the military is the Cooperative Communication System (CCS) [21, 24, 25]. The goal of the CCS system is to provide users with a tools that enhances patient care by (1) improving and optimizing the display of information presented to the user based on the status of the patient and the user role, and (2) leveraging machine learning technologies to assist providers in managing patients by matching the current patient's condition with historical patients that have been analyzed and matched a multivariate model of patient variables and characteristics. The CCS system is an example of a CDS implementation that leverages existing data sets to provide users with enhanced capability tailored to the needs to the patient and providers through the use of enhanced displays and data analysis/modeling.

Once a treatment plan is required, a CDS can also be used to optimize and manage the care of the patient by providing tools to assist both experts and non-experts through the treatment phase for the patient. The Burn Navigator™ system [26] is an example of a CDS that provides the user with recommendations and treatment options for a patient who has suffered a major burn. The system is an FDA cleared CDS to help manage the fluid resuscitation of a patient during the initial 24 to 48 h after a major burn injury. Using a set of mathematical models, the system provides the user with hourly fluid rate recommendations to optimize the resuscitation needs of the patient. In addition, the system provides an enhanced display that allows providers to better visualize the patient status and treatment effects and increase situation awareness.

Tools such as these have been shown to increase provider efficiency and improve patient outcomes by harnessing the power of information technology through several ways. First, providing users with enhanced displays and interfaces, these systems increase provider situational awareness by converting data into useful information through the use of trends, graphs, and other advanced display technologies. Second, the use of these enhanced visualization tools reduces the cognitive load for providers and increases their efficiency during patient care [27]. This is especially critical in combat casualty care when patients may need to be cared for by less educated or less experienced providers. Third, CDS systems that provide treatment recommendations and options may further improve the management of the patient by assisting the provider in guiding the appropriate recommendation in cases where the provider may not be aware of the appropriate course of action, or by reinforcing what the expert provider knows by validating a specific course of action.

One common characteristic of CDS systems is the need to maintain a "human in the loop" concept. CDS capabilities are used to enhance patient care by providing users with diagnostics and/or treatment recommendations to optimize patient care. However, once a CDS has been fully validated and shown to provide effective care, the natural evolution of these is as a "closed loop" system to allow for the CDS to fully automate the care of the patient [28–31]. Models and algorithms driving CDS systems can form the basis for developing fully automated interventional system that require little or no human interface. Bridging the concepts of CDS and full automation will provide users in austere and possibly modern care environments with additional capabilities for managing patients in difficult situations. Use of fully automated systems in scenarios

with extended patient care, exceptional resources limitations, mass casualties, or extended transport times is necessary to fully optimize patient management when providers may not have all the necessary resources to properly care for these patients.

2.3 Autonomy and Robotics

The emerging fields of autonomy and robotics represent areas of significant possibility to solve some of the great medical challenges faced in austere care environments [32]. Remote surgery enabled by tele-robotics currently exists using dedicated fiber optic networks. The daVinci surgical robot is currently used in military and civilian hospitals for minimally invasive surgery [33]. This same type of technology could be leveraged, when combined with advanced physiologic sensors, computer vision and other autonomous or semiautonomous systems (i.e. autonomous anesthesia [31]) to deliver effective casualty assessment, triage, and surgical intervention in the absence of local experts and timely evacuation even on unmanned ground and air systems (drones and robots) [6–8, 34]. Conceptually, one might imagine that a drone ambulance, fully equipped with a robotic attendant and interactive video communications with a distant emergency medical provider, could respond to an emergency on a remote mountaintop or deep in a jungle.

"Military funded research has demonstrated that surgical robotic systems can be successfully deployed to extreme environments and wirelessly operated via microwave and satellite platforms [35]. However, employment of these capabilities on the battlefield have not yet progressed beyond experimental proofs of concept, are not ruggedized, and are tele-operated component capabilities at best. Significant additional research is required to develop supervisory controlled autonomous robots that can overcome the operational communication challenges of limited bandwidth, latency, and loss of signal in the deployed combat environment. Addressing acute and life-threatening injuries such as major non-compressible vascular injury requires development of new surgical robots that move beyond stereoscopic, bimanual tele-manipulators and leverage advances such as computer vision and application of directed energy technologies already used in non-medical military robotic systems."

Additionally, since the successful surgery to install fully implantable artificial cardiac pacemakers occurred in 1958, society has slowly become more accustomed to "closed-loop" medical care in which machines perform their function without need for human involvement. However, reliable closed-loop intervention remains an unsolved Artificial Intelligence challenge. Current AI systems do not have the capacity for judgment in the way that medical caregivers do. While clearly a science, medicine remains, in many ways, an art form. In addition to teaching the science of medicine, medical training is an extended apprenticeship that provides vast amounts of experience about how to interpret and make judgment calls in the context of uncertainty [3, 4]. Computers today lack this capacity. Additionally, while some success has been documented with closed loop systems in controlled settings, the reality in the field is more complex and current systems do not have the ability to account for such variability effectively enough to be proven clinically safe [3, 4]. Interestingly, telemedical

technologies may offer a means to "bridge the gap" between current computer capabilities and human judgment in the context of uncertainty and might allow semi-autonomous systems to be utilized in the near term (Fig. 3) [36]. Telemedicine provides opportunities to record data that could help teach computers about human decisions: the data that remote human experts require to make informed decisions during telemedical encounters is the same data that autonomous systems will need to make similar "decisions" in the future. This real-time data includes information from interoperable, cyber secured medical devices (i.e. patient monitors, medication pumps, ventilators, robotics) and the data analytics to make these signals standardized, synchronized, and salient. As robotic capabilities continue to evolve, autonomous and tele-operated semiautonomous robotic patient support systems could enable closed-loop patient monitoring and triage as well as robotic intervention.

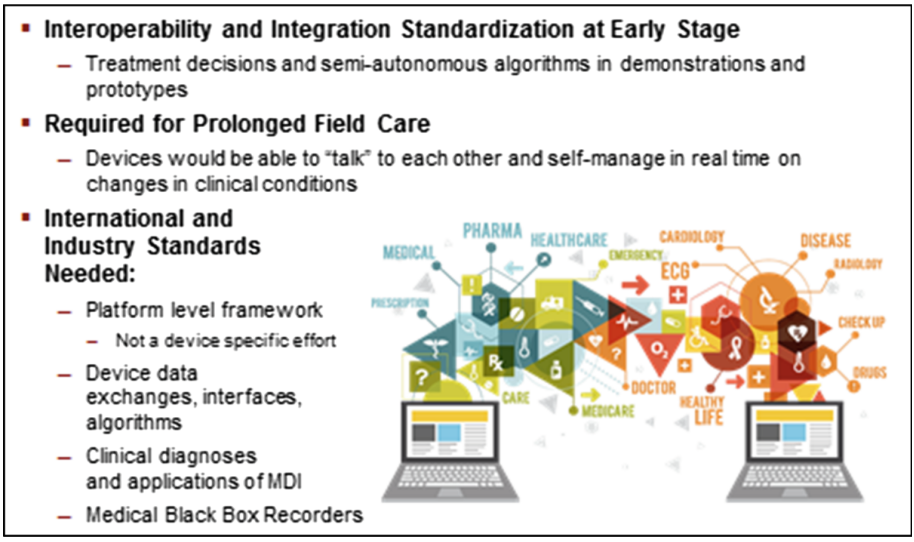

Fig. 3. Man vs. Machine, the progression of automation and how telemedicine can start to bridge the gap between artificial intelligence and expertise in medicine. Telemedicine brings the expert to the point of need in the absence of or in addition to the effective clinical decision making by inexperienced clinicians aided by automated systems and clinical decision support systems (CDS).

Transitioning between human in-the-loop systems and completely autonomous, closed loop medical care systems represent a major technical hurdle facing current medical care. In anticipation of combat casualty care in the future; the military is investing heavily in research portfolios to develop solutions to bridge this gap. Near-term prototypes will involve a hybrid of human and autonomous care models that shift between more human-intervention and more autonomous-care based on the clinical, technical, and logistical needs of a given situation. The military has identified the following as the most realistic research initiatives based on the complexities of the

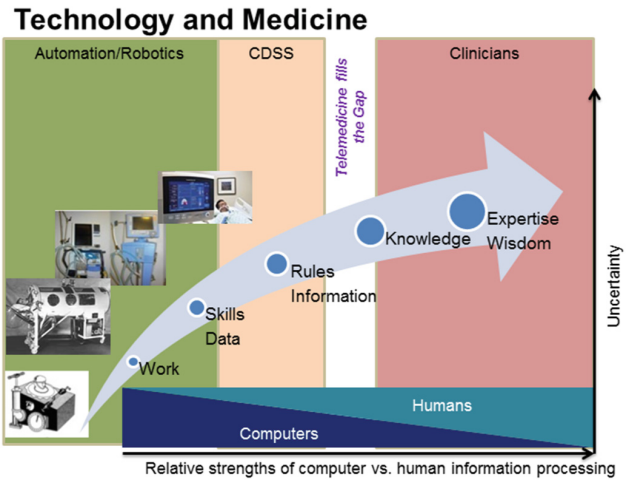

Fig. 4. Medical Device Interoperability Research Objectives to Support Telecritical Care. Source: Joint Program Committee 1; Medical Simulation and Information Sciences, US Army Medical Research and Materiel Command, Fort Detrick MD and the Defense.

technology involved, and the current state of the medical robotic and autonomous systems (Fig. 4):

(1) Development of an autonomous closed loop critical care systems-of-systems based on autonomous clustering and intelligent agents interacting with each other to provide care for multiple polytrauma patients. This type of system would act as a medical force multiplier by increasing medical care capacity during prolonged field care or mass casualty situations.
(2) Support continuous performance improvement and enable self-learning systems that can rapidly adapt to changing scenarios.
(3) Investigating artificial intelligence and machine perception systems for accurate detection and modeling of the human body to enable semi-autonomous robotic casualty extraction from complex environments and robot procedural intervention (e.g. placement of interosseous needle, needle thoracostomy, cricothyrotomy); these applications require high fidelity mapping of the human body in near-real time for safe physical contact with the casualty.

2.4 The Medical Fusion Center

Optimization of the complex care environment for both military operations and civilian emergency response requires a coordinated, real-time common operating platform, much like the control room for space flight. Alternatively, a virtual control room can be developed allowing interactions from around the globe. Real-time visibility of all medical assets in support of both civilian and military conventional and special

operations forces, particularly in kinetic multi-domain battlefields or disaster relief operations will facilitate optimal use of resources during evacuation, transport, patient care, tele-support/mentoring, and resupply. Ultimately the goal is a comprehensive "System of Systems" to support multiple end-users with medical fusion centers of global medical capabilities to facilitate a coordinated medical system and support all dimensions of healthcare. Key components of this type of center include: (1) real-time visualization of the area of operations; (2) a "medical intelligence" platform comprised of tele-support, decision support and analytics to facilitate clinical care, operational and logistic requirements, and evacuation/patient movement; and (3) robust reporting features for continuous performance improvement processes.

The primary aims of a functioning medical fusion center are to: (1) Send the appropriate resource (personnel, platform, and capabilities) to the right place at the right time to meet the specific patient requirements, (2) Support continuous performance improvement and enable self-learning systems, (3) Inform algorithms and decision support systems for tasking personnel assignments, medical treatment facility placement, and patient evacuation/movement ("Intelligent Tasking"), (4) Facilitate Virtual Health: tele support, just-in-time mentoring, remote patient mentoring, and reduced medical error rates, (5) Interface with allied military services, civilian and Government emergency response organizations.

Fundamental challenges to such a comprehensive, interconnected system include adequate connectivity and bandwidth. Assumptions of interruption of communications in a military or emergency setting are understood so optimizing local decision support tools will be essential. Streaming real-time information may not always be possible so AI will be needed to supplement decision support algorithms.

3 Future Directions

Medicine is woefully behind other industries, such as aviation and automotive industries, in adopting advanced semi and fully autonomous operations [37, 38]. The opportunity to provide high quality, reproducible, and effective monitoring for patients that informs local and remote care providers is an important, and achievable, goal for healthcare; it is a requirement for a military that needs to support isolated patients in austere locations or casualty evacuation in unmanned, pilotless systems. To this end, the Defense Health Agency and the military services, along with the other federal agencies such as the Food and Drug Administration and the National Institute of Standards and Technology, are moving toward investing in a comprehensive program of research. The overall objectives of this program are to develop an open, standards-based technical architecture and reference model for medical device data, sensors and actuators, communications enablers, algorithms, and knowledge representation that is suitable for informing autonomous, closed-loop, human-computer integrated CDS (Fig. 5).

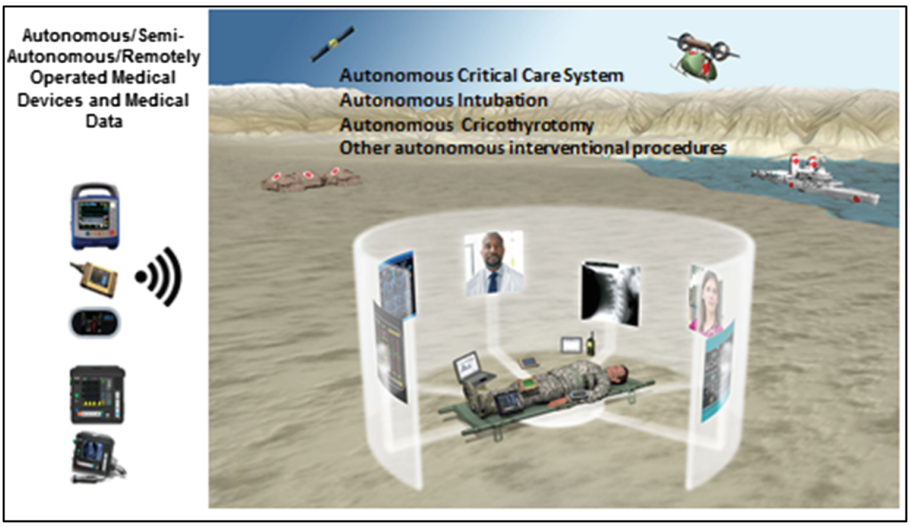

Fig. 5. Future Vision of Medical Device Interoperability and Telecritical Care. Source: Joint Program Committee 1; Medical Simulation and Information Sciences, US Army Medical Research and Materiel Command, Fort Detrick MD and the Defense Health Agency.

4 Conclusion

Rapid advancements in both medical and non-medical technologies offer the potential for improving clinical performance of both novice and advanced medical teams in resource limited contexts. Additionally, the correct application and integration of new medical combat casualty care doctrine with proven technologies offers significant improvements in both readiness and access to care while reducing unnecessary and risky evacuations. Beyond the potential for optimizing casualty care in the military, these efforts should be applicable to civilian healthcare systems in the future. Military human-computer models that support telemedicine and autonomous care in austere environments may help shape similar civilian healthcare solutions in similar environments. An open, standards-based infrastructure of medical devices could revolutionize healthcare by reducing research and development time, facilitating integration of new devices into the medical ecosystem, reducing costs, increasing healthcare reliability and safety, improving documentation of care, enabling more effective training, making logistical support more efficient, and, ultimately, improving patient outcomes. Efforts to develop and evaluate these solutions require continued military, academic, industry, cross -government, and international collaboration and technical ingenuity to be successful now and in the future.

References

1. Telemedicine Guide Article. https://evisit.com/what-is-telemedicine/. Accessed 22 Feb 2018
2. Rasmussen, T.E., Baer, D.G., Lein, B.C.: Ahead of the curve: sustained innovation for future combat casualty care. J. Trauma: Inj. Infect. Crit. Care 1–12 (2015)
3. Rasmussen, T.E., Reilly, P.A., Baer, D.G.: Why military medical research? Mil. Med. **179** (8S), 1–2 (2014)
4. Powell, D., McLeroy, R.D., Riesberg, J., Vasios, W.N., Miles, E.A., Dellavolpe, J.: Telemedicine to reduce medical risk in austere medical environments: the virtual critical care consultation (VC3) service. J. Spec. Oper. Med. **16**(4), 102–109 (2016)
5. Riesberg, J., Powell, D., Loos, P.: The loss of the golden hour. Spec. Warf. **30**(1), 49–51 (2017)
6. Perkins, G.D.G.: Multi-Domain Battle. Military Review [Internet], pp. 1–8, July–August 2017. http://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/July-August-2017/Perkins-Multi-Domain-Battle/
7. Meade, K., Lam, D.M.: A deployable telemedicine capability in support of humanitarian operations. Telemed. e-Health **13**(3), 331–340 (2007)
8. Vo, A.H., Brooks, G.B., Bourdeau, M., Farr, R., Raimer, B.G.: University of Texas medical branch telemedicine disaster response and recovery: lessons learned from hurricane Ike. Telemed. J. E-Health **16**(5), 627–633 (2010)
9. Garshnek, V., Burkle, F.M.: Applications of telemedicine and telecommunications to disaster medicine: historical and future perspectives. J. Am. Med. Inform. Assoc. **6**(1), 26–37 (1999)
10. Moughrabieh, A., Weinert, C.: Rapid deployment of international tele-intensive care unit services in war-torn Syria. Ann. Am. Thorac. Soc. **13**(2), 165–172 (2016)
11. Healthcare Management Article. www.healthdatamanagement.com/news/telemedicine-fills-the-gap-for-care-in-the-wake-of-harvey. Accessed 22 Feb 2018
12. U.S. Government Accountability Office Congressional Memorandum, 14th November 2017. https://www.gao.gov/products/GAO-18-108R
13. Simmons, S., Alverson, D., Poropatich, R., D'Iorio, J., DeVany, M., Doarn, C.R.: Applying telehealth in natural and anthropogenic disasters. Telemed. J. E-Health **14**(9), 968–971 (2008)
14. Crowther, M., Poropatich, L.: Telemedicine in the U.S. army: case reports from somalia and croatia. Telemed. J. **1**(1), 73–80 (2009). https://doi.org/10.1089/tmj.1.1995.1.73
15. Doarn, C.R., Merrell, R.C.: Telemedicine and the military. Telemed. J. E-Health **20**(9), 759–760 (2014)
16. Poropatich, R., Lappan, C., Lam, D.: Operational Use of U.S. Army Telemedicine Information Systems in Iraq and Afghanistan - Considerations for NATO Operations, 1 April 2010
17. Blanchet, K.D.: The U.S. army telemedicine and advanced technology research center (TATRC). Telemed. J. E-Health **12**(4), 390–395 (2006)
18. H.R. 4909 (114th): National Defense Authorization Act for Fiscal Year 2017, pp. 1–18 (2016)
19. Vasios, W.N., Pamplin, J.C., Powell, D., Loos, P.E., Riesberg, J., Keenan, S.: Teleconsultation in prolonged field care position paper. J. Spec. Oper. Med. **17**(3), 141–144 (2017)
20. Bennett, C.C., Hauser, K.: Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. Artif. Intell. Med. **57**(1), 9–19 (2013)
21. Nemeth, C., Blomberg, J., Argenta, C., Pamplin, J.C., Salinas, J., Serio-Melvin, M.J.: Support for salience: IT to assist burn ICU clinician decision making and communication. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1122–1126. IEEE (2015)

22. Lobach, D., Sanders, G.D., Bright, T.J., Wong, A., Dhurjati, R., Bristow, E.: Enabling health care decisionmaking through clinical decision support and knowledge management. Evid. Rep. Technol. Assess. (Full Rep.) **203**, 1–784 (2012)
23. Kawamoto, K., Houlihan, C.A., Balas, E.A., Lobach, D.F.: Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. Br. Med. J. Publ. Group **330**(7494), 765 (2005)
24. Nemeth, C., Pamplin, J., Anders, S.: Annual Report: A Cooperative Communication System for the Advancement of Safe, Effective, and Efficient Patient Care, pp. 1–232, February 2014
25. Nemeth, C., Anders, S., Strouse, R., Grome, A., Crandall, B., Pamplin, J.: Developing a cognitive and communications tool for burn intensive care unit clinicians. Mil. Med. **181** (5S), 205–213 (2016)
26. Salinas, J., Chung, K.K., Mann, E.A., Cancio, L.C., Kramer, G.C., Serio-Melvin, M.L.: Computerized decision support system improves fluid resuscitation following severe burns: an original study. Crit. Care Med. **39**(9), 2031–2038 (2011)
27. Ahmed, A., Chandra, S., Herasevich, V., Gajic, O., Pickering, B.W.: The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit. Care Med. **39**(7), 1626–1634 (2011)
28. Salinas, J., Drew, G., Gallagher, J., Cancio, L.C., Wolf, S.E., Wade, C.E.: Closed-loop and decision-assist resuscitation of burn patients. J. Trauma: Inj. Infect. Crit. Care **64**(4 Suppl.), S321–S332 (2008)
29. Pauldine, R., Beck, G., Salinas, J., Kaczka, D.W.: Closed-loop strategies for patient care systems. J. Trauma: Inj. Infect. Crit. Care **64**(Suppl.), S289–S294 (2008)
30. Lellouche, F., Mancebo, J., Jolliet, P., Roeseler, J., Schortgen, F., Dojat, M.: A multicenter randomized trial of computer-driven protocolized weaning from mechanical ventilation. Am. J. Respir. Crit. Care Med. **174**(8), 894–900 (2006)
31. Bibian, S., Dumont, G.A., Black, I.: Closed-loop target-controlled infusion systems: stability and performance aspects. Mil. Med. **180**(3 Suppl.), 96–103 (2015)
32. Linde, A.S., Thompson, D.M.: Robotic, Semi-Autonomous and Autonomous Medical Systems: Where will the soldier-medic fit in the future fight? http://smallwarsjournal.com/jrnl/art/robotic-semi-autonomous-and-autonomous-medical-systems-where-will-the-soldier-medic-fit-in. Accessed 22 Feb 2018
33. Yoo, A.C., Gilbert, G.R., Broderick, T.J.: Military robotic combat casualty extraction and care. In: Rosen, J., Hannaford, B., Satava, R. (eds.) Surgical Robotics. Springer, Boston (2011). https://doi.org/10.1007/978-1-4419-1126-1_2
34. The NATO RTO-HFM 182 Symposium: Advanced Technologies and New Procedures for Medical Field Operations, Essen, Germany, April 2010
35. Fisher, N., Gilbert, G.R.: Robotic and autonomous system technology enablers for the multi-domain battle 2030–2050. Small Wars J. (2017). http://smallwarsjournal.com/print/72735. Accessed 22 Feb 2018
36. Cummings, M.M.: Man versus machine or man+machine? IEEE Intell. Syst. **29**(5), 62–69 (2014)
37. Robots: Building New Business Models, 20 April 2016. https://www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/autonomous-systems-facts-and-forecasts.html. Accessed 22 Feb 2018
38. Innovation in Autonomous Systems: Royal Academy of Engineering, 22 June 2015. https://www.raeng.org.uk/publications/reports/innovation-in-autonomous-systems. Accessed 22 Feb 2018

Optimizing Team Performance When Resilience Falters: An Integrated Training Approach

Debbie Patton[1(✉)], Lisa Townsend[2], Laura Milham[2], Joan Johnston[1], Dawn Riddle[2], Amanda R. Start[3], Amy B. Adler[3], and Karen Costello[4]

[1] Army Research Laboratory, Aberdeen Proving Ground, Adelphi, MD, USA
{debra.j.patton.civ, joan.h.johnston.civ}@mail.mil
[2] Naval Air Warfare Center Training Systems Division (NAWCTSD), Orlando, FL, USA
{lisa.townsend, laura.milham, dawn.riddle}@navy.mil
[3] Walter Reed Army Institute of Research (WRAIR), Silver Spring, MD, USA
{amanda.r.start.ctr, amy.b.adler.civ}@mail.mil
[4] Army Resiliency Directorate (ARD), Arlington, USA
stephanie.k.costello@saic.com

Abstract. The U.S. Army strives to provide effective training for its soldiers. Part of this training is designed to build resilience, enabling soldiers and leaders to optimize personal readiness and performance in environments of uncertainty and persistent danger. Training complex tasks under high levels of stress is one way to support the development of resilience; another way is to train individuals in the use of specific resilience-based skills. Soldiers can use these skills not only to benefit their functioning but also to benefit the functioning of their teammates. The current paper reports on an innovative team-based approach to resilience training. Both the training content and training method provide novel approaches to addressing resilience in the context of high-stress scenarios. In terms of content, the training includes specific performance enhancement techniques that individuals can use to focus attention and optimize energy, and a method for intervening at the point-of-injury if a teammate experiences an acute stress reaction. In terms of method, the training includes classroom, virtual simulation, and live training. The resultant integrated training approach is Team Overmatch. This training milieu allows for the development, implementation, and evaluation of training modules fully embedded into tactical training. This paper discusses how innovative resilience strategies are integrated in a larger curriculum, including situational awareness, teamwork, and medical care, and how the training is being assessed in terms of knowledge and implementation.

Keywords: Resilience · Training · Military

Disclaimer: The views expressed herein are those of the authors and do not necessarily reflect the official position of the organizations with which they are affiliated.

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 339–349, 2018.
https://doi.org/10.1007/978-3-319-91467-1_26

1 Introduction

The Army recently reprioritized efforts to shift psychological skills training away from traditional didactic approaches and toward new approaches that integrate skills in tactical training. To codify this reprioritization, the Department of Defense (DoD) and Army has provided objectives and guidance.

First, the February 27, 2012 (updated October 2, 2013) DoD instruction (No. 64 90.09) for DoD Directors of Psychological Health includes language that highlights the importance of psychological health. Specifically, this document states that the goal is to:

> [i]nstitutionalize a culture and structure to **promote psychological health**, fitness, readiness, mission performance, and prevention of psychological health problems and mental health illness… and support efforts to **enhance psychological resilience**, not only to reduce injury and illness, but also to improve the success of the warfighter in the psychological performance domain. (DoD Instruction No. 6490.09, 2013) [1].

Second, an Army Regulation published in 2015 calls for combat and operational stress control universal prevention such as "surveillance and mitigation activities to reduce or avoid stressors and increase Soldiers' tolerance and resilience to severe stress" (Army Regulation 600-63 (Army Health Promotion), 2015) [2].

Third, The HQDA EXORD 086-16 (Human Dimension 22 DEC 15) reinforces the importance of optimizing human performance: "The Army has the capability and capacity to **optimize the human performance** of every Soldier and civilian in the total force to improve and thrive in the strategic environment of 2025 and beyond." In addition, the objective is to:

> [e]nhance Soldier and army civilian health and physical readiness through an **individualized comprehensive training system** …[and to] transition and implement psychological, social, and neurological science and other technological advancements to **train teams of army professionals to improve and thrive in ambiguity** and chaos [3].

Taken together, guidance pushes the Army to focus on "training on demand" formats that are amenable to the operational landscape. Previously, Squad Overmatch (SOvM) was developed as a training platform to implement and evaluate an operator-centered approach to resilience and performance training. This platform serves as a basis for developing new integrated training that optimizes team resilience.

1.1 SOvM Study

SOvM was a multi-year effort focused on designing a training platform to increase soldier performance under stress. Combining Army and Navy investment, SOvM involved the implementation of an integrated training approach [4]. At its core, SOvM relied on Stress Exposure Training, a training model that incorporates three stages: (1) an initial stage, in which didactic information is provided regarding stress and stress effects; (2) a skills training phase, in which specific cognitive and behavioral skills are acquired; and (3) the final stage of application and practice of these skills under conditions that increasingly approximate the criterion environment [5, 6]. This methodology emerged in 2015 when over 100 team members from across branches of service, other government agencies, industry, and academia collaborated to examine

improve training effectiveness for military personnel. In 2016, a scientific study tested the effectiveness of SOvM's integrated training approach on acquiring new knowledge and understanding concepts, shifting attitudes, and improving proficiency.

In the 2016 SOvM study, participants included eight Army squads, each augmented with a medic. The research team collected measures of learning, cognition, attitudes, and performance. Squads in the experimental condition participated in a three-and-a-half day integrated training curriculum consisting of classroom instruction, virtual simulation-based training, and three live mission training scenarios in the outdoor McKenna urban training facility. This criterion environment was augmented with live role players and advanced simulation technologies (e.g., non-pyrotechnic explosives, interactive avatars, and medical mannequins). Squads in the control condition participated in one day of live training on the M2 and M3 scenarios with the same role players and technologies, but without the classroom instruction and virtual simulation-based training.

Squads were randomized such that half were placed in the integrated training condition and half were placed in the control condition. Soldiers in both conditions reported high levels of confidence in their own ability and their squad's ability to perform well prior to two of the live scenarios [7].

Furthermore, regardless of condition, SOvM was well accepted. Soldiers in both conditions demonstrated a strong willingness to take part and considered participating in the training to be important. In addition, training as a whole appeared successful. Study results showed that all participants increased their basic knowledge of Resilience and Performance Enhancement (RPE), as well as Tactical Combat Care (TC3), Advanced Situation Awareness (ASA), Team Development (TD), and After Action Reviews (AAR) content areas. Soldiers in both conditions were also positive about their teamwork processes, efficacy, cohesion, and performance, and in general, these attitudes increased in positivity over time. Soldiers in both conditions also rated the AAR climate following the live scenarios as supportive and positive [4].

Overall, the integrated training condition appeared to be more successful than the control condition. Squads in the integrated training condition performed significantly more TC3, ASA, and TD tasks than the control condition squads [4]. Soldiers in the integrated training condition also demonstrated significantly more effective behaviors during the AARs than the control condition squads.

SOvM Study and Resilience. In terms of resilience and performance enhancement, significantly more gains in resilience knowledge were found for the integrated training condition compared to the control condition. Given that these skills were essentially internal skills, there was no objective measurement of how these skills were implemented.

Nevertheless, following training, soldiers in the integrated training condition rated themselves as highly competent to highly proficient, whereas soldiers in the control condition rated themselves as advanced beginner to moderately competent after the live exercises. Given this difference in self-rated proficiency and the improvement in knowledge scores, the resilience and performance enhancement materials showed promise.

Based on the findings from the 2016 SOvM study, a follow-on program has been developed: Team Overmatch (TOvM). While the training as a whole has been updated to account for lessons learned from SOvM, the resilience and performance materials have been expanded to include a new component focused on team resilience. Not only are the resilience and performance skills discussed in terms of how they can be used with teammates, but a new training module has been developed that introduces what individuals can do to intervene if a teammate experiences an acute stress reaction. This training module provides unique information about how to maintain team functioning under extraordinarily stressful conditions.

While SOvM focused on individual resilience, TOvM will focus on enhanced individual resilience as well as team resiliency training. This training will provide the necessary skills for soldiers to recognize the signs of stress in themselves and their squad mates. Not only will they be able to recognize these signs (increased breathing, rapid heart rate) but they will know how to interpret them and ultimately mitigate the effect to ensure mission success.

2 The Training Challenge

Soldiers and Marines make life and death decisions under extreme physical and psychological conditions. They are challenged with managing these stressors while maintaining high levels of collective performance. When a team member is injured and unable to perform, it is up to the team to serve as first responders and provide critical care at the point of injury. This care needs to be conducted quickly and effectively under complex and potentially traumatic conditions in order to keep the team functioning and to sustain the injured so that they can be transported for further care. Yet just as team members need to know how to apply tactical combat care to the physically wounded, they also need to be proficient at managing cases of acute psychological stress in their team members (Fig. 1).

3 Objective and Approach

TOvM trains individuals in providing an immediate intervention in the event that a team member experiences an acute stress reaction. This intervention is designed to be provided at the point of injury (either during or immediately after a combat-related event) and is expected to immediately return the affected battle buddy back to functioning, and thus sustain the team's fighting capabilities.

This intervention is based on a protocol developed by the Israeli Defense Force (IDF). The IDF protocol, YAHALOM, is embedded in a larger program called "Magen" (which means "Shield" in Hebrew). Magen offers soldiers training in a 5-step process to use if an individual team member experiences an acute stress reaction. This 5-step process is designed for rapid delivery in the field and should take 40–60 s to apply. Various versions of training in Magen have been piloted; the current version is taught in a one-hour class. This class includes a short video [8] and emphasizes practicing delivery of the intervention. The content has been so well received by the

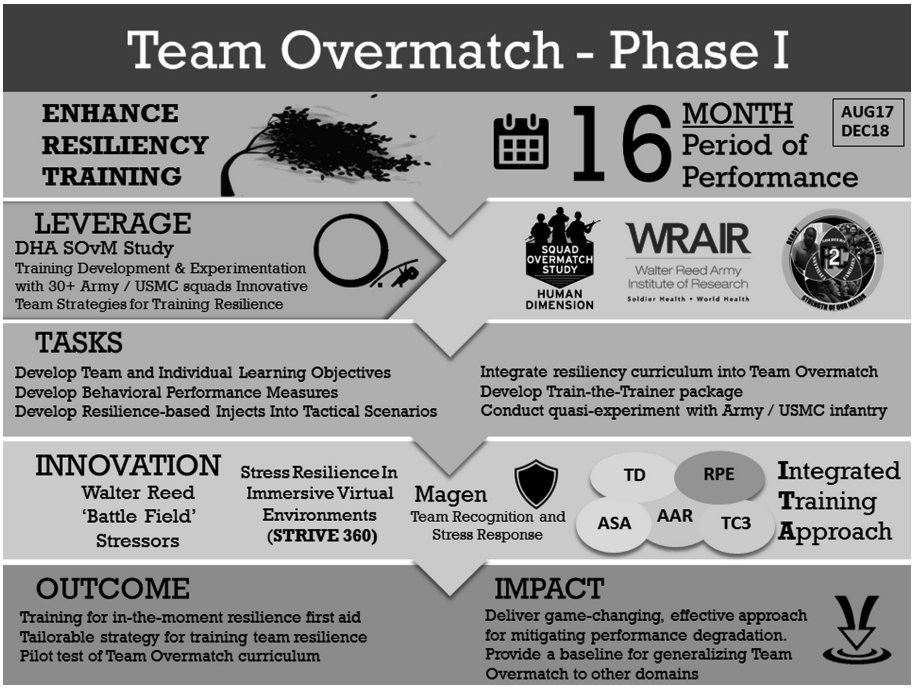

Fig. 1. Team Overmatch overview

IDF that as of 2017, Magen has become part of mandatory training that all recruits receive and all ground forces must demonstrate proficiency in the technique. Initial feedback from the field is positive. Anecdotal evidence from troop commanders indicates the training has already proved useful, and follow-on surveys are being administered (Lt. Col. Vlad Svetlitzky, personal communication).

The Magen material was originally shared in the 2017 US-Israeli Shoresh meeting, a biennial event designed to facilitate cooperation on medical research between the militaries of the two nations. Walter Reed Army Institute of Research (WRAIR) scientists serving as U.S. representatives to Shoresh discussed the potential for adapting this technique with the IDF. Following the meeting, WRAIR worked with their IDF counterparts to develop a US version of the Magen protocol.

Part of this discussion entailed the importance of recognizing the signs of an acute stress reaction. Magen, the IDF version of the intervention, covers 5 steps. In contrast, it became clear that the initial step of recognizing an acute stress reaction was an implied task. Thus, for the sake of clarity, the U.S. version of the training includes 6 steps. The first step is to identify an acute stress reaction in a team member; the remaining steps are equivalent to the Magen protocol.

The U.S. version of the training is called iCOVER. This acronym spells out the steps that a soldier should take to address an acute stress reaction in a team member (Fig. 2).

Fig. 2. iCOVER steps

The 6-step protocol is defined by the following procedures:

(1) **Identify a buddy in need:** look for signs of an acute stress reaction. While there are different types of how an acute stress reaction can be expressed, the common denominator is that the person is no longer functioning and this lapse is not due to physical injury
(2) **Connect:** break through the person's cognitive daze by making eye contact, squeezing their hand, and asking them to squeeze your hand back
(3) **Offer commitment:** break through the person's dissociation through a simple phrase reminding them that you are there
(4) **Verify facts:** ground the person in the present moment by asking 2 to 3 concrete questions related to the immediate situation
(5) **Establish order of events:** continue orienting the person through a brief situational report; provide one short sentence each about what happened, what is happening and what will happen
(6) **Request action:** restore the person by giving them a simple, externally-focused task that is relevant to the situation.

Training involves a brief explanation as to how the brain functions during an acute stress reaction, introducing the concept of an amygdala hijack, how iCOVER prompts the thinking part of the brain back into action, and what behaviors should be avoided during the intervention (e.g., shaking or shouting at the individual, delving into emotional topics).

The concepts behind Magen and iCOVER are based on a theoretical framework proposed by Hantman and Farchi [9]. This framework guides recommendations for how first responders should intervene during an acute stress reaction. Each of the model components are designed to help shift the affected person away from a position of helplessness and toward a return to functioning. The framework is defined by the six Cs: (1) cognitive, (2) communication, (3) challenge, (4) control, (5) commitment, and (6) continuity.

The first C, "Cognitive", is at the center of the model. "Cognitive" reflects the focus on returning the individual back to functioning by activating the prefrontal cortex. Each of the Cs help the individual's frontal cortex regain control from the limbic system and reduce disorientation. In the case of "Communication", verbal communication is used as a method to restore frontal cortex functioning. In the case of "Continuity", the goal of ordering events is to prompt reestablish the logical, chronological sequence.

The final three Cs ("Challenge", "Control", and "Commitment") are adapted from the work on psychological hardiness. These characteristics of hardiness are associated with more effective coping under stress and provide proactive ways to manage and perform under high-stress conditions [10]. In the framework, the first responder should intervene by leveraging strengths associated with hardiness. "Challenge" refers to prompting an individual to complete simple, specific tasks in order to increase their sense of self-efficacy and return to functioning. Traditionally, offering options enables the individual to reassert "Control" (although in the case of iCOVER, control is reasserted through providing an opportunity for the individual to respond to specific questions). By offering a "Commitment" to the individual, the first responder is reducing the individual's sense of isolation and dissociation.

Taken together, the six Cs framework provides a theoretical underpinning as to why Magen and iCOVER are expected to be effective. Furthermore, the essential steps of iCOVER are designed to be immediate and simple, consistent with the concepts of PIES [11] already used in the Army doctrine [9].

4 Hypotheses

The current study leverages the TOvM platform to assess the impact of training on ASR knowledge, recognition, treatment, and attitudes. In this quasi-experimental study, eight squads will be randomized into one of three groups: iCOVER Traditional (Experimental Group 1), iCOVER Simulation (Sim) (Experimental Group 2), and training as usual (Control Group).

To practice the recognition of and response to ASRs, the current Resilience Practical Application (within the MAGEN program) requires students to act out symptoms of ASRs and to verbally execute the ASR treatment protocol in groups. This traditional training approach will be utilized as iCOVER Traditional and compared to an alternative training approach that leverages the simulated environment to increase the fidelity of the practice. The alternative Resilience Practical Application (iCOVER Sim) will allow individual students to practice identifying simulated squad members who are experiencing an ASR and treating the ASR by selecting treatment options and

verbally executing the ASR treatment protocol in the simulated environment. The Control Group will not receive any training and practice in iCOVER.

There are two primary sets of comparisons in the current study. First, both Experimental Groups are expected to result in better ASR knowledge, recognition and treatment than the Control Group as measured by a Team Resilience Knowledge and Comprehension Test, a Team Resilience Practical Test, and observational ratings provided by Subject Matter Experts during live scenarios.

Second, the two different iCOVER Groups will be compared: iCOVER Traditional, which uses didactic and in-class practice with squad members, and iCOVER Sim, which uses didactic and simulation practice with avatars. The expectation is that iCOVER Sim Group will be more effective in cue recognition on a Team Resilience Practical Test, but those in the iCOVER Traditional Group will be more effective in providing treatment more quickly and accurately than those in the iCOVER Sim Group.

During training, the iCOVER Traditional Group engages in treatment verbally with squad members whereas the iCOVER Sim group will perform treatment through selecting a response from a list and executing treatment through avatars.

The use of a computerized virtual environment to provide consistent, repeatable cue patterns of ASR for avatars, instructional strategies (granular feedback provided on cue pattern recognition), and structured, consistent learning opportunities is expected to lead to more effective ASR cue recognition. While the fidelity of the most characteristics of the 'patient' team member is naturally higher in a traditional person-to-person setting, the fidelity of the ASR cues are expected to be lower. Soldier trainees, as patients, may not be able to accurately convey the complexity of an ASR. In this regard, avatars can provide more accurate timing and portrayal of ASRs.

5 Measures

Table 1 below lists and describes the subjective, objective, and performance measures to be collected during the TOvM study.

6 Method

The current study is to be conducted at Ft. Benning, GA. A combination of Army and Marine Squads were asked to participate in the study. The study will be covered by a protocol approved by the NAWCTSD institutional review board. All study participants will complete informed consent prior to participation.

Squads will be randomly assigned into one of the three conditions (iCOVER Sim, iCOVER Traditional, or Control). The **Control** group will receive the standard TO curriculum (ASA, TC3, RPE, and TD), with no team resilience (iCOVER training). In addition to the standard TOvM curriculum, the **iCOVER** Traditional group will also receive the team resilience classroom curriculum (iCOVER) and will complete the traditional practice of working in groups to switch off between exhibiting, observing, and treating ASRs. **iCOVER Sim** will receive the same classroom training as the

Table 1. Measures and descriptions of subjective, objective, and performance collected during the TO study.

iCOVER knowledge and comprehension test	A multiple choice test to assess participant knowledge on resilience learning objectives
Motivation	On a scale of 0 to 100 for each question, respondents rate the importance (1 item) of and their willingness (1 item) to successfully utilize the training. A score closer to 100 indicates greater importance and willingness
Multiple affect adjective checklist – revised (STATE)	Participants select terms that describe how they "feel right now" or "how you felt during the training you just completed"
Observable behaviors performance measures for individual and team Resilience iCOVER	Behaviorally Anchored Rating Scales (BARS) are scales used to assess performance based on observable behaviors. Subject matter experts rate each behavior on a 5-point scale (1 = Beginner; 5 = Highly Proficient) The Targeted Acceptable Responses to Generated Events or Tasks (TARGET) Checklist is a structured observation checklist method used to design the SOvM scenarios for both virtual and live training exercises. Events were identified that were expected to elicit demonstration of specific resilience and performance enhancement skills within the skill areas and team resilience; acceptable responses to each of the events were determined a priori
Psycho-physiology	Respiration Rate will be measured to assess breathing patterns associated during specific time points in the scenarios. Heart Rate will be measured to assess physical activity during specific time point in the scenarios. Will be measured using the Equivital EQ 2 system. Inter-beat-interval will be measured as a measure of the cognitive effort put forth during the scenarios
Resilience and performance enhancement post-event Assessment Team resilience skills post VBS and live assessment	Respondents reflect on the training exercise, indicating whether anyone on their team (themselves included) experienced an acute stress reaction or used any of the tactical stress care skills
iCOVER self-reported skill proficiency assessment	Respondents rate their current level of skill (beginner, advanced beginner, proficient, and expert) on learning objectives

(*continued*)

Table 1. (*continued*)

iCOVER knowledge and comprehension test	A multiple choice test to assess participant knowledge on resilience learning objectives
Team resilience practical test	Video-based situational judgment tests; participants will be shown a series of videos of actors who may be exhibiting signs of an Acute Stress Response (ASR) Participants then must decide whether the actor is experiencing an ASR and requires treatment. If the decision is made to treat, participants will complete the ASR treatment protocol. Afterwards, participants will be asked to identify which ASR cues/profiles were exhibited by the actor. Time and accuracy of recognition of and response to ASRs will be assessed
Team action Processes Attitudes: action processes, cohesion, efficacy, team resilience, and performance	Respondents rated 1 to 5 on a Likert-type scale the degree they agreed with statements that asked how well they thought their team performed together during the mission just completed. Processes such as coordination of actions and effective communication are probed. A higher score indicated better rated performance

iCOVER group but will engage in a simulated activity (with a simulated fire team) in which they are placed in a combat simulation and observe before, during, and after a stress event. During the event, one of the simulated fire team members will exhibit an ASR. Each trainee will work individually to recognize which simulated team members exhibited ASRs, and will select and treat stress care treatment options (Table 2).

Table 2. Procedure representation for the experiment

	Control (without iCOVER)	iCOVER traditional	iCOVER Sim
Baseline surveys	X	X	X
Training on traditional SOvM	X	X	X
Team resilience Sim test (pre)		X	X
iCOVER traditional			X
iCOVER Sim			X
Team resilience Sim test (post)	X	X	X
Simulated and live scenarios	X	X	X
Post-training surveys	X	X	X

7 Discussion

Squad Overmatch has been successfully implemented at several Army training sites. Based on lessons learned, the enhanced individual and team resilience modules will be integrated into the SOvM package, and will be included as part of the 2018–2019 operational implementation and transition efforts at additional Army sites. Furthermore, follow-on WRAIR studies are planned to assess the experience squads have with ASRs and implementation of iCOVER in real-world deployment contexts.

References

1. DoD Instruction No. 6490.09 (2012)
2. Army Regulation 600.63: Army Health Promotion (2015)
3. The HQDA EXORD 086-16: Human Dimension, 22 December 2015
4. Townsend, L., Milham, L., Riddle, D., Phillips, C.H., Johnston, J., Ross, W.: Training tactical combat casualty care with an integrated training approach. In: Schmorrow, D.D.D., Fidopiastis, C.M.M. (eds.) AC 2016. LNCS (LNAI), vol. 9744, pp. 253–262. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39952-2_25
5. Johnston, J.H., Cannon-Bowers, J.A.: Training for stress exposure. In: Driskell, J.E., Salas, E. (eds.) Stress and Human Performance. Erlbaum, Hillsdale (1996)
6. Johnston, J.H., Gamble, K., Patton, D.: Stress Assessment of Military Personnel during a SET training program (in progress)
7. Squad Overmatch (SOvM) Phase II Final Report: Optimizing Warriors…Achieving Squad Overmatch…Saving Lives. PEO STRI, Orlando (2017)
8. Magen 5 Step Process. https://www.youtube.com/watch?v=t-QZgZd-PJ4
9. Hantman, S., Farchi, M.: From helplessness to active coping in Israel: psychological first aid, Chap. 32. In: Schott, E., Weiss, E.L. (eds.) Transformative Social Work Practice. Sage Publication (2013)
10. Maddi, S.R.: Hardiness: the courage to grow from stresses. J. Posit. Psychol. **1**(3), 160–168 (2006). https://doi.org/10.1080/17439760600619609
11. Lewis, S.J.: Combat stress control: putting principle into practice. In: Adler, A.B., Castro, C.A., Britt, T.W. (eds.) Military Life: The Psychology of Serving in Peace and Combat. Operational Stress, vol. 2, pp. 121–140. Preager Security International, Westport (2006)

A Human Perspective on Maritime Autonomy

Tore Relling[1(✉)], Margareta Lützhöft[2], Runar Ostnes[1], and Hans Petter Hildre[1]

[1] Norwegian University of Science and Technology, Aalesund, Norway
{tore.relling,runar.ostnes,hans.p.hildre}@ntnu.no
[2] Western Norway University of Applied Sciences, Haugesund, Norway
mhl@hvl.no

Abstract. As for all of the transport segments, autonomy is gaining increasing interest by researchers and for development in the maritime industry, and introducing autonomy is expected to create new possibilities to increase efficiency and safety. Autonomy could lead to drastic changes in roles and responsibilities for involved agents (both technical systems and humans), and these changes will be an important driver for changing the rules which regulate the responsibilities of the involved actors in the maritime domain. This paper suggests a perspective of autonomy as a process of change as opposed to a defined state. The paper discusses three areas that warrant more attention in the development of autonomy in navigation in the maritime industry. Firstly; rather than the traditional reductionist safety models, it considers complexity in maritime systems with increased autonomy and explore systemic safety models to amplify positive human performance variability. Secondly; it argues that humans will be important also in systems with increased autonomy, and discusses the human involvement on *strategic, tactical* and *operational* levels. Thirdly; it discusses the importance of defining the concepts responsibility, authority and control from the perspective of *humans*, rather than that of the *vessel*.

Keywords: Human centred design · Maritime autonomy · Methods of control
Responsibility · Authority · Remote operations

1 Introduction

There is a belief that the future of maritime industry will see an increased use of autonomous solutions. The International Maritime Organisation (IMO) decided to include the issue of marine autonomous surface ships on its agenda in their Maritime Safety Committee in June 2017 [1], which is a solid sign of the importance of the topic. The dissension grows when discussing what autonomy *is*, and *how* it will affect maritime industry.

In contrast to what seems to drive the development of maritime autonomy, autonomy as a concept has no direct link with technology. Autonomy has been used since the early 17th century and comes from the Greek *autonomia*; from *autos* "self" + *nomos* "law". The Oxford Dictionaries explains autonomy as *the right or condition of self-government*, and further as the *freedom from external control or influence:*

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 350–362, 2018.
https://doi.org/10.1007/978-3-319-91467-1_27

independence [2]. In maritime history, we can easily find examples that fall under the definition of autonomy. The explorers who sailed into the unknown more than 700 years ago were self-governed from when they left the harbour, with no shipping company or authority to influence their choices. Another example would be fishermen sailing from Europe to the Antarctic 100 years ago, staying away for months, also with little or no influence from their owner in the home country. However, arguing that the maritime domain was more autonomous in the past than what we expect of the future might not be very helpful to reduce the dissension about what autonomy is, but it does show that we might interpret autonomy differently than the original meaning of the term. This paper discusses what autonomy is today, and further elaborates on the human role in the development of autonomy in maritime systems. A system is defined as "an assemblage or combination of functionally related elements or parts forming a unitary whole" [3]. The parts of the system could be technical or human agents, where an agent is defined as a "'thing' in an environment with capacities to sense states and effect aspects of the environment" [4].

2 What Is this Thing Called Autonomy?

Apparently, the term autonomy is used differently in colloquial language than in the technical definition. In addition it is interpreted in various ways both in the maritime industry and other industries. Automation and autonomy are often used interchangeably in the discussion of the technological development in the maritime industry. Parasuraman and Riley [5] define automation as "*the execution by a machine agent (usually a computer) of a function that was previously carried out by a human*". Some attempts have been made to create a more distinct difference between automation and autonomy by transforming *Levels of Automation* (LOA) developed by Sheridan and Verplank in 1978 into various *Levels of Autonomy* [6]. The attempts have neither been able to create a common understanding of the distinction between automation and autonomy, nor reach consent on whether using *levels* is suitable for describing concepts. Endsley [7] states that using *levels* is beneficial to communicating design options to stakeholders in both automated and autonomous systems, and especially for explaining the continuum between fully manual and fully automated. However using levels to describe autonomy has been criticised for being unidimensional and not reflecting the real problems in developing systems, and for not allowing for dynamically changing functions in various contexts [8]. Parts of this criticism is rejected by Kaber [9] who claims that it confuses automation with autonomy, and he states that the *"research focused on one construct (i.e automation as a technology) yet made criticism from the perspective of another (i.e autonomy as a state of being)"*. To distinguish between automation and autonomy is difficult, and the Society of Automotive Engineers (SAE) has decided to put the term "autonomous" in their section of deprecated terms. They state that the term has become synonymous to automation since the use of it has broadened to not only encompass decision-making but include the entire system functionality [10]. Parasuraman [11] presented his version on the Levels of Automation, where the highest level of automation is defined as the computer deciding everything, acting autonomously, and ignoring humans. According to this definition,

autonomy is gained at the highest level of automation, which could support the criticism of unidimensionality of the concept of autonomy.

There is no unified definition of what autonomy *is* in the context of developing systems. The challenge is apparently to define *a state of being* which includes all aspects of the benefits and complications within human and technology interaction in various contexts and changing scenarios. The reason for introducing autonomy in a system is to improve the performance of the system; hence, increased autonomy will not be a goal in itself. Relating autonomy to a change process based on system needs will give various answers of what autonomy is from system to system, and will vary over time. This paper suggests that autonomy, similar to Parasuraman and Riley's definition of automation, is a process rather than a defined state. Similar to automation, the use of technology is a main component in the change process, and autonomy especially implies the use of digitalization such as sensor fusion, control algorithms and communication and connectivity [12]. The other main component in the change process is the degree of involvement of humans in the operations, and the aim to reduce human presence in dangerous and hostile environments [13, 14].

The similarities between automation and autonomy are many, and several of the challenges Dekker and Woods [15] discuss about how humans and technology get along in highly automated systems, is the same challenges more recently discussed considering autonomous operations [16]. With an interchangeable use of the two terms, it is tempting to follow SAEs path by discarding autonomy as a term and stick to the use of automation. However, there is one solid argument for keeping autonomy as a term. While automation could span from a simple exchange of functions between humans and technology, to highly automated systems comparable with autonomy, autonomy implies a *significant* change to the system. This significant change also imply that understanding all effects of changes are more complex. Lee [17] describes autonomy from a network perspective, where automation and people are nodes in a network that produce emergent properties that are not predictable by looking at the nodes in isolation.

The main difference between autonomy and automation is therefore that autonomy implies a significant change to the system where emergent properties are expected to affect the performance of the system. This perspective takes into account that there is no single solution on what to change, nor is there a unified end-state of the change process. It opens for autonomy being different from system to system, varying over time and being affected by the context. This approach might seem complex and a rejection of the existing research on autonomy, but the purpose is the opposite. By agreeing on a significant level of the change, it is possible to discuss how complexity and emergent properties will affect performance of a system with increased autonomy. This will not limit autonomy to a few defined factors, but will be dynamic and adaptable based on context and the previous state/s of the system.

3 Humans and Autonomy in Maritime Systems

There are two main directions of development in the maritime domain that fall under the above-mentioned perspective of autonomy. One is the development of self-navigating vessels[1], and the other is remotely operated vessels. The similarity is the aim to reduce human presence on the bridge or even to reduce human presence on the entire vessel. Both directions could cause a *significant change* to maritime systems when humans are moved away from the bridge, to a position on shore (or elsewhere). The difference between remotely operated vessels and self-navigating vessels is *how* the humans are involved by remotely operating the vessel or taking a role of controlling the operations by monitoring or supervising from a distance. As discussed later in the paper, the two directions will create different challenges to overcome.

Increased autonomy is expected to provide benefits such as less environmental emission and increased efficiency and safety [18]. This paper is limited to discussing the challenges related to the effect on safety, and the navigation function, where the humans operate, monitor or supervise navigation from shore. The paper discusses the human involvement in three areas; a systemic approach which advocates for human as strengthening the system, the human role on strategic, tactical and operational levels and finally how humans will be responsible and remain in control in systems with increased autonomy.

3.1 Humans Will Strengthen the System

Increased safety is an expectation and a motivation for developing solutions with increased autonomy. It is claimed that increased safety will be achieved by reducing the likelihood of *human error* when introducing more autonomy [12, 19]. There is no reason to dispute the fact that reducing human error will increase safety, but it is necessary to be wary of the belief that introducing more technology is coherent to reduction of human error, and Bainbridge "Ironies of Automation" [20] is still as valid today as it was 35 years ago [21].

Since autonomy entails significant changes to systems, it could be compared to what Boy [22] defines as a typical twenty-first century problem, with *"global and non-linear"* problems where the number of components and interactions are far larger than in the twentieth century where the problems were *"local and linear"*. He claims that *complexity science* will be one of the most important sciences to understand these challenges. The term complexity science was introduced by Anderson in 1972 [23], and he has later defined complexity science to be:

> *"(..)the search for general concepts, principles and methods for dealing with systems which are so large and intricate that they show autonomous behavior which is not just reducible to the properties of the parts of which they are made"* [24].

He describes the developing discussion within physics science, where physics science has been subject to reductionism, in trying to reduce complexity to simplicity

[1] Self-navigating vessels refers to the development of vessels that are able to follow a pre-defined route and have a capability to detect and avoid obstacles en route.

by explaining the construction of the universe in smaller and smaller entities. This traditional modelling of decomposition into structural elements has been challenged for a time and Rasmussen [25] described the problem of reductionism by *"all work situations leave many degrees of freedom to the actors for choice of means and time for action"* and argued for a functional abstraction on a higher level rather than structural decomposition. Anderson [24] states that there is a growing interest to develop complexity out of simplicity. The new perspective highlights the importance of *emergent properties,* where emergence implies that there are new properties that did not pre-exist or were expected or pre-programmed in the system. Safety could be regarded as an emergent property, and safety is created from the interaction of system components [26]. This means that we need to understand and identify emergent properties to assess safety in complex socio-technical systems. When safety is an emergent property in complex socio-technical systems it is necessary to understand what affects safety in other terms than a traditional reductionist perspective. Since the complexity in socio-technical systems leads to gradually more intractable systems and work environments, it is stated that *performance variability* is a prerequisite for functioning systems [27]. It is especially important to study humans at work to understand the nature of performance variability, with the intention not to limit by constraining how people work, but by addressing reasons for variability, identifying ways to monitor variability and understanding consequences and means to control variability [28].

A fundamental change in how the maritime industry assesses safety is needed, considering the increased complexity. The immediate risk is that we choose an approach which limits the focus to the individual vessel alone. We could measure safety based on an assessment of technical components or isolated processes and by verifying that they are covered by an autonomous solution. Such an approach tends to use traditional risk assessment methods, but is a reductionist approach which does not take the whole system into consideration [29, 30]. A limited focus on the vessel alone could result in the conclusion that autonomy is as safe, or safer, than shipping is today. This may not be a correct safety assessment from a systems perspective, since vessels do not operate as single standalone vessels. A safety assessment needs to be elevated from the perspective of a single vessel (as a sub-system in a system) to an assessment of safety in a both a system and a system-of-systems perspective where the vessel interacts with other vessels, with Vessel Traffic Service Centres (VTS), with marine pilots and several other systems in the maritime industry as well as systems from other industries. Only by using this perspective, it will be possible to discuss safety as an *emergent property,* and find out more about what affects *performance variability* due to changes in the system.

Sheridan's statement *"overall design of large–scale human-automation systems (for example, design of modern airplanes or air traffic systems) will continue to be a matter mostly of experience, art, and iterative trial and error"* [31], indicates that there are difficulties in identifying challenges in novel systems. However, we must acknowledge the importance of emergent properties, and especially the ability to amplify positive performance variability and reduce negative performance variability. In recent years several systemic safety models have gained interest such as Functional Resonance Analysis Model (FRAM) [32], Systemic Theoretic Accident Models (STAMP) [26], AcciMap [25] and EAST broken-links approach [30]. The systemic

safety models aim to address the limitations of more traditional cause-effect chain models, which focus on blame and tend to search for single root causes after accidents. Systemic safety models create models that consider the entire socio-technical system, and relationships between parts of the system [26]. This paper will not elaborate on which systemic models are best suited for considering complexity in maritime operations with increased autonomy, but as all of them have a holistic approach to safety, they could all be candidates for assessing safety in designing novel systems.

The ownership of the challenge of assessing safety in a system perspective is not obvious. It remains to be seen if, and how, IMO, governmental authorities, shipping companies, technology groups or other stakeholders assume the responsibility to select methods with which to choose a systemic safety approach. Taking this responsibility implies performing large-scale testing and identifying new agents and interactions to assess safety. The other option, which is not sustainable and should be avoided, is to take the easy way out; concentrating on sub-systems and components and assuming there are no other solutions available.

3.2 Humans Will Be in the Loop, but There Will Be New Loops

Increased autonomy in navigation will impact the role of the master and will be a paradigm shift in maritime industry. To change the role of the master constitutes a drastic change to the maritime industry, not only in how to operate, but also regarding internal responsibilities for the state of the vessel, and external responsibilities towards other actors in the industry and society. A hasty and simplified approach to understanding the consequences of changing this role, could fail to uncover important aspects that affect safety, as the role of the master has a long tradition and includes many formal and informal tasks.

Autonomy aims to reduce human presence in dangerous and hostile environments, but in the maritime industry as for most of the other industries, this does not imply a total removal of humans in the system. Autonomy in the navigation function would most likely lead to relocating the humans from the bridge to a position on shore, and it is important to understand which role humans will have in such new systems.

When designing new concepts it is essential to understand *why* we need humans in the (new) loop. To create this understanding, we suggest using the terms *strategic*, *tactical* and *operational* levels to describe types of decisions in the system and where to expect change. The three terms do not have unified definitions but were initially introduced in the military literature [33]. Today they are widely used, for instance in on-road automated system development [10] where they are based on behavioural models and generalised to the problem solving task of the driver on three levels (strategical, tactical and operational) of skill and control [34]. Discussing maritime specific characteristics on each level could be a step towards understanding the human role in future maritime systems. The literature does not concur on the order of *tactical* and *operational* level, but in military doctrine the tactical level is often referred to as the lowest level of operation and operational level is the mid-level [35, 36]. In the Contextual Control Model Hollnagel describes the strategic level as being focused on the high-level goals, and is followed by the tactical level of beyond the present [37].

Coherent with this model, the paper choose to define the tactical level as subordinate to the strategic level.

Strategic level:
Strategic decisions set objectives for the organization as a whole, relatively long-range objectives, and formulate policies and principles intended to govern selection of means by which the objectives specified are to be pursued. [33]. Strategic decisions would fall under the three dimensions Boy [22] describe as important for an organisation; safety, efficiency and comfort.
Tactical level:
The tactical level could be described as the criteria derived from the goals set at the strategic level [34]. In navigation this would be both long-term and short-term planning on how to act, such as planning and deciding on the route, or weather routing during the voyage.
Operational level:
The operational level is the imminent response within strategic and tactical boundaries to occurring events. In navigation this would be the choice of whether to alter speed or heading in response to the immediate surroundings.

As illustrated initially in this article, autonomy is difficult to describe as a state-of-being. Since autonomy is a process of change, the role humans will play in the system will also change over time. However, the change of the role of humans is initially expected to occur mainly on the *operational* level, to a lesser extent on the tactical level and is not expected to affect the strategic level. The main reason for this expectation is based on the acknowledgement of maritime socio-technical systems being intractable and such systems work because people are able to adjust what they do [27, 38]. In particular this applies to managing the constant trade-off between objectives on the strategic and tactical levels, for example the balance between safety and efficiency, which is an area where humans are still superior to technology [27]. The prediction that change will occur mainly on the operational level will probably change, and a natural development would be that a successful implementation of increased autonomy on the operational level triggers an investigation of possible benefits of autonomy on the tactical level.

We do know that there will be humans in the loop, and even though the operational level will gain more autonomy, there will still be humans to take strategic and tactical decisions. *Why* we need humans in the new loops is fundamental to the understanding of the importance of taking humans into account in the entire concept design.

3.3 Humans Will Be Responsible and Will Remain in Control

The final challenge presented in this paper is to create an understanding of *how* the humans should be involved and kept in the loop. The similarities between automation and autonomy are many, and one similarity is the challenge of how to optimize the sharing of functions between humans and technology based on human strengths and weaknesses. In systems engineering a function refers to *"a specific or discrete action (or series of actions) that is necessary to achieve a given objective*" [3]. These functions are derived from the system requirements in a hierarchy where top-level functions are broken down to second-level functions and further to lower level functions. In the conceptual phase of systems design the purpose is to develop a top-level system

architecture and initially to identify *what* needs to be accomplished, and less focus is put on how to accomplish it [3].

A widespread concept in Human-Automation Interaction (HAI) is to combine Levels of Automation (LOA) with *function allocation* which uses a four-stage model of HAI; information acquisition, information analysis, decision selection and action implementation [39]. Each of these four stages is described in a continuum (the Levels of Automation) ranging from no technological involvement to a complete technological ownership. The concept is criticized for not taking into account the complexity of operating environments which leads to imprecise and unreliable predictions, which again leads to a concept which is difficult to apply in practice [40]. There is an on-going discussion between the defenders of the concept and those who challenge the concept. Both sides seem to agree that there are weaknesses such as difficulties in predicting human behaviour and imprecise behavioural constructs, and that there is a need for a more concise operational definition of the concept [9, 40, 41]. The solution to the problem is more contested, in that the defenders of the concept are suggesting an evolution of the model to get a more accurate and precise prediction of human-automation system performance [9], and the opponents are suggesting to leave the LOA paradigm entirely [40].

Those involved in the development of autonomy in the maritime industry need to pay attention to the limitations of the existing models, and the on-going debate on proposed solutions. A mutual agreement on a best practice to describe interactions between human and technology does not exist, which be a challenge for the practitioners that are designing novel systems with increased autonomy.

Bearing in mind the first challenge in this paper which argues for a holistic approach rather than a reductionist approach, it will not be beneficial to aim for complete functional allocation. There is a need to search for solutions that encompass complexity and a need for development of more dynamic models of HAI. A possible first step that does not contradict neither the defenders nor the opponents of the concept of LOA and function allocation could be to identify the system's top-functions, and then move on to exploring the human role in the top-function in terms of responsibility, authority and control.

Navigation could be a top-level function, and the discussion could start with exploring the responsibility, authority and control within this function. Amy R. Pritchett describes the relationship between responsibility and authority as *"authority is generally used to describe who is assigned the execution of a function in operational sense, responsibility identifies who will be held accountable in an organizational and legal sense for the outcome"* [42]. Execution in this definition is presumed to include all four stages from information acquisition to action implementation, and is not solely linked to the action implementation.

Control does not, as many of the other terms in this paper, fall under a uniform definition. Like the term autonomy, the Society of Automotive Engineers has placed the term control in their section of deprecated terms in their recommended practice. The reason is that the term has numerous meanings in technical, legal and popular language [10]. Taking a systems perspective, Leveson [43] states *that "control processes operate between levels to control the processes at lower level in the hierarchy. These control processes enforce the safety constraints for which the control process is responsible".*

Control is linked to both responsibility and authority, and control is the process where the *responsible* agent of the function ensures that the agent with given *authority* executes its function in accordance with the system's requirement (see Fig. 1).

SARUMS *"methods of control"* describes the relationship between human and vessel ranging from method 1 which is "Operated" (remote control, tele-operation or manual operation), to method 5 which is "Autonomous" [44, 45]. The *"methods of control*" are a valuable contribution to create a more accurate characterisation of control, however the approach is not the best fitted for discussing how humans will be involved in future systems with increased autonomy. The responsible agent of a top-function needs to ensure that the system's requirements, decided on the strategic level, are translated to safety constraints that then are complied with on the operational level. Both responsibility and control will, at least for the near future, be allocated to humans, and hence *methods of control* should be defined from the perspective of the *humans* rather than the *vessel*.

This perspective should be explored in depth, since even though most of the published documents of maritime autonomy address some human interaction, it is predominantly discussed from the perspective of the vessel. Scoping a system based on responsibility, authority and control from the *human perspective* will bring the human into a central role, and pave the way for a human-centred design approach.

A human perspective on *authority* will take into account the challenges of humans directly involved as "executor" but from a position on shore (e.g for remote operated vessels, or intervening if a self-navigating vessel is out of its constraints). The authority sharing will include many of the traditional challenges in HAI, which includes the discussion of what and how to share functions between human and technology.

A human perspective on *methods of control* will be able to describe different types of control to ensure that the function is executed in accordance with the system's requirement and human capabilities. The different methods of control will experience different challenges that the system needs to take into account. Examples of methods of control could be direct involvement (combined role with authority), monitoring (continuously assessing the executor's decisions) or supervising (intermittently assessing the executor's decisions). However, these methods of controls are simplified, and the best fitted methods of control for maritime autonomy should be further explored.

A human perspective on *responsibility* will discuss which responsibilities are linked to the top-function, and if there are areas of responsibility that are not accounted for in a new concept. It will contribute to the discussion of competencies and legal accountability, and it will be important for designing a concept that could be approved by authorities.

Further, it is important that we encounter for internal and external variations in the system. In practice, this means that we cannot develop a static concept with *one* agent given authority to execute and choose *one* method of control. In navigation of a vessel we will see different requirements in congested waters than in open waters. The terminology we use needs to be able to describe a dynamic concept, and handle the complexity that follows with changing authority and methods of control during an operation.

How the humans are involved will change, and we will see that technical agents will be given more authority to execute functions. However, humans will remain

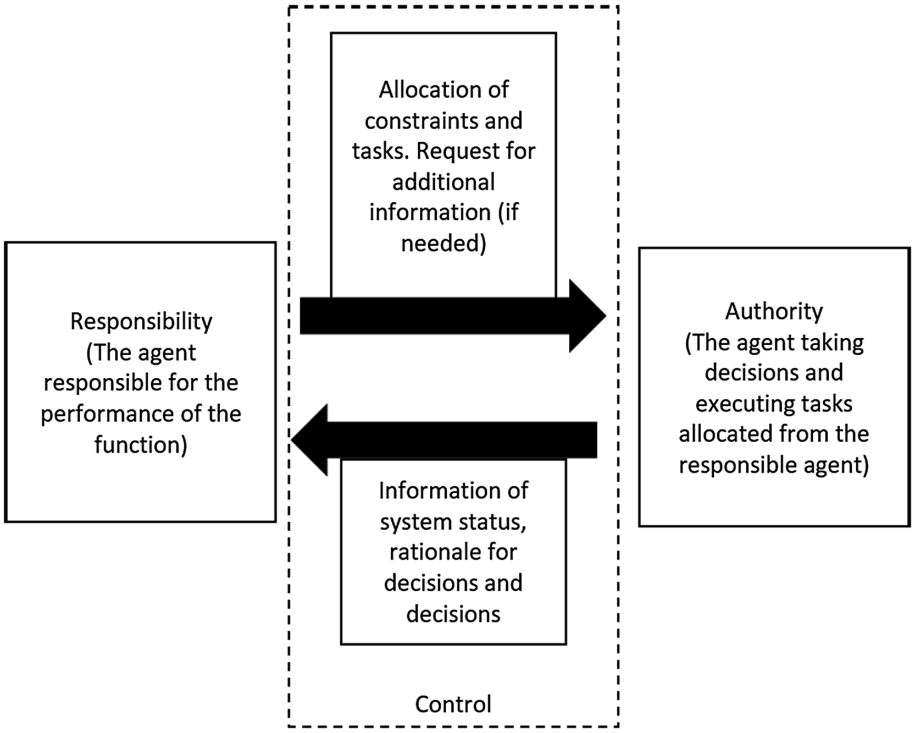

Fig. 1. Control links the responsibility and authority

responsible and humans will remain in control. It is therefore imperative that we develop a terminology that is best fitted to describe responsibility, control and authority from the human perspective. The way humans will be involved in future autonomous operation leads to new challenges, and these challenges need to be overcome to prove the safety status of novel system.

4 Conclusion

The increasing interest in autonomy in transport segments is also present in the maritime industry. Even though it is gaining a lot of attention, there is no unified definition of what autonomy is. This paper argues that agreeing on a defined state of being for autonomy would not be possible, and focuses on the autonomy as *a process of change.* As for automation, autonomy is about how to increase the use of machine agents in functionalities previously done by humans. The use of levels of autonomy as a state of being would be imprecise since what is defined as a high level of autonomy today (as self-navigating vessels monitored from shore) will be a lower level of autonomy in few years (if the machine agents are replacing humans on shore). This perspective acknowledges that autonomy is different from system to system, and will vary over time and be affected by the context. The purpose of changing the focus from a state of

being to a change process is to learn from many aspects of autonomy and allow for different factors based on context and previous state of the system.

The paper discusses the importance of considering humans in the development of autonomy in three areas concerning the safety of the navigation function. The first area is to leave the traditional safety approach, where systems are reduced to components and these components are assessed in isolation. The paper argues for a systemic approach to safety with a holistic perspective, where safety is an emergent property of the system, and human performance variability is essential for improving safety.

The second area is to understand why there will be humans in the new loops of systems with increased autonomy. The paper uses the levels strategic, tactical and operational to argue that autonomy would initially be experienced on the operational level, while the human ability to perform trade-offs between strategic and tactical objectives is still superior to the technology. System designers need to understand the importance of humans in the loop of future systems.

The third area is to know how to involve humans in the system, and for system designers it will be essential to follow the on-going discussion of the validity of the concept of function allocation and levels of automation to describe HAI. Both improving the concept and leaving the concept will lead to major implications within HAI. Independently of this discussion the paper argues for taking a human perspective on responsibility, authority and control of the top-function, such as navigation. As humans will be involved in the loop, at least on tactical and strategic levels, they will also be responsible and be involved in control processes of the execution of function, and the paper highlights the importance of developing *methods of control* from the human perspectives in the development of autonomy in the maritime industry.

References

1. IMO: MSC 98th session (2017). http://www.imo.org/en/MediaCentre/MeetingSummaries/MSC/Pages/MSC-98th-session.aspx. Accessed 5 Oct 2017
2. English Oxford Living Dictionaries: Definition of autonomy. https://en.oxforddictionaries.com/definition/autonomy. Accessed 19 Dec 2017
3. Blanchard, B.S., Fabrycky, W.J.: System Engineering and Analysis, 5th edn. Pearson Education Limited, London (2013)
4. Kaber, D.B.: A conceptual framework of autonomous and automated agents. Theor. Issues Ergon. Sci., 1–25 (2017). https://doi.org/10.1080/1463922X.2017.1363314
5. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. Hum. Factors: J. Hum. Factors Ergon. Soc. **39**, 230–253 (1997). https://doi.org/10.1518/001872097778543886
6. Lloyds Register: LR Code for Unmanned Marine Systems (2017)
7. Endsley, M.R.: Level of automation forms a key aspect of autonomy design. J. Cogn. Eng. Decis. Mak. (2017). https://doi.org/10.1177/1555343417723432
8. Bradshaw, J.M., Hoffman, R.R., Johnson, M., Woods, D.D.: The seven deadly myths of "autonomous systems". IEEE Intell. Syst. **28**, 54–61 (2013). https://doi.org/10.1109/mis.2013.70

9. Kaber, D.B.: Issues in human-automation interaction modeling: presumptive aspects of frameworks of types and levels of automation. J. Cogn. Eng. Decis. Mak. (2017) https://doi.org/10.1177/1555343417737203
10. SAE International: Surface Vehicle Recommended Practice J3016 (2016)
11. Parasuraman, R.: Designing automation for human use: empirical studies and quantitative models. Ergonomics **43**, 931–951 (2000). https://doi.org/10.1080/001401300409125
12. Rolls-Royce: Autonomous ships: the next step. In: AAWA: Advanced Autonomous Waterborne Applications, vol. 7 (2016)
13. Insaurralde, C.C., Petillot, Y.R.: Capability-oriented robot architecture for maritime autonomy. Robot. Auton. Syst. **67**, 87–104 (2015). https://doi.org/10.1016/j.robot.2014.10.003
14. Wahlström, M., Hakulinen, J., Karvonen, H., Lindborg, I.: Human factors challenges in unmanned ship operations – insights from other domains. Procedia Manuf. **3**, 1038–1045 (2015). https://doi.org/10.1016/j.promfg.2015.07.167
15. Dekker, S.W.A., Woods, D.D.: MABA-MABA or abracadabra? Progress on human-automation co-ordination. Cogn. Technol. Work **4**, 240–244 (2002)
16. Van Den Broek, H., Schraagen, J.M., Brake, G., Van, J.: Approaching full autonomy in the maritime domain: paradigm choices and human factors challenges, pp. 1–11 (2017)
17. Lee, J.D.: Perspectives on automotive automation and autonomy. J. Cogn. Eng. Decis. Mak. (2017) https://doi.org/10.1177/1555343417726476
18. Munin: MUNIN's Rationale | MUNIN (2016). http://www.unmanned-ship.org/munin/about/munins-rational/. Accessed 6 Oct 2017
19. Allianz: Safety and Shipping Review 2015, p. 36 (2015)
20. Bainbridge, L.: Ironies of automation. Automatica **19**, 775–779 (1983). https://doi.org/10.1016/0005-1098(83)90046-8
21. Strauch, B.: Ironies of automation: still unresolved after all these years. IEEE Trans. Hum.-Mach. Syst., 1–15 (2017). https://doi.org/10.1109/thms.2017.2732506
22. Boy, G.A.: Orchestrating Human-Centered Design. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4339-0
23. Anderson, P.W.: More is different. Science **177**(4047), 393–396 (1972)
24. Anderson, P.W.: More and Different: Notes from a Thoughtful Curmudgeon. World Scientific Publishing Company, Singapore (2011)
25. Rasmussen, J.: Risk management in a dynamic society - a modelling problem. Saf. Sci. **27**, 183–213 (1997)
26. Leveson, N.: A new accident model for engineering safer systems. Saf. Sci. **42**, 237–270 (2004). https://doi.org/10.1016/S0925-7535(03)00047-X
27. Hollnagel, E.: A tale of two safeties. Nucl. Saf. Simul. **4**, 1–9 (2012)
28. Hollnagel, E.: Coping with complexity: past, present and future. Cogn. Technol. Work **14**, 199–205 (2012). https://doi.org/10.1007/s10111-011-0202-7
29. Stanton, N.A., Salmon, P.M., Rafferty, L.A., Walker, G.H., Baber, C., Jenkins, D.P.: Human Factors Methods - A Practical Guide for Engineering and Design, 2nd edn. Ashgate Publishing Limited, Farnham (2013)
30. Stanton, N.A., Harvey, C.: Beyond human error taxonomies in assessment of risk in sociotechnical systems: a new paradigm with the EAST "broken-links" approach. Ergonomics **60**, 221–233 (2017). https://doi.org/10.1080/00140139.2016.1232841
31. Sheridan, T.B.: Comments on "issues in human–automation interaction modeling: presumptive aspects of frameworks of types and levels of automation" by David B. Kaber. J. Cogn. Eng. Decis. Mak. (2017) https://doi.org/10.1177/1555343417724964
32. Hollnagel, E.: FRAM: The Functional Resonance Analysis Method: Modelling Complex Socio-Technical Systems. Ashgate Publishing Ltd., Farnham (2012)
33. Ackoff, R.L.: Strategy. Syst. Pract. **3**, 521–524 (1990). https://doi.org/10.1007/bf01059636

34. Michon, J.A.: A critical view of driver behavior models. In: Evans, L., Schwing, R.C. (eds.) Human Behavior and Traffic Safety, pp. 485–520. Springer, Boston (1985). https://doi.org/10.1007/978-1-4613-2173-6_19
35. Forsvaret: Forsvarets fellesoperative doktrine - Norwegian Armed Forces Joint Operational Doctrine (2014)
36. NATO: Nato Standard AJP-01 Allied Joint Doctrine (2017)
37. Hollnagel, E.: Time and time again. Theor. Issues Ergon. Sci. **3**, 143–158 (2002). https://doi.org/10.1080/14639220210124111
38. Hollnagel, E.: Safer Complex Industrial Environments. CRC Press, Boca Raton (2009)
39. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. - Part A: Syst. Hum. **30**, 286–297 (2000). https://doi.org/10.1109/3468.844354
40. Jamieson, G.A., Skraaning, G.: Levels of automation in human factors models for automation design: why we might consider throwing the baby out with the bathwater. J. Cogn. Eng. Decis. Mak. (2017) https://doi.org/10.1177/1555343417732856
41. Wickens, C.: Automation stages & levels, 20 years after. J. Cogn. Eng. Decis. Mak. (2017) https://doi.org/10.1177/1555343417727438
42. Pritchett, A.R., Kim, S.Y., Feigh, K.M.: Modeling human-automation function allocation. J. Cogn. Eng. Decis. Mak. **8**, 33–51 (2014). https://doi.org/10.1177/1555343413490944
43. Leveson, N.G.: Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, Cambridge (2011)
44. Ornfelt, M., (SARUMS): Safety and regulations for unmanned maritime systems. In: Unmanned Surface Vessel Regulation Conference (2016). http://www.ukmarinealliance.co.uk/sites/default/files/SARUMS-2016-MASRWG.pdf
45. Lighthouse- Swedish Maritime Competence: Autonomous safety on vessels (2016)

Improving Understanding of Mindfulness Concepts and Test Methods

Melissa M. Walwanis[1(✉)] and Derek S. Bryan[2]

[1] Naval Air Warfare Center Training Systems Division, Orlando, FL, USA
Melissa.Walwanis@navy.mil
[2] Ingenia Services, Wake Forest, NC, USA
dbryan@ingeniaservices.com

Abstract. The concept of mindfulness is largely dependent on one's theoretical perspective but, in general, there is agreement that it involves open receptive attention, present moment awareness, and de-automization in thought processes. As a contemplative training intervention, mindfulness has been especially lauded by many practitioners as making improvements to performance ranging from increased productivity to enhanced decision making [11]. While some of these results are backed by empirical evidence, the scientific community lags in comprehensively validating these claims [19]. This has resulted in calls from the science community to establish a comprehensive research agenda across disciplines of Psychology to address the need to underpin practical prescriptions with empirically derived principles and guidelines [5, 6, 12, 19]. This paper reviews some of the criticisms of the existing body of literature and provides recommendations for moving towards a rigorously informed evidence-based practice. Next, we integrate frameworks for mindfulness concepts across disciplines and offer consideration of how Modeling and Simulation, in combination with proven statistical methods, can be utilized to understand the relationships and significance of mindfulness factors. Finally, we discuss the plausibility of further mindfulness research and test methods with the potential to improve human performance across a wide variety of activities.

Keywords: Mindfulness · Modeling and Simulation · Design of Experiments

1 Introduction

The concept of mindfulness is largely dependent on one's theoretical perspective but, in general, there is agreement that it involves open receptive attention, present moment awareness, and de-automization in thought processes. As a contemplative training intervention, mindfulness has been especially lauded by many practitioners as making improvements to performance ranging from increased productivity to enhanced decision making [11]. While some of these results are backed by empirical evidence, the scientific community lags in comprehensively validating these claims [19]. Despite these techniques ancient origins in religious practices (e.g., Zen Buddhism) and use in Clinical Psychology settings, this is an emergent field of scientific inquiry in a nascent state [4, 10, 20]. This has resulted in calls from the science community to establish a

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10916, pp. 363–374, 2018.
https://doi.org/10.1007/978-3-319-91467-1_28

comprehensive research agenda across disciplines of Psychology to address the need to underpin practical prescriptions with empirically derived principles and guidelines [5, 6, 12, 19].

Some criticisms of the existing body of empirical research are that there is no single operational definition for mindfulness, an over reliance on subjective recall measures leaving common method bias as a concern, an ill-defined nomological network, a failure to control for confounds, and an inability to replicate results found [1, 4, 19]. Additionally, Clinical Psychology scholars are pointing out potential negative outcomes of mindfulness-based interventions with contraindications for certain populations emerging in the empirical literature [19]. Certainly, this points to the need to more fully understand the boundary conditions of contemplative strategies. Acknowledgement of these challenges provides opportunities to employ rigorous methods driven by theory to arrive at an informed evidence-based practice. Further, these types of studies will assist practitioners with answering what the return-on-investment is for interventions such as mindfulness practice.

Historically, Modeling and Simulation (M&S) test environments have supported the development of principles and guidelines based on multitrait-multimethod approaches in other contexts and may provide a similar supporting role for mindfulness concepts [3]. When combined with a statistical methodology such as Design of Experiments (DoE), M&S offers an effective and efficient strategy for determining and evaluating key system and human performance parameters. M&S and DoE have been successfully applied to a wide variety of industries including medical, agricultural, e-commerce, and defense [13]. One of the fundamental concepts of DoE – replication – is well-suited for M&S applications. Replication via M&S increases test data and confidence in test results, allowing for comparisons across samples and techniques that would be difficult, impractical, or too expensive otherwise [17].

Given the lack of empirical evidence for the effectiveness of contemplative training interventions, could a similar analytical approach (M&S plus DOE) be used to validate current or find alternative results? The ability to identify and scope significant factors and control the test environment remains paramount. Certainly, factors such as the background of the individual, the quantity and type of contemplative (e.g., mindful breathing, focused thought, meditation) or other training interventions experienced, task expertise level, and nature of the task require further examination and control, which M&S offers [4]. What other factors have the potential to affect the results and to what degree do those factors interact? A recent meta-analysis of trait mindfulness, the average/baseline level of a person's mindfulness absent a mindfulness practice or intervention, suggests the existence of mediating variables between this construct and work effort and perceived job stress respectively. An M&S environment would surely offer opportunities to identify such relationships and interaction effects in a controlled setting [16]. What test environment features (M&S or otherwise) are necessary to realistically stimulate the system under test and conduct data collection? We posit that an M&S testbed, along with a statistically-designed test approach, could provide empirical results to these questions for a given task or activity and system under test. Prior to a discussion of M&S, we focus on introducing mindfulness definitions and

conceptualizations from multiple academic disciplines. Next, we discuss methodological shortfalls in mindfulness research and propose measures for assessment across human functions and performance categories. We then demonstrate basic M&S and DoE principles to a sample task (driving) to show their ability to better understand the relationships and significance of the mindfulness categories and associated measures. Finally, we draw preliminary conclusions from the presented research and propose foundations for future mindfulness research and testing.

2 Mindfulness Definition and Concept Confusion

Globally, what is meant by the term mindfulness is largely dependent upon theoretical perspective, leading to an incohesive literature base, which is further compounded by interest in the topic across disciplines. While cross-disciplinary interest provides some exciting prospects, it also presents challenges when crosstalk is stilted. The theoretical perspectives underpinning work in mindfulness can be crudely dichotomized into those that nest neatly within Eastern Philosophy and those that have been adapted to fit within Western Philosophy. Regardless of philosophical stance, there is agreement that mindfulness involves open receptive attention, present moment awareness, and de-automization in thought processes. Beyond this, there are significant departures that serve as sources of debate. Table 1 below provides some popular definitions for mindfulness across academic disciplines. It is evident from this list that it is largely considered as a state as opposed to trait construct. However, during our review of the literature for preparation of this manuscript we noted on several occasions that surveys psychometrically validated to assess trait mindfulness were used to assess state mindfulness. Unfortunately, this is not an uncommon occurrence in the study of mindfulness [9].

Table 1. Mindfulness operational definitions in the literature

Mindfulness definition	Academic discipline	Citation
"Self-regulation of attention so that it is maintained on immediate experience, thereby allowing for increased recognition of mental events in the present moment" and "adopting a particular orientation toward one's experiences in the present moment, an orientation that is characterized by curiosity, openness, and acceptance"	Clinical psychology	p. 232, [1]
"A state of consciousness in which attention is focused on present-moment phenomena occurring both externally and internally"	Business	p. 997, [4]
"State of consciousness characterized by receptive attention to and awareness of present events and experiences, without evaluation, judgment, and cognitive filters"	Industrial/organizational psychology	p. 119, [7]

(*continued*)

Table 1. (*continued*)

Mindfulness definition	Academic discipline	Citation
"The awareness that emerges through paying attention on purpose, in the present moment, and non-judgmentally to the unfolding experience moment by moment"	Medicine	p. 146, [14]
State based processing whereby new categories are created, or the re-creation of existing categories, one is open to receiving new information, and one is aware of more than one perspective	Social psychology	[15]
"A state involving the simultaneous arising of a particular intention, attention, and attitude"	Clinical psychology	p. 383, [18]

One of the greatest challenges in the empirical literature is the lack of a consistent conceptualization for mindfulness. Good et al. [9] conducted a review of the mindfulness literature to understand the effects in the workplace. Results of this review revealed that the term "mindfulness" has been used to refer to trait mindfulness, state mindfulness, mindfulness practice, and mindfulness interventions. While all of these uses are valid, the use of the umbrella term "mindfulness" is not recommended for facilitating a coherent scientific and technical base to advance understanding. Rather, specificity of which conceptualizations are under consideration in any given study is imperative. In alignment with this recommendation, we offer Table 2 below to facilitate selection of terminology. Regardless of concept(s) undergoing test, research methodology remains a concern across disciplines.

Table 2. Mindfulness conceptualizations.

Mindfulness concept	Definition	Citation
Trait mindfulness	Individual predisposition to engage in receptive attention to and awareness of present events and experiences or the average/baseline level of a person's mindfulness absent a mindfulness practice or intervention	[2]
State mindfulness	State of experiential processing focused on attention to internal and/or external stimulus to register the facts observed in the present moment	[9]
Mindfulness practice	Actively practicing mindfulness activities such as focused attention or monitoring of sensory stimuli	[9]
Mindfulness intervention	An organizational intervention such as a lecture, discussion, or policy/procedure designed with a specific organizational outcome (e.g., wellness, enhanced decision making)	[9]

3 Addressing Methodological Shortfalls in Testing Mindfulness Concepts

Recently, Goldberg et al. conducted a systematic review of the methodological quality of the Clinical Psychology mindfulness literature base, which revealed modest improvements over the last 17 years [8]. However, they did identify needed methodological improvements: (1) active control conditions, (2) larger sample sizes, (3) longitudinal studies, (4) treatment fidelity assessment, and (5) reporting of instructors/instruction certification/validation. Indeed, these shortfalls can be leveled on the Work Psychology literature as well adding an overreliance on cross-sectional methods leaving common method bias a concern and causation in the existing nomological network unanswered [9]. Further, a failure to replicate results has been noted. All told, this presents opportunities to easily remedy the many methodological deficiencies noted (e.g., conducting a Power Analysis can assist with identifying the right sample size to adequately test a concept in any given study).

Recently, there have been efforts across both the Clinical and Work Psychology disciplines to provide frameworks to organize existing research and define points of departure for future research [9, 19]. We integrated these frameworks in Table 3 below where there was convergence and added a category where one should naturally exist (i.e., attitudes). Additionally, we culled existing measures that have been used to test mindfulness concepts in the literature demonstrating that researchers are spanning beyond the surveys used in cross-sectional studies. This list is in no way exhaustive but rather is intended to serve as a point of departure to inspire future directions. Working with these measures, studies to test antecedents, correlates, and proximal/distal outcomes can easily be conceived. In this vein, such an organizing framework lends itself to development of testable theories of mindfulness, where few exist. Further, rigorous methodologies, and understanding of variables that may have substantial pay off one could engage in Experimental Design to rapidly define a research agenda.

M&S offers a strong resource for rigorous empirical test of mindfulness concepts. First, M&S offers the suspension of reality through the creation of contexts that research participants experience. For example, in a clinical setting one could easily conceive of modeling a series of anxiety inducing environments in which the efficacy of various mindfulness practice could be tested. Similarly, in a work setting, environments that simulate task settings could be developed to test the effectiveness of mindfulness on worker's performance. These types of environments could be used to support laboratory, quasi-experimental, longitudinal, and computational experimental methods while also providing a measure of control that has been missing in many past efforts. Further, it is plausible that these measures can be combined in a myriad of ways dependent upon the research questions of interest.

Table 3. Mindfulness categories, potential measures, & potential utilization of M&S testbed

Categories of mindfulness lines of scientific inquiry	Potential measures	M&S testbed utility	Citations
Human functioning			
Cognition	Cognitive capacity Cognitive flexibility	Present situations/tasks that allow assessment of capacity/flexibility	[9, 19]
Emotional	Reactivity Valence	Present situations that elicit a variety of emotional responses	[9, 19]
Behavioral	Self regulation Reduced automaticity	Provide situations/tasks with branching ipsative choices to allow for assessment of effects of mindfulness practice on self-regulation/automaticity	[9, 19]
Physiological	Neural plasticity Cortisol levels Brain response Heart rate Respiration	Enable collection of psychophysiological data in a controlled setting with high fidelity situations/tasks	[9, 19]
Human performance			
Social/interpersonal relationships	360° feedback reports Communications Quality of interactions Conflict management Empathy/compassion Leadership Team performance	Present a series of situations that allow for situational judgements & assessment of responses Provide virtual role players as a reliable consistent stimulus	[9, 19]
Performance	Productivity Job/task Safety	Present situations/tasks that enable assessment of job performance	[9, 19]
Well being	Psychological	Present varied situations to assess state of well being	[9, 19]
Attitudinal	Job satisfaction Organizational citizenship behaviors Deviance	Present scenarios/vignettes that allow for assessment of work related attitudes	N/A
Attention	Stability Control Efficiency	Present tasks, such as vigilance tasks, that enable the assessment of attention	[9, 19]

4 Leveraging M&S and Experimental Design to Test a Sample Mindfulness Research Agenda

M&S environments are typically very good at computationally-based problems and can often be executed many times and very quickly. In doing so, M&S can produce large sets of results, generally for much less time, effort, and resources than would otherwise be required. It is these basic characteristics that often lead people to use M&S to address their research questions. But how do you know which research questions can best be addressed by M&S? How should you interpret your results? And how should you use your results to refine your model and to improve system performance? It is these and related questions that led to the development of a statistical methodology for planning, conducting, and analyzing experiments, including those that use computer-based M&S, known as Design of Experiments (DoE).

Believed to have begun in the 1920s in the agricultural industry, DoE uses statistical methods to efficiently identify key factors and obtain the most information with as few trials as possible. Maximizing these efficiencies becomes very important when dealing with limited, expensive, or high-risk resources. The process to identify what factors or combinations of factors impacts the desired response variable(s) is called *screening*. In its simplest form, screening can be implemented by a *factorial design* that includes all combinations of factors at all levels – two factors, each with two levels would produce $2^2 = 4$ trials. If one factor has a different effect (response variable outcome) at different levels on another factor, this is called an *interaction* [13]. The existence of an interaction, along with an understanding of the desired response variable(s), can be used to make more efficient experimental designs (fewer trials). One can imagine that complex experiments with many factors and non-continuous levels would produce an unmanageable number of trials. To deal with this situation, experimenters can intelligently reduce the number of factors and levels based on their higher order interactions through a *fractional factorial design* [17].

Additional principles for experimental designs can be used to ensure the objectivity and efficiency of trials. *Randomization* implies running trials in random order to reduce bias to the degree possible. This principle is especially useful when human participants are involved. *Replication* is the repeat of one or more trials in order to estimate the experimental error (typically minor differences in response variables due to unimportant factors e.g., accuracy or consistency of a scale) and *blocking* attempts to suppress the impact of high-variance factors on the experimental error [13, 17].

In the late 1990's the National Highway Traffic Safety Administration along with the National Center on Sleep Disorders Research conducted a comprehensive study on driver drowsiness and fatigue. While not directly related to mindfulness, the study provided a framework for understanding various effects on driving that could be extended and applied to mindfulness. Further, the study provides a context (driving) that is already well-represented within the civilian, commercial, and military M&S community. The following discussion is offered as an example of how M&S and DoE could be applied to a mindfulness context with drivers.

Table 4. Full factorial run matrix

Trial	Driver age	Periodicity	Traffic level	Distraction level	Trial	Driver age	Periodicity	Traffic level	Distraction level
1	L	D	L	L	42	M	W	M	H
2	L	D	L	M	43	M	W	H	L
3	L	D	L	H	44	M	W	H	M
4	L	D	M	L	45	M	W	H	H
5	L	D	M	M	46	M	M	L	L
6	L	D	M	H	47	M	M	L	M
7	L	D	H	L	48	M	M	L	H
8	L	D	H	M	49	M	M	M	L
9	L	D	H	H	50	M	M	M	M
10	L	W	L	L	51	M	M	M	H
11	L	W	L	M	52	M	M	H	L
12	L	W	L	H	53	M	M	H	M
13	L	W	M	L	54	M	M	H	H
14	L	W	M	M	55	H	D	L	L
15	L	W	M	H	56	H	D	L	M
16	L	W	H	L	57	H	D	L	H
17	L	W	H	M	58	H	D	M	L
18	L	W	H	H	59	H	D	M	M
19	L	M	L	L	60	H	D	M	H
20	L	M	L	M	61	H	D	H	L
21	L	M	L	H	62	H	D	H	M
22	L	M	M	L	63	H	D	H	H
23	L	M	M	M	64	H	W	L	L
24	L	M	M	H	65	H	W	L	M
25	L	M	H	L	66	H	W	L	H
26	L	M	H	M	67	H	W	M	L
27	L	M	H	H	68	H	W	M	M
28	M	D	L	L	69	H	W	M	H
29	M	D	L	M	70	H	W	H	L
30	M	D	L	H	71	H	W	H	M
31	M	D	M	L	72	H	W	H	H
32	M	D	M	M	73	H	M	L	L
33	M	D	M	H	74	H	M	L	M
34	M	D	H	L	75	H	M	L	H
35	M	D	H	M	76	H	M	M	L
36	M	D	H	H	77	H	M	M	M
37	M	W	L	L	78	H	M	M	H
38	M	W	L	M	79	H	M	H	L
39	M	W	L	H	80	H	M	H	M
40	M	W	M	L	81	H	M	H	H
41	M	W	M	M					

The purpose of our sample research project is to determine the mindfulness-related effects on driving. Our dependent variables (measured response outcomes) will be both physiological (heart rate, respiration), and attention (stability, control, efficiency). Our independent variables (factors) will be driver age (16–25 [L], 26–55 [M], 56+ [H]), periodicity (how often the driver completes this route – daily [D], weekly [W], monthly [M]), traffic level (low [L], medium [M], high [H]), and distraction level (occurrence of unanticipated events - low [L], medium [M], high [H]). Therefore, we have up to $3^4 = 81$ trials if we were to conduct a full factorial design of this experiment as further detailed in Table 4.

Conducting a live experiment with 81 trials including human drivers, different traffic levels, and different distraction levels would be difficult to control and potentially very time consuming. For these reasons, we have decided to use a virtual simulator (human operator using simulated equipment) to conduct our experiment. We estimate that each trial will take approximately 45 min (~60 h total) to complete. Unfortunately, we only have access to the driving simulator for a maximum of 30 h (~40 trials) so we will have to find ways to reduce the number of required trials by half while still maintaining confidence in our results.

Table 5. Sampling run matrix

Trial	Driver age	Periodicity	Traffic level	Distraction level
1	L	D	L	L
2	L	D	L	H
3	L	D	H	L
4	L	D	H	H
5	L	M	L	L
6	L	M	L	H
7	L	M	H	L
8	L	M	H	H
9	M	D	L	L
10	M	D	L	H
11	M	D	H	L
12	M	D	H	H
13	M	M	L	L
14	M	M	L	H
15	M	M	H	L
16	M	M	H	H
17	H	D	L	L
18	H	D	L	H
19	H	D	H	L
20	H	D	H	H
21	H	M	L	L
22	H	M	L	H
23	H	M	H	L
24	H	M	H	H

The first step in trying to reduce the number of trials is to identify which factors do and do not interact (have an impact on the response outcomes). Unless there is an existing data set for the factors of interest, one will need to find a method to determine if and to what degree there are interactions. One of the most straightforward ways of determining interactions is by sampling and executing a subset of the trials. Using the full factorial run matrix above, we have decided to sample Driver Age [L, M, H], Periodicity [D, M], Traffic Level [L, H], and Distraction Level [L, H]. This is a reasonable approach because we are sampling all levels of Driver Age and the boundaries of all other factors. These samples will use 24 of the 40 available trials as detailed in the Table 5 but should give us strong indications of interactions.

Through execution of the selected trials we have learned that Driver Age and Traffic Level has minimal interactions (the effect is not significant on the desired response outcomes). Based on this information and the number of available trials remaining (16), we have refined our run matrix in Table 6 as follows:

Table 6. Refined run matrix

Trial	Driver age	Periodicity	Traffic level	Distraction level
25	M	D	M	L
26	M	W	M	M
27	M	M	M	H
28	M	D	L	L
29	M	D	M	L
30	M	W	M	M
31	M	M	M	H
32	M	D	M	M
33	M	D	M	L
34	M	W	M	M
35	M	M	M	H
36	M	D	H	H
37	L	D	L	L
38	L	M	H	H
39	H	D	H	H
40	H	M	L	L

The refined run matrix largely keeps driver age and traffic constant while iterating all levels of periodicity and distraction level. The final four trials (36–40) are "extra" and are replications of trials from our sampling matrix to ensure continuity of results.

The notional results of our sample research project indicate that periodicity has the largest physiological effect while distraction level has the largest attention effect and these factors do interact. Specifically, we now know that higher periodicity combined with lower distraction levels decreases our physiological (heart rate, respiration) and attention (stability, control, efficiency) factors, while lower periodicity and higher distraction levels increases our physiological and attention factors. Furthermore, due to

the appropriate application of DoE methods, we are able to state that our results have statistical significance and can be used to refine our current model or used as input to future studies. Finally, the use of M&S allowed us to conduct many more trials, with much more control of the experimental environment than would be possible in a live experiment.

5 Conclusion

While there is general agreement that improving "mindfulness" has positive effects, the research summarized in this paper confirms that there is no consensus regarding the various mindfulness theories, definitions, or measures. This suggests that the discipline and context to which one is applying mindfulness concepts must be strongly considered. This also suggests that discipline-specific approaches to mindfulness concepts may need to be further researched and developed.

Mindfulness categories and measures are proposed across human functions and performance that are essential for testing and assessing any mindfulness theory or definition. The sample research project uses those measures, along with a statistically-significant and replicable methodology (DoE plus M&S), to determine and evaluate mindfulness factors for a given context (driving). This example and approach is significant because it can be improved and applied to other contexts and can assist with the research and evolution of additional theories and definitions of mindfulness.

Collectively, this research assists the community in achieving a broader understanding and body of knowledge of the state of mindfulness concepts. Additional research is recommended to further understand and refine discipline-specific mindfulness concepts and test those concepts using proven experimental methods. If conducted, these activities are expected to result in improved mindfulness theories, definitions, and measures that can be used for the benefit of individual and collective human performance across a spectrum of disciplines.

Acknowledgement. The views expressed herein are those of the authors and do not necessarily reflect the official position of the organizations with which they are affiliated.

References

1. Bishop, S.R., Lau, M., Shapiro, S., Carlson, L., Anderson, N.D., Carmody, J., Segal, Z.V., Abbey, S., Speca, M., Velting, D., Devins, G.: Mindfulness: a proposed operational definition. Clin. Psychol. Sci. Pract. **11**(3), 230–241 (2004). https://doi.org/10.1093/clipsy/bph077
2. Brown, K.W., Ryan, R.M., Creswell, J.D.: Mindfulness: theoretical foundations and evidence for its salutary effects. Psychol. Inq. **18**(4), 211–237 (2007). https://doi.org/10.1080/10478400701598298
3. Campbell, D.T., Fiske, D.W.: Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull. **56**, 81–105 (1959)
4. Dane, E.: Paying attention to mindfulness and its effects on task performance in the workplace. J. Manag. **37**(4), 997–1018 (2011). https://doi.org/10.1177/0149206310367948

5. Davidson, R.J., Dahl, C.J.: Outstanding challenges in scientific research on mindfulness and meditation. Perspect. Psychol. Sci. 1–4 (2017). https://doi.org/10.1177/1745691617718358
6. Davidson, R.J., Kaszniak, A.W.: Conceptual and methodological issues in research on mindfulness and meditation. Am. Psychol. **70**(7), 581–592 (2015). https://doi.org/10.1037/a0039512
7. Glomb, T.M., Duffy, M.K., Bono, J.E., Yang, T.: Mindfulness at work. In: Research in Personnel and Human Resources Management, vol. 30, pp. 115–157 (2011). http://doi.org/10.1108/S0742-7301(2011)0000030005
8. Goldberg, S.B., Tucker, R.P., Greene, P.A., Simpson, T.L., Kearney, D.J., Davidson, R.J.: Is mindfulness research methodology improving over time? A systematic review. PLoS ONE **12**(10), 1–16 (2017). https://doi.org/10.1371/journal.pone.0187298
9. Good, D.J., Lyddy, C.J., Glomb, T.M., Bono, J.E., Warren Brown, K., Duffy, M.K., Baer, R.A., Brewer, J.A., Lazar, S.W.: Contemplating mindfulness at work: an integrative review. J. Manag. **42**(1), 114–142 (2016). https://doi.org/10.1177/0149206315617003
10. Harrington, A., Dunne, J.D.: When mindfulness is therapy: ethical qualms, historical perspectives. Am. Psychol. **70**(7), 621–631 (2015). https://doi.org/10.1037/a0039512
11. Hess, P.: Psychologists express growing concern with mindfulness meditation It's not bare-knuckle, that's for sure. Inverse Science (2017). Accessed https://www.inverse.com/article/37291-mindfulness-meditation-criticism-psychologists
12. Hyland, P.K., Lee, R.A., Mills, M.J.: Mindfulness at work: a new approach to improving individual and organizational performance. Ind. Organ. Psychol. **8**(4), 576–602 (2015). https://doi.org/10.1017/iop.2015.41
13. Johnson, R.T., Hutto, G.T., Simpson, J.R., Montgomery, D.C.: Designed experiments for the defense community. Qual. Eng. **24**, 60–79 (2012). https://doi.org/10.1080/08982112.2012.627288
14. Kabat-Zinn, J.: Mindfulness-based interventions in context: past, present, and future. Clin. Psychol. Sci. Pract. **10**(2), 144–156 (2003). https://doi.org/10.1093/clipsy/bpg016
15. Langar, E.J.: Mindfulness. Da Capo Press, Boston (2014)
16. Mermer-Magnus, J., Manapragada, A., Viswesvaran, C., Allen, J.W.: Trait mindfulness at work: a meta-analysis of the personal and professional correlates of trait mindfulness. Hum. Perform. **30**(2–3), 79–98 (2017). https://doi.org/10.1080/08959285.2017.1307842
17. Sanchez, S.: Work smarter, not harder: guidelines for designing simulation experiments. In: Proceedings of the 2005 Winter Simulation Conference (2005). Accessed http://www.dtic.mil/dtic/tr/fulltext/u2/a491983.pdf
18. Shapiro, S., Carlson, L., Astin, J.A., Freedman, B.: Mechanisms of mindfulness. J. Clin. Psychol. **62**(3), 373–386 (2006). https://doi.org/10.1002/jclp.20237
19. Van Dam, N.T., Van Vugt, M.K., Vago, D.R., Schmalzl, L., Saron, C.D., Olendzki, A., Meissner, T., Lazar, S.W., Kerr, C.E., Gorchov, J., Fox, K.C.R., Field, B.A., Britton, W.B., Brefczynski-Lewis, J.A., Meyer, D.E.: Mind the hype: a critical evaluation and prescriptive agenda for research on mindfulness and meditation. Perspect. Psychol. Sci. 1–26 (2017a). https://doi.org/10.1177/1745691617709589
20. Van Dam, N.T., Van Vugt, M.K., Vago, D.R., Schmalzl, L., Saron, C.D., Olendzki, A., Meissner, T., Lazar, S.W., Gorchov, J., Fox, K.C.R., Field, B.A., Britton, W.B., Brefczynski-Lewis, J.A., Meyer, D.E.: Reiterated concerns and further challenges for mindfulness and meditation research: a reply to Davidson and Dahl. Perspect. Psychol. Sci. 1–4 (2017b). https://doi.org/10.1177/1745691617727529

Author Index