Don Harris
Wen-Chin Li (Eds.)

LNAI 12767

Engineering Psychology and Cognitive Ergonomics

18th International Conference, EPCE 2021
Held as Part of the 23rd HCI International Conference, HCII 2021
Virtual Event, July 24–29, 2021, Proceedings

HCI 2021
INTERNATIONAL

Springer

Lecture Notes in Artificial Intelligence 12767

Subseries of Lecture Notes in Computer Science

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

Don Harris · Wen-Chin Li (Eds.)

Engineering Psychology and Cognitive Ergonomics

18th International Conference, EPCE 2021
Held as Part of the 23rd HCI International Conference, HCII 2021
Virtual Event, July 24–29, 2021
Proceedings

Springer

Editors
Don Harris
Coventry University
Coventry, UK

Wen-Chin Li
Cranfield University
Cranfield, UK

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-77931-3 ISBN 978-3-030-77932-0 (eBook)
https://doi.org/10.1007/978-3-030-77932-0

LNCS Sublibrary: SL7 – Artificial Intelligence

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

Foreword

Human-Computer Interaction (HCI) is acquiring an ever-increasing scientific and industrial importance, and having more impact on people's everyday life, as an ever-growing number of human activities are progressively moving from the physical to the digital world. This process, which has been ongoing for some time now, has been dramatically accelerated by the COVID-19 pandemic. The HCI International (HCII) conference series, held yearly, aims to respond to the compelling need to advance the exchange of knowledge and research and development efforts on the human aspects of design and use of computing systems.

The 23rd International Conference on Human-Computer Interaction, HCI International 2021 (HCII 2021), was planned to be held at the Washington Hilton Hotel, Washington DC, USA, during July 24–29, 2021. Due to the COVID-19 pandemic and with everyone's health and safety in mind, HCII 2021 was organized and run as a virtual conference. It incorporated the 21 thematic areas and affiliated conferences listed on the following page.

A total of 5222 individuals from academia, research institutes, industry, and governmental agencies from 81 countries submitted contributions, and 1276 papers and 241 posters were included in the proceedings to appear just before the start of the conference. The contributions thoroughly cover the entire field of HCI, addressing major advances in knowledge and effective use of computers in a variety of application areas. These papers provide academics, researchers, engineers, scientists, practitioners, and students with state-of-the-art information on the most recent advances in HCI. The volumes constituting the set of proceedings to appear before the start of the conference are listed in the following pages.

The HCI International (HCII) conference also offers the option of 'Late Breaking Work' which applies both for papers and posters, and the corresponding volume(s) of the proceedings will appear after the conference. Full papers will be included in the 'HCII 2021 - Late Breaking Papers' volumes of the proceedings to be published in the Springer LNCS series, while 'Poster Extended Abstracts' will be included as short research papers in the 'HCII 2021 - Late Breaking Posters' volumes to be published in the Springer CCIS series.

The present volume contains papers submitted and presented in the context of the 18th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2021), an affiliated conference to HCII 2021. I would like to thank the Co-chairs, Don Harris and Wen-Chin Li, for their invaluable contribution to its organization and the preparation of the proceedings, as well as the members of the Program Board for their contributions and support. This year, the EPCE affiliated conference has focused on topics related to human performance and error, cognitive processes, cognitive psychology in aviation and air traffic control, and cognitive psychology in interaction design.

I would also like to thank the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences for their contribution towards the highest scientific quality and overall success of the HCI International 2021 conference.

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

July 2021 Constantine Stephanidis

HCI International 2021 Thematic Areas and Affiliated Conferences

Thematic Areas

- HCI: Human-Computer Interaction
- HIMI: Human Interface and the Management of Information

Affiliated Conferences

- EPCE: 18th International Conference on Engineering Psychology and Cognitive Ergonomics
- UAHCI: 15th International Conference on Universal Access in Human-Computer Interaction
- VAMR: 13th International Conference on Virtual, Augmented and Mixed Reality
- CCD: 13th International Conference on Cross-Cultural Design
- SCSM: 13th International Conference on Social Computing and Social Media
- AC: 15th International Conference on Augmented Cognition
- DHM: 12th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- DUXU: 10th International Conference on Design, User Experience, and Usability
- DAPI: 9th International Conference on Distributed, Ambient and Pervasive Interactions
- HCIBGO: 8th International Conference on HCI in Business, Government and Organizations
- LCT: 8th International Conference on Learning and Collaboration Technologies
- ITAP: 7th International Conference on Human Aspects of IT for the Aged Population
- HCI-CPT: 3rd International Conference on HCI for Cybersecurity, Privacy and Trust
- HCI-Games: 3rd International Conference on HCI in Games
- MobiTAS: 3rd International Conference on HCI in Mobility, Transport and Automotive Systems
- AIS: 3rd International Conference on Adaptive Instructional Systems
- C&C: 9th International Conference on Culture and Computing
- MOBILE: 2nd International Conference on Design, Operation and Evaluation of Mobile Communications
- AI-HCI: 2nd International Conference on Artificial Intelligence in HCI

List of Conference Proceedings Volumes Appearing Before the Conference

1. LNCS 12762, Human-Computer Interaction: Theory, Methods and Tools (Part I), edited by Masaaki Kurosu
2. LNCS 12763, Human-Computer Interaction: Interaction Techniques and Novel Applications (Part II), edited by Masaaki Kurosu
3. LNCS 12764, Human-Computer Interaction: Design and User Experience Case Studies (Part III), edited by Masaaki Kurosu
4. LNCS 12765, Human Interface and the Management of Information: Information Presentation and Visualization (Part I), edited by Sakae Yamamoto and Hirohiko Mori
5. LNCS 12766, Human Interface and the Management of Information: Information-rich and Intelligent Environments (Part II), edited by Sakae Yamamoto and Hirohiko Mori
6. LNAI 12767, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris and Wen-Chin Li
7. LNCS 12768, Universal Access in Human-Computer Interaction: Design Methods and User Experience (Part I), edited by Margherita Antona and Constantine Stephanidis
8. LNCS 12769, Universal Access in Human-Computer Interaction: Access to Media, Learning and Assistive Environments (Part II), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 12770, Virtual, Augmented and Mixed Reality, edited by Jessie Y. C. Chen and Gino Fragomeni
10. LNCS 12771, Cross-Cultural Design: Experience and Product Design Across Cultures (Part I), edited by P. L. Patrick Rau
11. LNCS 12772, Cross-Cultural Design: Applications in Arts, Learning, Well-being, and Social Development (Part II), edited by P. L. Patrick Rau
12. LNCS 12773, Cross-Cultural Design: Applications in Cultural Heritage, Tourism, Autonomous Vehicles, and Intelligent Agents (Part III), edited by P. L. Patrick Rau
13. LNCS 12774, Social Computing and Social Media: Experience Design and Social Network Analysis (Part I), edited by Gabriele Meiselwitz
14. LNCS 12775, Social Computing and Social Media: Applications in Marketing, Learning, and Health (Part II), edited by Gabriele Meiselwitz
15. LNAI 12776, Augmented Cognition, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
16. LNCS 12777, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Human Body, Motion and Behavior (Part I), edited by Vincent G. Duffy
17. LNCS 12778, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: AI, Product and Service (Part II), edited by Vincent G. Duffy

18. LNCS 12779, Design, User Experience, and Usability: UX Research and Design (Part I), edited by Marcelo Soares, Elizabeth Rosenzweig, and Aaron Marcus
19. LNCS 12780, Design, User Experience, and Usability: Design for Diversity, Well-being, and Social Development (Part II), edited by Marcelo M. Soares, Elizabeth Rosenzweig, and Aaron Marcus
20. LNCS 12781, Design, User Experience, and Usability: Design for Contemporary Technological Environments (Part III), edited by Marcelo M. Soares, Elizabeth Rosenzweig, and Aaron Marcus
21. LNCS 12782, Distributed, Ambient and Pervasive Interactions, edited by Norbert Streitz and Shin'ichi Konomi
22. LNCS 12783, HCI in Business, Government and Organizations, edited by Fiona Fui-Hoon Nah and Keng Siau
23. LNCS 12784, Learning and Collaboration Technologies: New Challenges and Learning Experiences (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
24. LNCS 12785, Learning and Collaboration Technologies: Games and Virtual Environments for Learning (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
25. LNCS 12786, Human Aspects of IT for the Aged Population: Technology Design and Acceptance (Part I), edited by Qin Gao and Jia Zhou
26. LNCS 12787, Human Aspects of IT for the Aged Population: Supporting Everyday Life Activities (Part II), edited by Qin Gao and Jia Zhou
27. LNCS 12788, HCI for Cybersecurity, Privacy and Trust, edited by Abbas Moallem
28. LNCS 12789, HCI in Games: Experience Design and Game Mechanics (Part I), edited by Xiaowen Fang
29. LNCS 12790, HCI in Games: Serious and Immersive Games (Part II), edited by Xiaowen Fang
30. LNCS 12791, HCI in Mobility, Transport and Automotive Systems, edited by Heidi Krömker
31. LNCS 12792, Adaptive Instructional Systems: Design and Evaluation (Part I), edited by Robert A. Sottilare and Jessica Schwarz
32. LNCS 12793, Adaptive Instructional Systems: Adaptation Strategies and Methods (Part II), edited by Robert A. Sottilare and Jessica Schwarz
33. LNCS 12794, Culture and Computing: Interactive Cultural Heritage and Arts (Part I), edited by Matthias Rauterberg
34. LNCS 12795, Culture and Computing: Design Thinking and Cultural Computing (Part II), edited by Matthias Rauterberg
35. LNCS 12796, Design, Operation and Evaluation of Mobile Communications, edited by Gavriel Salvendy and June Wei
36. LNAI 12797, Artificial Intelligence in HCI, edited by Helmut Degen and Stavroula Ntoa
37. CCIS 1419, HCI International 2021 Posters - Part I, edited by Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa

38. CCIS 1420, HCI International 2021 Posters - Part II, edited by Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa
39. CCIS 1421, HCI International 2021 Posters - Part III, edited by Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa

http://2021.hci.international/proceedings

18th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2021)

Program Board Chairs: **Don Harris,** ***Coventry University, UK*** **and Wen-Chin Li,** ***Cranfield University, UK***

- Vincent Ferrari, France
- Maik Friedrich, Germany
- Yueh-Ling Hsu, Taiwan
- Hannu Karvonen, Finland
- Peng Liu, China
- Lothar Meyer, Sweden
- Lei Wang, China
- Jingyu Zhang, China

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-2021.php

HCI International 2022

The 24th International Conference on Human-Computer Interaction, HCI International 2022, will be held jointly with the affiliated conferences at the Gothia Towers Hotel and Swedish Exhibition & Congress Centre, Gothenburg, Sweden, June 26 – July 1, 2022. 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://2022.hci.international/:

General Chair
Prof. Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
Email: general_chair@hcii2022.org

http://2022.hci.international/

Contents

Cognitive Psychology in Aviation

Cognitive Psychology in Air Traffic Control

Studies on Cognitive Processes

Human Error and Human Performance

Cognition and Design

Cognitive Psychology in Aviation

Validation of Virtual Reality Cognitive Assessment for Pilots Across the Lifespan

Oluchi Audu(✉), Kathleen Van Benthem(✉), and Chris M. Herdman(✉)

Carleton University, Ottawa, ON K1S 5B6, Canada
oluchiaudu@cmail.carleton.ca, {kathyvanbenthem,
chrisherdman}@carleton.ca
https://carleton.ca/ace/

Abstract. This research reports on the acceptance and responses to a Virtual Reality (VR) cognitive health screening tool for younger and older pilots. Currently, there are few cognitive assessments that examine the cognitive health of pilots across the lifespan. The cognitive assessments tools for pilots that do exist index a variety of domain-independent functions which do not translate to real-world risks during flights. Furthermore, domain-independent functions such as processing speed, are negatively affected by age thereby making these tools potentially biased against older pilots. CANFLY, a 3-Dimensional (3D) virtual reality simulator, addresses the need for a domain-specific cognitive assessment tool which assesses cognitive functions that pertain to real-world flight such as situation awareness and prospective memory. While CANFLY addresses the problem of validity and generalization to real-world risk, it is also important to ensure that older pilots do not experience the systematic bias that can occur with the use of domain-independent cognitive assessment tools. Some possible age-related issues that could potentially arise with VR cognitive assessment tools include unexpected negative effects (such as simulator sickness or discomfort), or a general lack of acceptance of 3D flight simulation devices amongst older pilots. To examine age effects in VR flight, forty-seven pilots (four females), between the ages of 17 and 71, flew two sessions, the first in a standard full-scale simulator and the second in a VR flight simulator. The tasks in the VR flight were designed as the cognitive health screening tool and indexed key domain-dependent cognitive factors such as situation awareness and prospective memory. After the two sessions, the pilots were also asked to describe their experience with the 3D VR simulator compared to the standard flight simulator. Thus, the present study examined whether or not simulation environment and pilot age had an effect on flight performance. Interactions between the pilot age and effect of the simulation environment were also investigated. Age was also explored as a factor in the acceptance of the VR flight environment and the presence of cybersickness after the VR flight. The results showed that older pilots performed worse for a number of flight tasks, but that there was no interaction effect between age and flight simulation environment on situation awareness and prospective memory. There was a preference for the VR simulation over the full-scale simulator, and this was seen in both age groups. No effect of age was found for the cybersickness measures, although there was a small trend for pilots of all ages to experience slightly increased symptoms associated with queasiness after the VR flight. Findings from the present research show that

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 3–18, 2021.
https://doi.org/10.1007/978-3-030-77932-0_1

older pilots are not likely to experience bias from the VR technology or cybersickness symptoms in a VR cognitive health screening tool. Results support the use of VR as a useful platform for evaluating domain-dependent cognition for pilots across the lifespan.

Keywords: Cognitive-assessment older pilots · VR flight simulator · CANFLY · Human-computer interaction · Cognition · Aging

1 Introduction and Background

The general aviation pilot population is growing older. As of 2019, the average age of U.S. general aviation pilots stood at 44.2 years [1]. The average age of Canadian general aviation pilots in 2017 had risen to 60 years, which is 5 years older than the mean age reported in 2012 [2]. Furthermore, it is currently typical for general aviation pilots to fly beyond the age of 70, even into their 80's and 90's [3]. This aging cohort of pilots raises important questions about the relationship between age and piloting performance as well as viable means with which older pilot's could be assessed and deemed fit for flying. In an attempt to address the notable trend of age advancement in the aviation sector, the Federal Aviation Administration (FAA) established the "Age-60" rule that required airline pilots to retire upon reaching their 60th birthday. However, there has been a persistent debate regarding the efficacy of this rule. The major argument that has been posed against this notion is based on the fact that The Age 60 Rule was not based on any scientific data showing that general aviation pilots aged 60 and older were any less safe than younger pilots [4]. Furthermore, various research has shown no apparent medical, scientific or safety justification for The Age 60 Rule and pilots in general aviation have also stated that the use of age as a single criterion of older pilot fitness represents age discrimination in the aviation sector [5].

Alternative ways to assess pilot's fitness would be to develop individually-based pilot performance assessment. In terms of physical assessment, the FAA requires the examination of general aviation pilots every year and this involves typically vision tests, Electrocardiogram (ECG), lung function, chest X-ray, ear function and so on [6]. On the other hand, there are currently few assessments that examine the cognitive health of pilots across the lifespan despite the significant role cognition plays in pilot performance. To address the need of a cognitive assessment tool specifically for pilots in general aviation, the CANFLY, a virtual reality cognitive screening tool, was developed and is undergoing validation tests. In the CANFLY, the cognitive screening is embedded within standard flying tasks which in turn are presented within realistic flight scenarios for general aviation pilots. While CANFLY addresses the need for a domain-dependent cognitive assessment tool, it is important to ensure that this system accommodates older pilots, who, in the past, have experienced systemic bias when tested with standard cognitive assessments [5]. Thus, the goal of the present research was to examine if either flight simulation environment (standard full-scale versus VR flight simulator) or age group had any main or interaction effects on domain-dependent cognitive performance. This research also explored whether age affected the acceptance of a VR flight environment and if the VR flight simulation produced any cybersickness in pilots. This research aims to identify any potential disadvantages the VR simulator may have towards older pilots.

1.1 Cognition and Flying

In general aviation, the significance of screening for cognitive dysfunction in pilots cannot be downplayed, as this risk management strategy could reduce accidents and potentially save the lives of travellers. The association between pilot cognition and accident risk is apparent in the annual reports linking pilot-related to 75% of accidents within the General Aviation sector [7]. Furthermore, various research has shown that pilot judgement was associated with 66% of crashes [8] while most of the pilot-related accidents occur during takeoff or landing, which are both situations where demands on pilot cognition are greatest [9].

Domain dependent cognitive factors associated with the field of general aviation include: Prospective memory and situation awareness. Prospective memory refers to the cognitive function that involves remembering to carry out a planned action or recall a planned intention in the future [10]. There are generally three types of prospective memory, the first, prospective memory proper, involves holding an intention in memory over a period of time (e.g. remembering to file a flight plan). The second type of prospective memory involves memory for vigilance which relies on maintaining an intention in one's consciousness while simultaneously engaging in other tasks (e.g. checking altitude at regular intervals while flying). Lastly, habitual or routine prospective memory involves remembering to perform tasks that are performed routinely, such as remembering to complete radio calls at a regular location in a flight pattern. Dismukes (2007) found that pilots performed all three types of prospective memory tasks during routine flights thereby providing evidence for the significance of prospective memory while flying [11]. Studies have shown that flight-related prospective memory failures typically involve conditions involving high workload [12].

Situation awareness is another significant cognitive factor involved in pilot performance. Situation awareness can be simply described as one's knowledge and perception of environmental elements and events [13]. Situation awareness is not just being aware of what the current events entail, but also being able to understand what is likely to occur in the nearest future given the current situation and this is a significant prerequisite for decision making, particularly in dynamically changing situations. In order to establish the meaning of "knowing what is going on" and what that entails, Endsley (1995) developed a model which posits three distinct levels of situation awareness [14]. The first level is the perception of elements in the environment which involves the knowledge and perception of objects, events, people, and environmental factors. Perception of elements also involves the current state of these elements such as their location, condition, and actions. In general aviation, an example of this level of situation awareness would be the pilots accurately perceiving their aircraft and it's system, as well as perceiving the weather conditions and emergency information if need be [15]. The next level of situation awareness is the comprehension of the current situation which involves integrating the information perceived so as to understand the significance of these elements to the desired goal or outcome. For example, a pilot forming a mental picture using the information gathered while perceiving elements in the environment. The final level of situation awareness is the projection of future status which is the highest level of situation awareness and involves the ability to project the future status of elements in the environment. After perceiving and comprehending the situation, it is important to use this information to predict likely future status of the environment which would potentially be important for decision making. For example, a pilot anticipating a need to refuel and making the decision to land in the next airfield based on that anticipation.

1.2 Ecological Validity of Cognitive Assessment Tools

Despite the significant role cognition plays in pilot performance, there are currently only a few cognitive assessments that examine the cognitive health of pilots across the lifespan. The most prominent one is CogScreen AE [16] which is a cognitive screening tool that was designed to measure the underlying cognitive and information processing abilities associated with flying [17]. The CogScreen battery indexes a variety of *domain-independent* functions, such as processing speed and working memory. The problem with this tool and other assessments of *domain-independent* cognitive functions is their limited ecological validity and weak application to real-world risk during flight. Furthermore, domain-independent cognitive functions (e.g., processing speed) are notoriously negatively affected by older age [18]. One of the prominent arguments presented by older pilots in a 2005 lawsuit against the FAA's 'Age 60' rule stated that CogScreen - AE does not sufficiently identify age-related cognitive functions that would impact pilot performance [19].

Rather, what is needed in the aviation sector are cognitive screening tools that index *domain-dependent* cognitive functions which are factors that actually pertain to real world flight such as pilot situation awareness and prospective memory. It is important that the cognitive screening tool shows strong associations with critical incidents and are not age-biased [20]. To address the need of an ecologically valid cognitive assessment tool in the general aviation sector, Van Benthem & Herdman (2014) developed and examined the efficacy of a virtual reality cognitive health screening tool integrated into simulated flight scenarios [20]. CANFLY [20] is a virtual-reality based flight simulator which indexes domain-dependent cognitive factors such as the aforementioned prospective memory and situation awareness. The cognitive screening process is embedded within standard flying tasks which in turn were in turn presented within realistic flight scenarios for General Aviation pilots.

1.3 VR Symptoms and Age

It is important to ensure that older pilots do not experience systematic bias that typically occurs while using cognitive assessment tools. With virtual reality, some age-related issues that could potentially arise include VRISE (virtual reality induced symptoms and effects) which is an aspect of the VR user experience that includes nausea, queasiness, dizziness, disorientation or a general discomfort while using the Virtual reality tool [21]. Various studies have shown that the use of head mounted devices while using VR interfaces increases the chances of a user experiencing VRISE [22, 23]. In addition, physiological changes over the lifespan, such as a decline in the visual and vestibular senses, could potentially make older people more prone to simulator sickness while using virtual reality tools [24].

1.4 Present Research

The present work examined whether age had an effect on pilot user experience with a virtual reality cognitive assessment tool. The major goals for this study included: examining whether age had an effect on both pilot performance and interaction with a VR cognitive assessment tool, comparing the physical responses and acceptance of the VR flight simulator to that of a standard 2-dimensional full-scale flight simulator, and lastly identifying any potential disadvantages the VR flight simulator may have posed, particularly towards older pilots. As mentioned above, the flight activities in the VR flight simulator were designed for cognitive screening, whereby key domain-dependent cognitive factors were assessed during the flight.

2 Method

2.1 Participants

Forty-seven pilots (four females) flew two sessions, the first in a standard full-scale flight simulator and the second in a VR flight simulator. The pilots were between the ages of 17 and 71 years, certified with a minimum of a current student permit and solo cross-country experience. Pilots were recruited from local flying clubs associations, and flight training units. Inclusion criteria included having a valid aviation medical certification and having logged a minimum of one pilot-in-command hour in the previous 24 months. Participants completed informed consent procedures in accord with the university ethics board requirements.

2.2 Study Instruments

Full-Scale Flight Simulator. Participants flew an aircraft simulator constructed from a Cessna 172 Fuselage and cockpit. The cockpit displayed flight information via a virtual ‘6-pack’ (i.e., the primary flight controls) and was equipped with a variety of controls including yoke, throttle, and flaps. The graphics were produced by Prepar3D (Lockheed Martin) “on the fly” and were modeled after real world aerodromes and surrounding terrain (Fig. 1).

Virtual Reality Flight Simulator. The 360° virtual flight environment included the interior/exterior of the Cessna 172 and all the exterior terrain details were delivered via an Oculus Rift VR headset. The virtual environment was rendered using Lockheed Martin’s Prepar3D software. As shown in Fig. 2, to complement the functionality of the virtual cockpit, a prototype flight control unit was designed and built in-house. The location of the simulator controls mirrored the location of controls present in the VR Cessna 172 cockpit (e.g., yoke, throttle, and flaps) (Fig. 3).

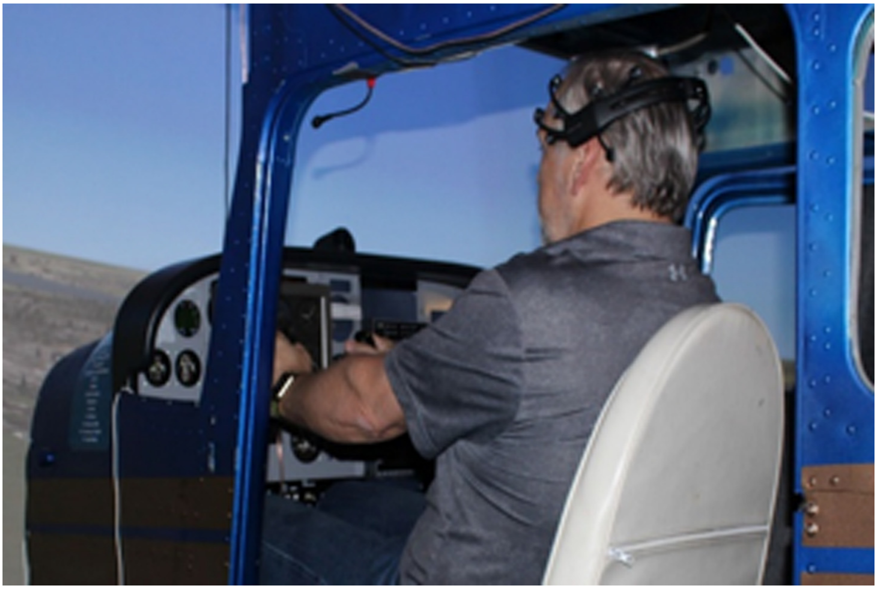


Fig. 1. Environment and interior of full-scale Cessna 172 flight simulator (left door removed).

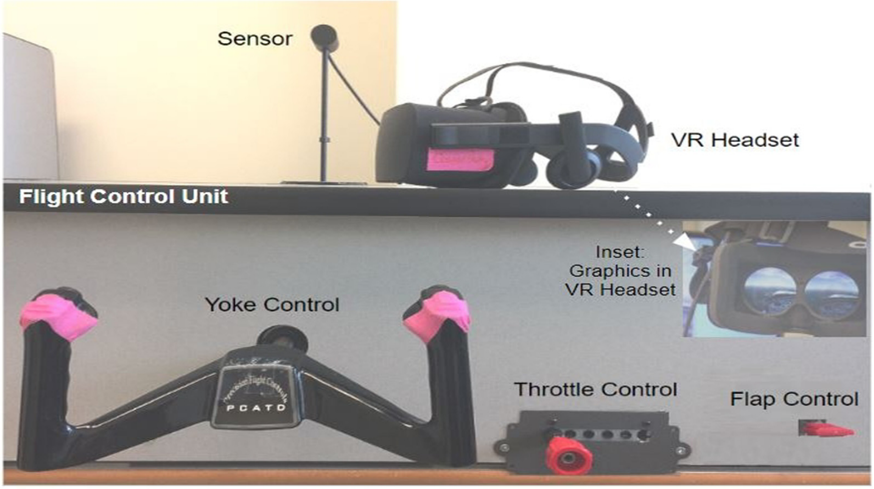


Fig. 2. Flight control unit: An insert figure (top right of flight control unit) shows the VR goggles and sample graphical display. Labels in figure are added

2.3 Procedure

Participants flew two sessions, the first in the full-scale flight simulator and the second in the VR flight simulator. In session one, after a full briefing regarding the flight procedures, demographics and pilot experience questionnaire and simulator training flight

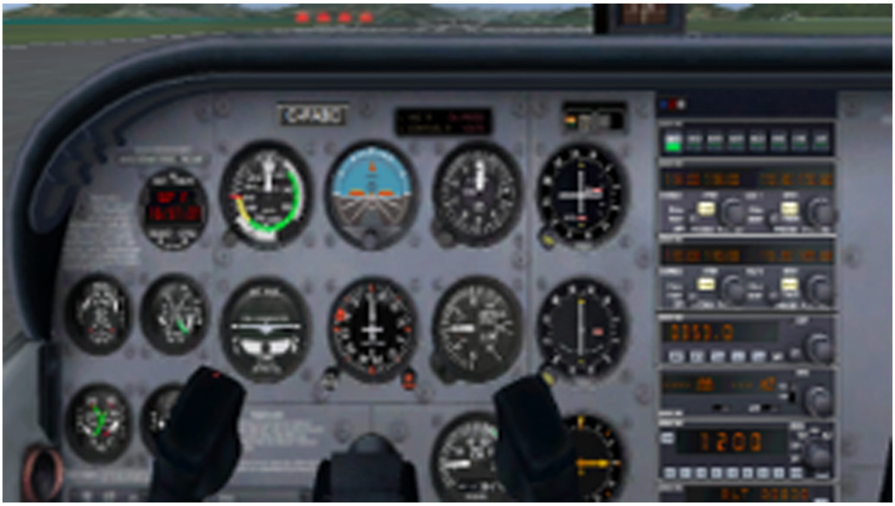


Fig. 3. Views of the virtual cockpit as seen through the VR goggles.

and practice, participants flew a one-hour cross country flight. The flight tasks were designed to represent a search and rescue familiarization, which included changes in altitude and flight over both a level and hilly terrain. Throughout the flight, participants were required to make as well as listen to radio calls from other aircrafts while completing tasks that measured prospective memory and situational awareness. As part of a larger research agenda on aging, pilot performance, and biometric indices of mental workload, researchers monitored for all occurrences of critical events during the one-hour flight and participants were also provided with a wireless electroencephalography (EEG) headset and biometric measures wristband to wear during the flight simulation. In the second session, participants received training and practice in the 3D VR flight simulator. They practiced the prospective memory task, the situation awareness queries and how to follow the prescribed four legs of the test simulation route. During the VR flight simulation, participants wore head-mounted VR goggles, which also delivered engine sounds and messages from other aircrafts and air traffic controllers (via external headset speakers). The pilots flew four legs, which included take-off, pattern activity, cross-country flight, approach and landing. The tasks in the second session flight were designed as the cognitive health screening tool and indexed key domain-dependent cognitive factors (situation awareness, prospective memory, task management, and task-related flight performance). Afterwards, the participants were asked a series of user-experience questions pertaining to the VR flight simulator. Queries indexed their preference for either the first standard simulator versus the VR simulator, and their history with motion sickness and other VR products. Before and immediately after the VR flight, participants were also asked to rate (on a scale from 1 to 6) their perceived level of cybersickness symptoms: queasiness, dizziness, and disorientation. A cybersickness variable was derived from the difference between the post- and - pre-flight summed symptom score.

2.4 Study Measures

Age. For the purposes of age-related analyses, the age groups were split at age 55, with pilots aged 17 to 54 belonging to the younger age group.

Situation Awareness. Situation awareness performance was measured using queries presented during brief pauses occurring between the segments of each flight. In the full-scale simulator there were two sets of queries and in the VR simulator there were four sets of queries. For level 1 SA the queries focused on the pilots ability to detect and retain information regarding the call signs, intentions, and aircraft type for the other aircraft. In each leg of the flight there were two or three other aircraft that they heard making radio calls, via their headset. Scoring reflected the amount of detail that could be recalled, with scores ranging from 0 (no correct details) to 4 (complete details). An example of a level 1 SA query is "Please report the call signs for all the aircraft you heard during this segment of flight". The level 2 SA measure was based on the accuracy of the pilots reported ownship (where they located themselves on a map) and the position of all other aircraft they heard during that segment of the flight. The level 2 SA scores ranged from 0 (beyond an established perimeter of the actual location) and 2 (very close to the actual position). These level 2 SA queries were conducted following the level 1 SA queries, during the pauses. The percent correct scores for each session were totalled separately for levels 1 and 2 SA. The levels 1 and 2 SA scores were standardized separately (using a mean of 1 and standard deviation of 1) to permit statistical comparison of each level of SA across both simulator conditions.

Prospective Memory. Prospective memory was indexed via the pilots ability to remember to make radio calls when predetermined events occurred during the flight. In the full-scale simulator, the prospective memory task related to making the one of two pre-set radio calls whenever they detected a right facing arrow on a display to the right side of the cockpit (but out of view). In the VR environment, pilots were instructed to push the radio call button whenever they heard the word "traffic" being spoken by other aircraft or ground services personnel on their radio. Both prospective memory scores were converted to the percentage of correctly completed calls. While these indices of prospective memory differed in their associated trigger (one was visual and one was auditory) both relied on detecting peripheral cues in the environment, and then completing a radio communication-like task. The percent correct scores for each session were totalled and a percent correct was calculated. All scores for both sessions were standardized (mean of 1 and standard deviation of 1) to permit statistical comparison across both simulator conditions.

Simulator Environment Preference. Simulator preference was indexed via two queries that were posed at the end of the second session. This query asked pilots to report on their preferred simulator environment and then provide open-ended comments as to why they selected either the full-scale or the VR simulator as their preferred simulation environment.

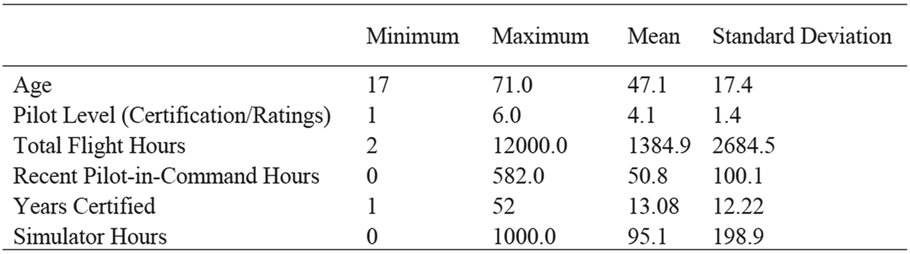
Cybersickness. Cybersickness symptoms were based on Kennedy et al. (1993) Simulator Sickness Questionnaire (SSQ) [25]. The full SSQ queries 16 criteria associated with cybersickness (e.g., salivation, nausea, dizziness, disorientation, vertigo). However, factor analysis has shown that all of the cybersickness symptoms from the SSQ load onto three distinct factors that can be characterized as related to nausea (or queasiness), dizziness, and disorientation [26]. Thus for expediency, in this experiment only three queries were presented at the beginning and the end of session 2 (pre and post VR simulator flight). The three queries used a seven-point self-rating scale and asked participants to indicate their perceived level of queasiness, dizziness, and disorientation.

3 Results

3.1 Pilot Age and Expertise

The age range of the pilots was 17 to 71 years. Pilot certification levels were categorized into six groups, ranging from 1-student pilot, 2-recreational permit, 3-visual flight rules (VFR) private pilot with no additional ratings, 4-VFR private pilot with additional ratings, 5-Instrument flight rules (IFR) private pilot, and 6-airline, air transport, or military pilot. The most common certification was "4-VFR private pilot with additional ratings". As shown in Table 1, total flight hours, recent pilot-in-command hours, number of years certified and total simulator hours were also collected.

Table 1. Age and experience ranges



	Minimum	Maximum	Mean	Standard Deviation
Age	17	71.0	47.1	17.4
Pilot Level (Certification/Ratings)	1	6.0	4.1	1.4
Total Flight Hours	2	12000.0	1384.9	2684.5
Recent Pilot-in-Command Hours	0	582.0	50.8	100.1
Years Certified	1	52	13.08	12.22
Simulator Hours	0	1000.0	95.1	198.9

3.2 Comparison of Performance in Full-Scale Versus VR Simulator

A series of univariate factorial ANOVAs were conducted to investigate whether standardized scores for situation awareness and prospective memory would differ between simulator environments. The second between-group factor was age group, and was used to determine whether there was a main effect of age on flight performance and whether age might interact with simulator environment.

Situation Awareness Level 1

There was a significant effect of age on level 1 SA, such that older pilots performed with lower accuracy than the younger pilots, $F(3,84) = 21.30$, $p < .001$. There was no main effect of simulator environment ($p > .1$). However, as shown in Fig. 4, there was a trending, but non-significant interaction between simulation environment and age, where in the VR simulator the effect of age was less pronounced than in the full-scale simulator, $F(3, 84) = 3.087, p = .08$.

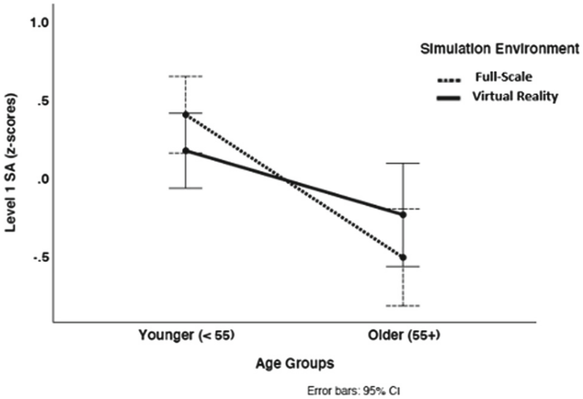


Fig. 4. Interaction between age and simulator environment on level 1 situation awareness

Situation Awareness Level 2
For level 2SA, there were no significant main effects of age, $F(1, 89) = 2.07, p = .15$, or simulator environment, p > .1. There was also no significant interaction between simulation environment and age, $F(1, 89) = 1.763, p = .19$.

Prospective Memory
There was a significant trend regarding the effect of age on prospective memory, such that older pilots had more prospective memory failures than the younger pilots, $F(1{,}94) = 3.16, p = .08$. However, this effect appeared to be driven by a significant interaction between simulation environment and age, $F(1, 94) = 6.113, p = .02$: as shown in Fig. 5, the negative effect of age was seen primarily in the Full-Scale simulator condition.

3.3 Simulator Preference and Cybersickness

As shown in Fig. 6, pilots expressed a preference for the VR simulator (25/45) over the full-scale simulator (17/45). The preference for the VR simulator was seen in both older and younger pilots, with no effect of age found for simulator preference, $X^2 = 1.62, p = 0.44$. Two participants were ambivalent regarding their simulator preference and were not included in this analysis.

Open-ended Responses to Simulator Preference. A qualitative review of the open-ended responses to the simulator preference item was conducted. With respect to participant experience flying in the VR environment wearing the head mounted device (HMD) compared to standard full-scale simulator, three themes emerged: realism/fidelity, enjoyment/ease, and immersiveness. Most of the pilots described the VR simulation as more enjoyable and realistic and a more veridical representation of a real-life flight. Eight

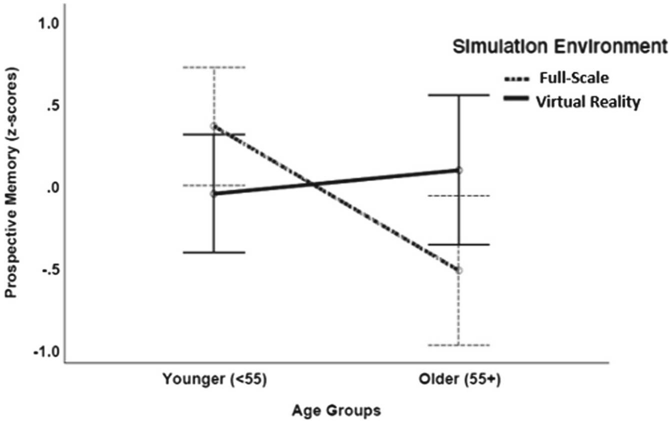


Fig. 5. Interaction between age and simulator environment on prospective memory

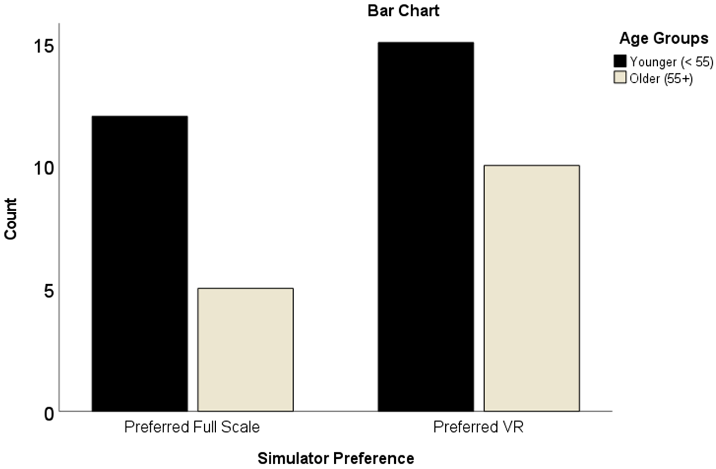


Fig. 6. Simulator preference by age group

pilots indicated that their experience in the VR simulation was quite similar to that of the standard full-scale simulator. Negative comments about the VR flight simulator were relatively rare and tended to reflect discomfort with the HMD and some visual symptoms such as blurred graphics.

Cybersickness. With respect to the cybersickness queries, a series of repeated measures ANOVAs were conducted to investigate the effect of the VR flight simulation on queasiness, dizziness, and disorientation (pre and post flight). Age group was used as a between group factor to investigate whether pilot age would interact with any VR flight simulation effects.

Queasiness. There was a slight increase in queasiness symptoms after the VR flight, however this trend was not significant, $F(1, 45) = 3.87, p = .06, \eta_p^2 = .08$. As illustrated in Fig. 7, there was no main effect or age or significant interaction between pre and post VR flight and age group, $p > .1$.

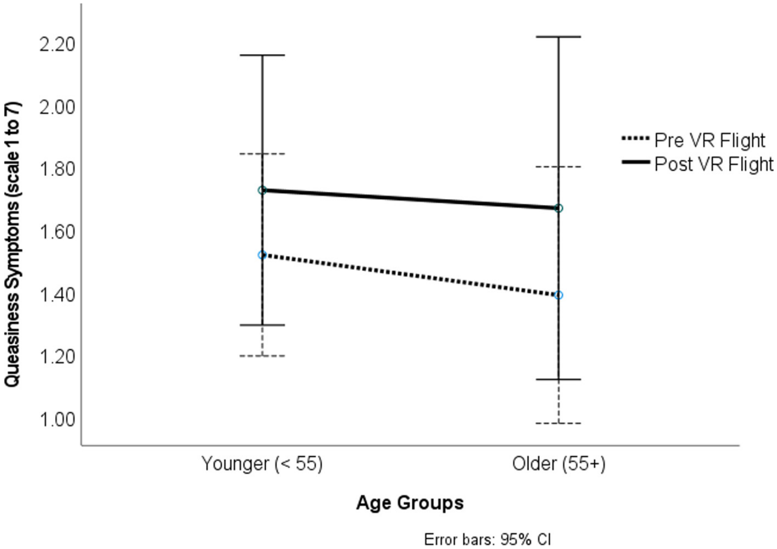


Fig. 7. Effects of VR flight simulation on Queasiness

Dizziness
There was no significant main effect of the VR flight simulation on dizziness, $F(1, 45) = 0.69, p = .41, \eta_p^2 = .02$. There was also no main effect or age or significant interaction between pre and post VR flight and age group, $p > .1$.

Disorientation
There was a significant main effect of the VR flight simulation on disorientation, such that the post flight disorientation (M = 1.68) was higher than the pre-flight disorientation (M = 1.36), $F(1, 45) = 6.89$, $p = .012$, $\eta_p^2 = .13$. There was no main effect of age or significant interaction between pre and post VR flight and age group, $p > .1$.

4 Discussion

As part of an effort to ensure that older pilots do not experience systematic bias that typically occurs while using cognitive assessment tools, the present work provides evidence towards validating a virtual reality cognitive assessment tool for pilots. User-experience queries were incorporated into the cognitive screening process to investigate pilot age might have an effect on preference of simulation environment or cybersickness and thus disadvantage older pilots during the VR cognitive screening. Before and after the VR cognitive screening, participants rated their symptoms of queasiness, dizziness, and disorientation. Also, pilot performance (measured by tasks relating to situation awareness and prospective memory) in both the VR and full scale flight simulators was observed.

The results show that there was a significant deleterious effect of age on level 1 SA. This implies that in the level 1 SA tasks, in which the participants were required to answer questions about aircraft call signs, type and intention, older pilots answered these queries with lower accuracy compared to younger pilots. On the other hand, there was a non-significant interaction between simulation environment and age thereby implying that the simulation environment played no role in the age effect seen on performance in level 1 SA. However, as shown in Fig. 4, there was a trending but non-significant interaction between simulation environment and age such that the difference in performance caused by age was slightly less prominent in the VR simulation. This was an expected result as it suggests that the VR simulator may serve to eradicate issues that older pilots could have experienced in the full scale simulator. This result pattern is quite unique as previous studies in the field typically highlight challenges VR poses towards older users [21, 27].

Furthermore, the results showed no significant effect of age on level 2 SA, suggesting that in the level 2 SA tasks, in which the participants were asked to answer questions regarding the location of themselves and others in the simulation, older pilots performed with similar accuracy as the younger pilots. There was also no significant interaction between simulation environment and age meaning that the older and younger pilots performed with similar accuracy in both the full scale simulator and VR simulator.

The prospective memory tasks required the participants to remember to make radio calls at intervals. The results showed that there was a significant trend regarding the effect of age on performance such that older pilots had more prospective memory failures than the younger pilots. This negative effect of age appeared to be driven by the significant interaction between simulation environment and age as the effect was primarily seen in the Full-Scale simulation. The detrimental effects of age on prospective memory are quite evident in the literature as various studies have shown that older people are more likely to forget intended actions compared to younger people [28, 29].

There was a general preference for the VR simulator over the full-scale simulator. This preference most likely stems from the various advantages of VR simulation. Firstly, compared to a standard simulator, VR allows for a fully immersive and more engaging experience for the participants such that the simulation greatly resembles a real-life flight. Also, being a relatively novel technology, the general preference for VR could have been as a result of the enthusiasm from experiencing a VR flight simulation. This acceptance of the VR flight simulation compared to the full scale also extended to the physical symptoms involved while using the simulation as there was no significant effect of time in the VR simulator on queasiness or dizziness. However, there was a small but significant increase in disorientation symptoms after the VR flight. There were no significant interactions between pre and post VR flight cybersickness symptoms and age group; thereby, implying that older pilots were not disadvantaged (in terms of experiencing cybersickness) while using the VR flight simulation.

Concerns over the use of virtual reality with older populations stem from the possibility that physiological changes over the lifespan such as a decline in the visual and vestibular senses could potentially make older people more prone to simulator sickness while using virtual reality tools [24]. Excessive optic flow in the periphery and the incongruence between visual and vestibular messaging could lead to negative physical symptoms amongst older people using VR systems [24, 30]. In addition, various studies have shown that older participants in highly immersive virtual environments were more likely to experience cybersickness when compared to younger participants [24, 30, 31]. The present work is encouraging in that cybersickness symptoms were minimal not only amongst younger participants, but also older participants.

There are explanations for why the VR flight simulation did not elicit negative physical symptoms among the participants. The first pertains to the short duration of the test. Excluding the 15-min briefing and practice, the participants spent approximately 30 min within the Virtual reality flight simulation. Studies have shown that the longer one spends within a virtual environment, the more likely the chances are of getting simulator sickness [32]. Hence, the relatively short amount of time spent within the VR simulator could be a potential explanation as to why participants did not experience cybersickness. It is also significant to note that the participants were provided with brief intermittent breaks where no visual information was displayed and all virtual movement was suspended which possibly reduced the chances of experiencing simulator sickness.

5 Conclusion

Although there is interest and motivation in the field to improve older pilot safety, there has been relatively little development in ecologically valid cognitive assessment of older pilots. The results of this study show that the head-mounted VR system was well accepted by both older and younger pilots. These findings imply that VR could potentially be a fair and ecologically valid cognitive assessment tool for older pilots. Also, this acceptance of VR simulation can be applied to other domains such as driving, where domain-specific cognitive testing would be as significant.

References

1. Statista Homepage. https://www.statista.com/statistics/910024/average-age-pilots-in-the-united-states/#:~:text=This%20statistic%20illustrates%. Accessed 11 Feb 2021
2. Keith, C.: Report on the COPA 2017 membership survey. Ottawa, ON. https://copanational.org/sites/copanational.org/wp-content/uploads/2017/02/2017_Membership_survey_eng.pdf. Accessed 11 Feb 2021
3. Aging and the General Aviation Pilot. https://www.aopa.org/-/media/Files/AOPA/Home/Pilot-Resources/Safety-and-Proficiency/Accident-Analysis/Special-Reports/1302agingpilotreport.pdf. Accessed 11 Feb 2021
4. U.S. Department of Transportation. https://www.transportation.gov/testimony/faa%E2%80%99s-age-60-commercial-pilot-rule. Accessed 11 Feb 2021
5. Odenheimer, G.: Function, flying, and the age-60 rule. J. Am. Geriatr. Soc. **47**, 910–911 (1999). https://doi.org/10.1111/j.1532-5415.1999.tb03854.x
6. Guide For Aviation Medical Examiners. https://www.faa.gov/about/office_org/headquarters_offices/avs/offices/aam/ame/guide/. Accessed 11 Feb 2021
7. Bazargan, M., Guzhva, V.: Impact of gender, age and experience of pilots on general aviation accidents. Accident Anal. Prevent. **43**, 962–970 (2011). https://doi.org/10.1016/j.aap.2010.11.023
8. Oster, C., Strong, J., Zorn, C.: Why Airplanes Crash. Oxford University Press, New York (1992). https://global.oup.com/academic/product/why-airplanes-crash-9780195072235?cc=ca&lang=en&. Accessed 11 Feb 2021

9. Rebok, G., Qiang, Y., Baker, S., Li, G.: Pilot age and error in air taxi crashes. Aviat. Space Environ. Med. **80**, 647–651 (2009). https://doi.org/10.3357/ASEM.2369.2009
10. Finstad, K., Bink, M., McDaniel, M., Einstein, G.: Breaks and task switches in prospective memory. Appl. Cogn. Psychol. **20**, 705–712 (2006). https://doi.org/10.1002/acp.1223
11. Dismukes, R.: Prospective memory in workplace and everyday situations. Curr. Direct. Psychol. Sci. **21**, 215–220 (2012). https://doi.org/10.1177/0963721412447621
12. Van Benthem, K., Herdman, C., Tolton, R., LeFevre, J.: Prospective memory failures in aviation: effects of cue salience, workload, and individual differences. Aerosp. Med. Hum. Perform. **86**, 366–373 (2015). https://doi.org/10.3357/amhp.3428.2015
13. Endsley, M.: Situation awareness misconceptions and misunderstandings. J. Cogn. Eng. Decis. Making **9**, 4–32 (2015). https://doi.org/10.1177/1555343415572631
14. Endsley, M.: Towards a new paradigm for automation: designing for situation awareness. IFAC Proc. Vol. **28**, 365–370 (1995). https://doi.org/10.1016/s1474-6670(17)45259-1
15. Nguyen, T., Lim, C., Nguyen, N., Gordon-Brown, L., Nahavandi, S.: A review of situation awareness assessment approaches in aviation environments. IEEE Syst. J. **13**, 3590–3603 (2019). https://doi.org/10.1109/JSYST.2019.2918283
16. Kay, G.: CogScreen aeromedical edition professional manual (professional manual). 824 Florida: Psychological Assessment Resources Inc. (1995)
17. Chee, S., Bigornia, V., Logsdon, D.: The application of a computerized cognitive screening tool in naval aviators. Milit. Med. **186**, 198–204 (2021). https://doi.org/10.1093/milmed/usaa333
18. Eckert, M., Naom, K., Roberts, D., Calhoun, V., Harris, K.: Age-related changes in processing speed: unique contributions of cerebellar and prefrontal cortex. Front. Hum. Neurosci. (2010). https://doi.org/10.3389/neuro.09.010.2010
19. Examining the federal aviation administration's age 60 rule. https://www.govinfo.gov/content/pkg/CHRG-109shrg63516/html/CHRG-109shrg63516.htm. Accessed 11 Feb 2021
20. Van Benthem, K., Herdman, C.M.: CanFly! A distinctly canadian approach to promoting safe aviation in the third age. Presentation for the Canadian Owners and Pilots Association. Flight 8 Meeting, 23 September 2014, Ottawa, ON (2014)
21. Arns, L., Cerney, M.: The relationship between age and incidence of cybersickness among immersive environment users. In: IEEE Proceedings of the Virtual Reality, VR 2005 (2005). https://doi.org/10.1109/VR.2005.1492788
22. Sharples, S., Cobb, S., Moody, A., Wilson, J.: Virtual reality induced symptoms and effects (VRISE): comparison of head mounted display (HMD), desktop and projection display systems. Displays **29**, 58–69 (2008). https://doi.org/10.1016/j.displa.2007.09.005
23. Cobb, S., Nichols, S., Ramsey, A., Wilson, J.: Virtual reality-induced symptoms and effects (VRISE). Presence: Teleoper. Virtual Environ. **8**, 169–186 (1999). https://doi.org/10.1162/105474699566152
24. Bermúdez Rey, M., Clark, T., Wang, W., Leeder, T., Bian, Y., Merfeld, D.: Vestibular perceptual thresholds increase above the age of 40. Front. Neurol. **7**, 162 (2016)
25. Kennedy, R., Lane, N., Berbaum, K., Lilienthal, M.: simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int. J. Aviat. Psychol. **3**, 203–220 (1993). https://doi.org/10.1207/s15327108ijap0303_3
26. Balk, S., Bertola, M., Inman, V.: Simulator sickness questionnaire: twenty years later. In: Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: driving assessment 2013 (2013). https://doi.org/10.17077/drivingassessment.1498
27. Knight, M., Arns, L.: The relationship among age and other factors on incidence of cybersickness in immersive environment users. In: Proceedings of the 3rd symposium on Applied perception in graphics and visualization - APGV 2006 (2006). https://doi.org/10.1145/1140491.1140539

28. Chen, S., Zhou, R.: Age-related declines in prospective memory: modulation of the prospective and retrospective components. Acta Psychol. Sinica. **42**, 640–650 (2010). https://doi.org/10.1007/978-3-540-73281-5_72
29. West, R., Craik, F.: Age-related decline in prospective memory: the roles of cue accessibility and cue sensitivity. Psychol. Aging **14**, 264–272 (1999). https://doi.org/10.1037/0882-7974.14.2.264
30. Ramkhalawansingh, R., Butler, J., Campos, J.: Visual–vestibular integration during self-motion perception in younger and older adults. Psychol. Aging **33**, 798–813 (2018). https://doi.org/10.1037/pag0000271
31. Stauffert, J., Niebling, F., Latoschik, M.: Latency and cybersickness: impact, causes, and measures. A review. Front. Virtual Reality **1** (2020). https://doi.org/10.3389/frvir.2020.582204

Culture's Consequences on the Categorisation of Causal Factors in Aviation Accident Reports

Wesley Tsz-Kin Chan(✉) and Wen-Chin Li

Safety and Accident Investigation Centre, Cranfield University, Cranfield, UK
wesley.chan@cranfield.ac.uk

Abstract. In aircraft accident investigations, the classification of causal factors to different parts of the organisational hierarchy ranging from upper-level management to frontline operators has been known to differ depending on accident investigators' national culture backgrounds. There however remains a research gap in assessing how within-group, intra-cultural variations between individual values and attitudes can affect the attribution of accident causes, particularly in culturally diverse environments. The present research involved thirty (N = 30) accident investigators' categorisation of causal factors of the same incident using the Human Factors Analysis and Classification System (HFACS) [1]. Rather than using geopolitical countries to categorise participants into groups, individual cultural values were assessed using Hofstede's six-dimension model in order to capture intra-cultural variations [2]. The frequency of participants selecting a HFACS category as contributory were compared with their six Hofstede's dimensions. It was discovered that the attribution of accident causes to categories in the *Unsafe Acts of Operators* level (HFACS Level 1) were driven by the investigators' *Power Distance* and *Long-Term Orientation* dimensions; categories in the *Preconditions of Unsafe Acts* level (HFACS Level 2) were unaffected by investigators' Hofstede's dimensions; and the higher HFACS levels concerning supervisory and organisational influences (Levels 3 and 4) were moderated by *Uncertainty Avoidance* and *Long-Term Orientation*. Findings of the present research can provide insight into sources of bias in accident investigations and inform adjustments for consequential training and operational changes.

Keywords: Accident investigation · Aviation training · Cultural differences

1 Introduction

A primary goal of aviation accident investigations is to objectively highlight human factors errors in the lead up to the accident. Whilst the existing literature indicates that approximately 60 to 80% of accidents were at least partly caused by human error [3], it is inappropriate to simply attribute aircrew error as the sole cause of aircraft accidents. Errors committed by the crew should be considered as the end result of a number of underlying factors in the causal sequence of events. The objective of accident investigations is to identify these underlying deficits in task performance, so that remedial training and design strategies can be formulated accordingly [4].

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 19–27, 2021.
https://doi.org/10.1007/978-3-030-77932-0_2

The Human Factors Analysis and Classification System (HFACS), diagrammatically described in Fig. 1, is an analytical framework for the categorisation of underlying human causes of aviation accidents [1]. The model provides 18 categories of errors, spread across four levels, for which causal factors of accidents can be attributed. The first level of HFACS, Unsafe Acts of Operators, classifies the active failures of frontline operators. The second level, Preconditions for Unsafe Acts, encompasses both latent and active failures in the context of substandard conditions and practices. The third level of Unsafe Supervision traces the latent failures up to the front-line supervisory level. The fourth and highest level, Organisational Influences, involves latent failures at the upper levels of management, such as unsatisfactory supervisory practices and decisions. As the framework provides a taxonomy representing system-wide inadequacies leading to human errors, ranging from the conditions of frontline operators to failures at the organisational level, it provides users with the ability to develop targeted and data-driven intervention strategies [5].

Although HFACS encompasses system-wide aspects of human error, accident investigators' categorisation of accident causes into the HFACS model remains a subjective pursuit. For example, the same error may be categorised by some investigators as a front-line operator fault, whereas others may consider it a supervisory deficiency. Investigators' own backgrounds may change the yardsticks and norms used to assess behaviours of other groups, and the interpretation of language and culturally accepted ways of doing things may become 'lost in translation' [6]. In a comparison of British and Taiwanese investigators' attribution of causal factors using HFACS, Li et al. [7] found significant differences in the utilisation of Perceptual Error, Adverse Mental States, Planned Inadequate Operations, and Organisational Climate categories. They found British investigators more likely to attribute shortcomings to higher-level management than their Taiwanese counterparts, possibly due to cultural differences in respect and deference to the organisational hierarchy. They also found that Taiwanese participants underused the category of Adverse Mental States, instead opting to use with greater frequency the Perceptual Error category, which carried less social stigma. However, the use of nationality as the unit of analysis disregards intra-cultural, individual-level differences in values and preconceptions [8]. For example, British and Taiwanese nationals differed inconsistently across six dimensions of culture (see Table 1) [2]. This is particularly relevant in aviation as accident investigation teams are often multi-national, and even in domestic investigations it is necessary to work with organisations or professions with different cultural preconceptions [6]. To gain a clearer understanding of the cultural mechanisms influencing accident error classification, there is a need to measure cultural behaviours at the individual level.

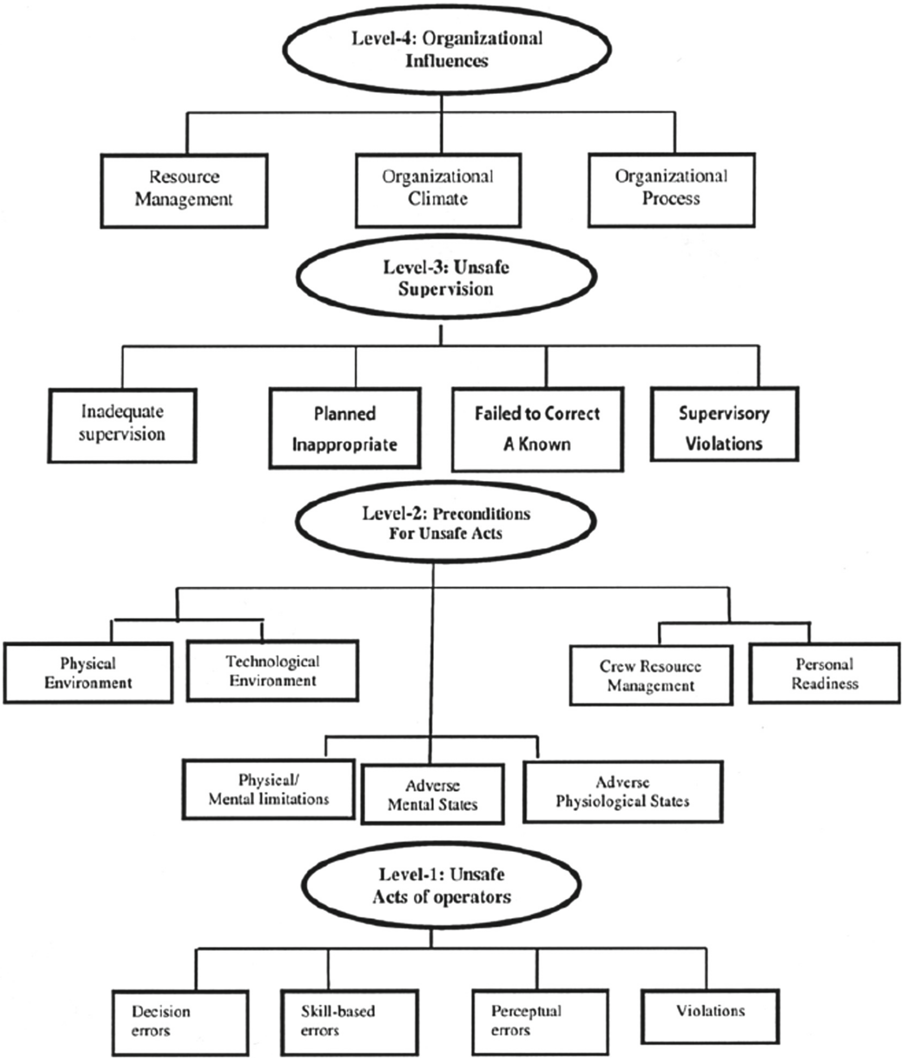


Fig. 1. The HFACS framework [7].

Hofstede's six-dimension model of culture (see Table 1) provides validated measures of individual preferences and behaviours across six dimensions of Power Distance (PD), Individualism-Collectivism (IDV), Masculinity-Femininity (MAS), Uncertainty Avoidance (UA), Long- vs. Short-Term Orientation (LTO), and Indulgence vs. Restraint (IVR) [2]. These dimensions provide validated, scalar measures of specific, individual-level cultural traits, providing in greater detail intra-cultural differences which may be lost in conventional, nationality guided categorisations of culture. Therefore, the goal of this study was to assess whether accident investigators' own cultural dimensions had any influence on their attribution of causal factors across the 18 HFACS categories.

Table 1. Characteristics defined by Hofstede's cultural dimensions [9] and examples of country differences [2].

Dimension	High	Low	Taiwan	United Kingdom
Power Distance (PD)	Acceptance of a hierarchical order between people, where everybody has a place	Desire for equal distribution of power and demands justification of inequalities	58	35
Individualism-Collectivism (IDV)	Preference to take care of one's self and their immediate family only	Unquestioned loyalty towards strong, cohesive in-groups	17	89
Masculinity-Femininity (MAS)	Achievement, assertiveness, and competition	Cooperation, modesty, and caring for others	45	66
Uncertainty Avoidance (UA)	Uncomfortable with uncertainty and ambiguity, desires strong codes of behaviour and intolerant of unorthodox ideas	More relaxed attitudes. Practice is more important than principles	69	35
Long- vs. Short-Term Orientation (LTO)	Fosters adaptation, perseverance, and thrift. Oriented towards future rewards	Look towards the past, preferring to maintain traditions whilst being suspicious of societal changes	93	51
Indulgence vs. Restraint (IVR)	Relatively free gratification of desires	Human drives suppressed by strict social norms and regulations	49	69

2 Method

2.1 Participants

This analysis utilised the responses from 30 accident investigators comprising of aviation industry professionals (pilots, airline safety managers) as well as collegiate aviation students and faculty. Participation was voluntary, no identifying information was collected, and participants had the right to terminate the survey at any time. Ethics approval was provided by the Cranfield University Research Ethics system (CURES/12290/2020).

2.2 Research Design

Data collection was by a survey instrument, hosted digitally on the Qualtrics system. Participants were recruited by referral sampling, with an anonymous hyperlink to the survey sent via email. The hyperlink first directed participants to a participant information and consent page. If the participants provided their consent, they were then redirected to the main survey to begin the coding process.

The survey included 24 five-point Likert-scale questions drawn from the Values Survey Module (VSM: Hofstede and Minkov, 2013) which were used to compute Hofstede's six cultural dimensions for each respondent (see Table 1). These questions were followed by the HFACS coding process, where contributing factors underlying the mid-air collision between a Tupolev Tu-154 and a Boeing 757 aircraft over Uberlingen on July 1, 2002, adapted from the official accident report (Bundesstelle fur Flugunfalluntersuchung [10]), were listed alongside checkboxes for each of the 18 HFACS categories. Participants were instructed to tick the boxes corresponding to the HFACS category or categories for which they wish to attribute the contributing factor.

3 Results and Discussions

3.1 Statistical Analysis

Data collected from September to December 2020 was processed and analysed using SPSS. The VSM items were compiled into scores for the six cultural dimensions. A median split for each dimension was carried out to separate responses into high and low groups for the PD, IDV, MAS, UA, LTO, and IVR dimensions. The presence (coded 1) or absence (coded 0) of each HFACS category were assessed, with each category counted a maximum of only once to avoid over-representation from any single causal factor. Chi-square analyses were then performed to determine the strength of association between the six cultural dimensions and each of the 18 HFACS categories.

3.2 Errors at the Operator Level

The results indicate that the PD and LTO dimensions were influential in attributing errors to the Unsafe Acts of Operators level (HFACS level 1; see Fig. 2). High-PD respondents were significantly less likely than low-PD respondents to indicate the Decision Error category as a factor ($\chi 2(1, N = 30) = 7.23, p < .01$), and longer-term orientation (high-LTO) was linked to decreased utilisation of the Perceptual Error category ($\chi 2(1, N = 30) = 4.82, p < .05$).

These findings are complementary to previous research comparing HFACS attribution between accident investigators of different nationalities, and supports the present use of cultural dimensions, rather than nationality, to categorise different respondent groups. In a comparison between British and Taiwanese investigators on the use of HFACS, Li et al. [7] found that operator-level error categories (HFACS level 1) were more frequently utilised by Taiwanese investigators, whilst British investigators were more likely to be critical of higher-level management (HFACS level 3 and 4). Aside from nationality, what were the variations in cultural values between the British and the Taiwanese that caused

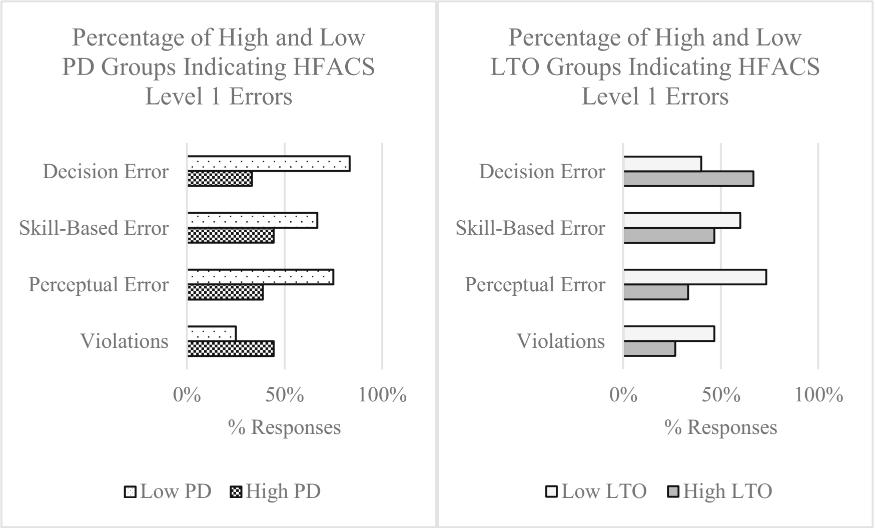


Fig. 2. Percentage of respondents indicating causal factors at HFACS level 1, against PD and LTO dimensions.

this interpretive difference? According to Hofstede et al. [2], British culture is low-PD and low-LTO, whereas Taiwanese culture is high-PD and high-LTO (see Table 1). When this information is integrated with the findings of the present research, which found high-PD linked to fewer selections of HFACS level 1, and high-LTO linked to more selections of HFACS level 1, it is possible to infer that Taiwanese investigators' greater use of HFACS level 1 was influenced by their LTO values.

3.3 Supervisory and Organisational Errors Determined by Long-Term Orientation

On the third and fourth HFACS levels of Unsafe Supervision and Organisational Influences, the classification of causal factors were shaped by investigators' UA and LTO dimensions. There were significant differences between the high and low UA groups in attributing errors to the Failure to Correct a Known Problem category, with high UA investigators more likely to attribute errors to this category ($\chi 2(1, N = 30) = 7.03, p < .01$; see Fig. 3). LTO was found to influence the level of attribution to the Organisational Process category, with longer-term oriented (high-LTO) respondents selecting this category to a significantly greater extent than their low-LTO colleagues ($\chi 2(1, N = 30) = 7.50, p < .01$; see Fig. 3).

To make sense of these findings, it is necessary to delve into the values that each cultural dimension and HFACS category represents. High-UA individuals have an inclination for the work process to be dictated by rules and regulations [2], and the HFACS category of Failure to Correct a Known Problem refers to instances where supervisory failures may be considered to implicitly foster an atmosphere which encourages rule-violation [1]. The greater number of attributions of supervisory failures by high-UA

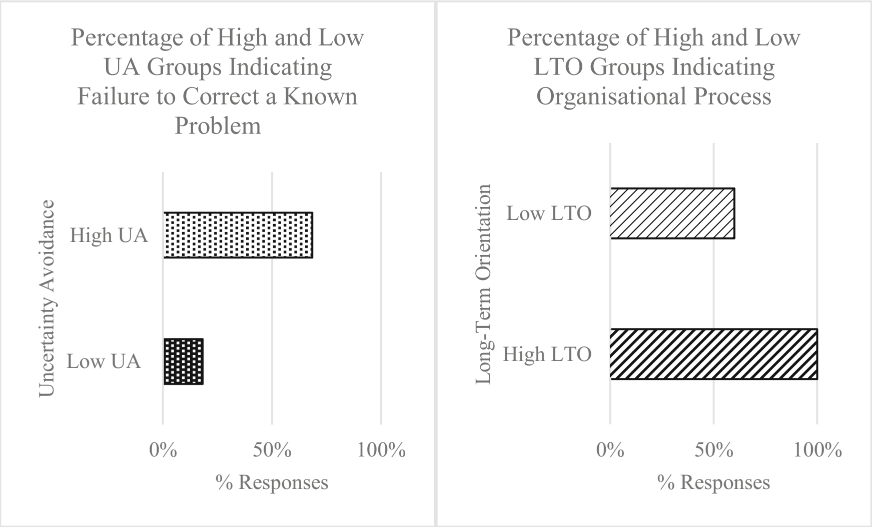


Fig. 3. Percentage of participants of high and low UA groups attributing factors to the *Failure to Correct a Known Problem* category, and percentage of high and low LTO groups attributing factors to *Organisational Process* issues.

respondents may therefore reflect investigators' greater volition for following rules and order. The finding of high-LTO respondents' greater frequency of indicating Organisational Process as a causal factor may be reflective of longer-term oriented (high-LTO) individuals' workplace values. According to Hofstede et al. [2], high-LTO values at work are reflected by beliefs such as that disagreement does not hurt, and that managers and workers share the same aspirations. Thus, higher-LTO accident investigators may be more likely to attribute errors to categories at the higher organisational levels (HFACS levels 3 & 4).

However, the finding of high-UA and high-LTO investigators' greater tendency to attribute errors to the supervisory and organisational levels were unexpected when considered in conjunction with the other cultural dimensions. This is because high-UA, high-LTO cultures are typically also high-PD (e.g. Taiwan, see Table 1), and in high-PD cultures individuals are typically reluctant to criticise higher-level authorities (HFACS levels 3 & 4). A review of previous literature offers several possible explanations. First, the propensity to attribute causal factors to higher HFACS levels is said to be influenced, amongst other things, by the investigators' personal desire to be told what to do [7]. As high-UA signifies a fear of ambiguous situations and a desire for strict rules, it is possible that in situations of high ambiguity, high-UA investigators may be more inclined to blame management for not providing adequate rules and guidance. Secondly, there exists a body of research questioning the efficacy of PD in aviation professions, with shared training experiences and in-group membership known to override the effects of power distance hierarchies [8, 11]. Thirdly, the LTO dimension is reflective of Confucian authority, where the unequal status, power-distant relationships come with an expectation of reciprocation [12]. Viewed from this angle, the attribution of factors to the higher

HFACS levels may possibly be considered as positive feedback to ensure team success, rather than a criticism of supervisory and organisational management.

3.4 Application of Findings

Data aggregated from HFACS analyses can be utilised for the identification of potential training interventions. The findings of cultural dimensions effects on HFACS categorisation helps to paint a clearer picture of biases in the analysis phase of training design. For example, the finding of high-LTO respondents' lesser tendency to select Perceptual Error but a greater tendency to select Organisational Process can be taken into account when analysing aggregated data on commonly indicated deficits. In this case, when the accident analysis teams or organisational culture are known to be high-LTO, a low tally of Perceptual Error should not be mistaken as an indicator of non-issue, and similarly a high tally of Organisation Process factors should not be considered literally. This can help to counteract the effects of subjective biases in performance evaluation, thereby ensuring that subsequent training is well-targeted and operationally-relevant [4].

The results also makes it possible to infer the needs and desires of trainees with different cultural dimensions. This is a rich source of information for the design and development of training and intervention programmes. Consider the five intervention approaches in the Human Factors Intervention Matrix (HFIX) which includes organisational, human, technology, task, and physical environment interventions [13]. As reflected by their more frequent use of the Failure to Correct a Known Problem category, it can be implied that participants with higher UA desired more supervisory guidance and more active supervisory management. If the training audience is high in UA, then a greater focus on organisational interventions within HFIX may possibly be more effective than designing intervention strategies based on the other four approaches.

4 Conclusion

Objectivity is a primary goal of aircraft accident investigations. However, previous research on how accident investigators of different nationalities interpreted identical events differently were insufficient in highlighting specific cultural variations in values and preferences which create interpretive differences. Using Hofstede's six-dimensions of national culture as the basis for comparison, the results indicate that the attribution of causal factors to acts at the frontline operator level (HFACS level 1) were influenced by accident investigators' own PD and LTO levels, whereas the use of higher levels representing supervisory and organisational errors (HFACS levels 3 & 4) were driven by investigators' UA and LTO. These findings can be applied in training development, with biases in the attribution of error taken into account during the analysis phase of training design. The needs and desires of the target audience can also be inferred from the results and utilised for the design and development of instructional systems.

References

1. Wiegmann, D.A., Shappell, S.A.: A Human Error Approach to Aviation Accident Analysis, A Human Error Approach to Aviation Accident Analysis (2016). https://doi.org/10.4324/9781315263878

2. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software of the Mind, Intercultural Cooperation and Its Importance for Survival, Cultures and Organizations (2010). https://doi.org/10.1007/s11569-007-0005-8
3. Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.A.: Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Hum. Factors (2007). https://doi.org/10.1518/001872007X312469
4. Li, W.C., Harris, D.: Identifying training deficiencies in military pilots by applying the human factors analysis and classification system. Int. J. Occup. Saf. Ergon. **19**, 3–18 (2013). https://doi.org/10.1080/10803548.2013.11076962
5. Shappell, S., Wiegmann, D.: The human factors analysis and classification system – HFACS, USDOT/FAA/AM-00/7 Office of Aviation Medicine (2000)
6. Williams, P.R.: Some cultural considerations for investigators. In: Australian and New Zealand Societies of Air Safety Investigators, Regional Air Safety Seminar. Canberra, ACT: Australian and New Zealand Societies of Air Safety Investigators (2010)
7. Li, W.-C., Young, H.-T., Wang, T., Harris, D.: International cooperation and challenges: understanding cross-cultural issues. ISASI Forum - Air Saf. Through Invest. **40**(4), 16–21 (2007)
8. Chan, W.T.-K., Harris, D.: Third-culture kid pilots and multi-cultural identity effects on pilots' attitudes. Aerosp. Med. Hum. Perform. **90**(12), 1026–1033 (2019). https://doi.org/10.3357/AMHP.5397.2019
9. Hofstede, G., Minkov, M.: Values survey module 2013 manual (2013)
10. Bundesstelle fur Flugunfalluntersuchung: Investigation Report AX001-1-2/02 (Accident 1 July 2002 near Ueberlingen/Lake of Constance/Germany) (2004)
11. Helmreich, R.L., Merritt, A.C.: Culture at Work in Aviation and Medicine. Routledge, New York (1998)
12. Jing, H.S., Batteau, A.: The dragon in the cockpit: a Chinese view about the differences between China and the west in aviation. In: International Air Safety Seminar Proceedings (2015)
13. Shappell, S., Wiegmann, D.: A methodology for assessing safety programs targeting human error in aviation. Int. J. Aviat. Psychol. **19**, 252–269 (2009). https://doi.org/10.1080/10508410902983904

Floating Iceberg Model of Psychological Competence Towards Airline Transport Pilots' Professionalism Lifecycle Management System

Shu Li[1](✉), Lei Wang[1], and Ming Zeng[2]

[1] Flight Academy, Civil Aviation University of China, Tianjin, China
[2] Department of Flight Standard, Civil Aviation Administration of China, Beijing, China

Abstract. Based on the concept of Pilots' Professionalism Lifecycle Management System put forward by Civil Aviation Administration of China, this paper aims to develop a Floating Iceberg Model of Pilots' Psychological Competence, which shows that professional adaptability of psychological competence consist of professional competence and social competence and formed on the basis of personality traits and cognitive abilities, meanwhile, mental health status fluctuates with the changes of professional adaptability and surrounding environment in a certain range above and below the water. There are flow channels among the modules in the iceberg, so it is necessary to explore the key flow factors within cognitive abilities and personality traits that contributes to pilots' professional adaptability, as well as the ways of their influence on professional competence and social competence, and to clarify that whether gates can be set to control and guide the flow of key flow factors. This paper lists the dimensions and indicators, assessment methods and stages, the key psychological indicators closely related to core competency, work style competency, and abilities of coping with "black swan" and "grey rhino" during the whole pilots' professionalism lifecycle. It provides some suggestions for the future research direction and the challenge of pilots' psychological competence construction and psychological competency assessment.

Keywords: Psychological competence model · Psychological competency assessment · Professionalism lifecycle management system · Professional adaptability of psychological competence

1 Introduction

It is impossible to foresee all plausible accident scenarios, especially in today's aviation system where the system's complexity and high reliability mean that the next accident may be something completely unexpected [1]. Evidence-Based Training (EBT) addresses this by moving from pure scenario-based training to prioritizing the development and assessment of defined competencies, leading to better training outcomes. Eight pilot competencies [2] (application of procedures, communication, aircraft flight path management-automation, aircraft flight path management- manual control, leadership

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 28–37, 2021.
https://doi.org/10.1007/978-3-030-77932-0_3

and teamwork, problem solving and decision making, situation awareness and management of information, workload management) proposed by International Civil Aviation Organization (ICAO) and another competency [3] (application of knowledge) proposed by European Aviation Safety Agency (EASA) are being used for competency-based training and assessment (CBTA).

Civil Aviation Administration of China (CAAC) issued the "Guidance on Comprehensively Deepening Flight Training Reform of Transport Airlines" in June 2019, and issued the "Roadmap for the Construction and Implementation of the Airline Transport Pilots' Professionalism Lifecycle Management System in China" [4] in December 2020. Professionalism Lifecycle Management (PLM) System [4] is a qualification management system covering theory, personnel, equipment, procedures and support system, which is proposed by CAAC. PLM takes career competency as the center, evidence-based training as the driving force, work style construction as the traction, assessment evidence of core competency and professional adaptability (of psychological competence) as the basics, covering the whole elements and whole cycle of pilot training as features, and continuously improving the "grey rhino" and "black swan" risk management and control ability of pilot.

PLM integrates three evaluation dimensions which are core competency, psychological competency and work style competency into career competency, and establishes an evaluation model with multi-dimensional, full cycle and big data integration based on the same underlying methodology of index system construction. In PLM, EBT is the foundation of training methodology, and EBT integration will be steadily carried forward in the five stages of Pilot Professionalism Lifecycle from ab initial, F/O (first officer), PIC (pilot in command), TRI (type rating instructor) to DPE (designed pilot examiner). CBTA constructs nine core competency and behavior indicators of pilots, which can be obtained from training operational data. To make sure that the individual psychological competences of pilots meet the requirements of psychological competency, the mental health status and professional adaptability of pilots can be obtained from the psychological competency assessment, which is an effective tool to control the increment of entering the industry and screen pilots already in training. Work style competency puts forward relevant requirements from pilots' personal behavior standards, personal guarantee of safety and self-improvement, which can be obtained from the work style competency evaluation. Through training, pilots' three competency to cope with the two risks are continuously strengthened.

2 Research Progress on Psychological Competency

The psychological competency proposed by CAAC refers to the conformity of mental health and professional adaptability with the requirements of career competency [4]. It is formed on the basis of heredity, under the influence of education and environment, through the practical training.

2.1 Mental Health

Studies have determined that almost 80% of aviation accidents are caused by human errors. Human factor has been widely recognized as a key factor to ensure aviation safety and efficiency [5]. As the key safety personnel, pilots' mental health has become one of the key factors to ensure aviation safety.

The World Health Organization (WHO) [6, 7] defines health as a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. Mental health is further defined as a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community.

Aviation is a high risk, high reliability, safety critical industry. Pilots' poor health may pose a risk to the safety of the system and the public. The safety management system (SMS), as provided for by ICAO Annex 19, is intended to support the identification of hazards and the mitigation of risks associated with the potential outcomes of those hazards. The effectiveness of an SMS depends on a positive safety culture and an environment, which supports the health and performance of all persons working within the aviation industry [8]. Psychological problems amongst aircrew present a threat to flight safety, given the ensuing impairments to task performance. Three years after the Germanwings crash in which a pilot deliberately flew a jet into a mountainside, the regulations made by EASA seek to prevent a similar tragedy by providing all pilots with access to a support program in case of mental health problems and by making European airlines perform a psychological assessment of pilots before they start work [9].

2.2 Professional Adaptability of Psychological Competence

Professional adaptability (of psychological competence) refers to a person who is engaged in the professional pilot work must have certain psychological competence [4]. Through psychological assessment and selection, the talents suitable for commercial transportation flight are selected for training, and then the qualified pilots are trained, which can lay the foundation for flight safety and effectively save the training cost. Psychological selection, including basic cognitive ability, psychomotor ability and personality evaluation etc., has become a necessary link to recruit ab-initio pilot all over the world [10].

EASA puts forward that psychological assessment should include cognitive abilities, personality traits, operational and professional competencies, and social competences in accordance with crew resource management principles [9]. In the pilot aptitude test (PAT) guidance material [11] developed by IATA, six dimensions (English language proficiency, basic mental ability, composite mental abilities, operational abilities, social-interpersonal abilities, personality traits) should be measured covering pilots' professionalism lifecycle which are all directly related to the eight pilot competencies [2] developed by ICAO.

To describe the transfer of ab-initio pilot selection methods from Germany to China, the different phases of test development and evaluation were carried out by DLR in cooperation with Civil Aviation Flying College of China, which found that the reliabilities for the personality questionnaire as well as for the tests of memory, mental concentration, and spatial abilities seem to be equivalent with the German versions [12]. The number of correct responses had significant positive correlations for both ab initio and licensed pilots with measures of quantitative and spatial abilities, visual memory, multi-tasking, and psychomotor performance, but not with personality, at the same time the number of incorrect responses was correlated with emotional instability and openness for licensed pilots but not with measures of cognitive ability [13]. There was a strong relationship between visuospatial ability, field-independence and spatial orientation in flight [14]. Trait mindfulness had a direct and negative effect on airline pilots' incident involvement, but an indirect effect on incident involvement through influencing risk perception, and flight experience has the negative and direct effect of trait mindfulness on incident involvement [15]. The best predictors of success in training were instrument comprehension, mechanical principles, and aviation information based on the meta-analysis [16]. Emotional quotient has a significant impact on theoretical performance and flight performance and shows moderate level of correlation with crisis response ability for the psychological selection of student pilots [17]. The new scales should be used in those instances in which accurate measurement of the attitudes is needed, specifically to assess the contribution of pilots' attitudes to accident involvement and to assess of the impact of training [18]. Hazardous attitude is a psychological risk element that has significant negative effects on severe exceedance behaviors of pilots and the flight operation performance levels of pilots were also affected by the psychological risk level [19].

2.3 Relationship Between Mental Health and Professional Adaptability

Neuroticism within the Five Factor Model of personality is mostly strongly associated with poor stress regulation [20, 21], that is linked to mental health [22]. Personality studies has demonstrated some particular variables have strong effects on mental health, such as self-esteem [23] and self-monitoring [24]. During recent decades, special interest has developed in the positive rather than the negative aspects of mental health [25], such as resilience [26, 27], hardiness [27–29] and mental toughness [30] are factors which act as protective ones and improve mental health. Better perception, understanding, and management of emotion of individuals with higher emotional intelligence make it less likely that they will experience mental health problems and emotional intelligence has useful additional predictive information over and above the Big Five Dimensions for mental health functioning [31, 32]. Emotional intelligence and proactive coping had protective potential in the prevention of depression and anxiety [33]. Mindfulness primarily influences pilots' anxiety indirectly by affecting burnout [34].

It is time to pay more attention to healthy aspects of personality and mental process or in other words good mental health in order to find what kind of features are prominent in healthy individuals rather than what kind of features should not be seen in individuals and also in order to include wellness and mental health promoting factors [35].

3 Floating Iceberg Model of Pilots' Psychological Competence

Based on previous studies, this paper proposes a floating iceberg model of pilots' psychological competence, as shown in Fig. 1.

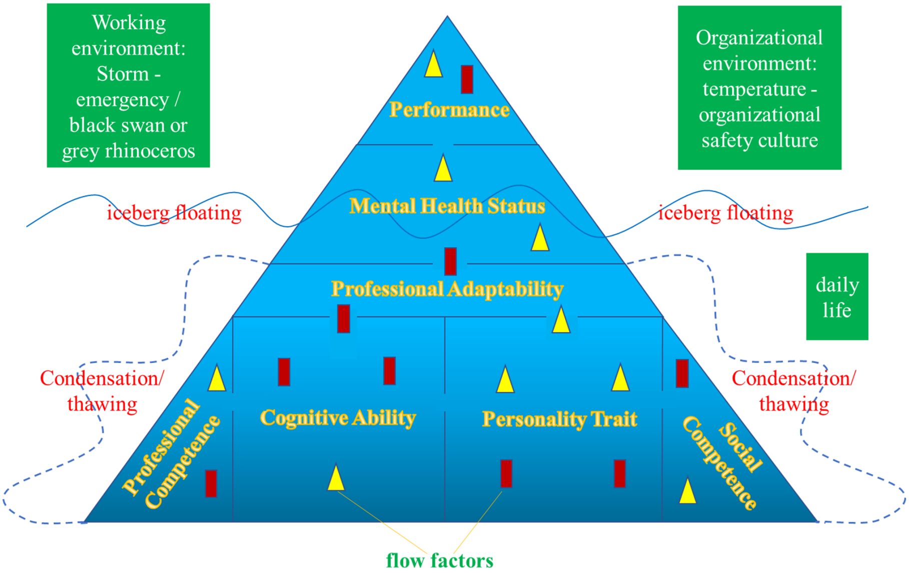


Fig. 1. Floating iceberg model of pilots' psychological competence

3.1 Structure of the Model

The structure of the model is as follows:

- Bottom of the iceberg - professional adaptability
- Professional adaptability is a kind of professional competences and social competences formed on the basis of personality traits and cognitive abilities. These are relatively stable after formation, but may become stronger (iceberg condensation) or worse (iceberg thawing) due to the influence of external environment.
- Middle of the iceberg - mental health status
- The mental health status fluctuates with the changes of professional adaptability and surrounding environment within a certain range above and below the water. We need to explore the relationship between mental health status and the elements of each module of the professional adaptability as well as the surrounding environment.
- Tops of the iceberg - performance
- Pilots' professional adaptability and mental health, such as sub-health and high stress, directly affect their performance.
- Around the iceberg - environment
- Environment includes organizational environment (temperature - organizational safety culture), working environment (Storm-emergency/black swan or grey rhino) and daily life environment can make the iceberg floating, condensing or thawing.

3.2 Mechanism of the Model

There are flow channels among the modules in the iceberg, so it is necessary to explore the key flow factors within cognitive abilities and personality traits that contributes to pilots' professional adaptability, as well as the ways of their influence on professional competences and social competences, and to clarify that whether gates can be set to control and guide the flow of key flow factors. For example, pilots with certain personality traits have stronger communicative competence, and pilots with good information processing ability have better situational awareness.

The personality traits and cognitive abilities of an adult have been formed before the recruitment to be a pilot, and they can be screened out more suitable for the flying profession through the evaluation. On this basis, their professional competences and social competences have certain plasticity, like the condensation and thawing of iceberg, which are necessary to be strengthened during the training and be selected when upgrading.

As for the floating iceberg model, one's professional adaptability psychological energy (iceberg volume) needs to reach a certain value. The higher the value is, the better the mental health state is, the better the performance is, and the more resistant to the disturbance and erosion of external environment. Meanwhile, the better the external environment is, the higher the professional adaptability psychological energy is (iceberg condensation is strong), and the more stable the mental health state is (iceberg floating is weak).

4 Professional Adaptability Assessment of PLM

Psychological competency assessment is to assess one's psychological and mental quality through a series of scientific assessment tools, so as to make the match between the individual and profession more scientific and reasonable, reduce accidents and improve work efficiency. Some inherent risks of professional adaptability defects cannot be controlled or mitigated by flight training, or the training cost is unacceptable. Psychological assessment is the first defense line to be professional pilots, and it is also a tool for us to screen the talents.

Based on psychological competency index system, psychological competency assessment is to assess what specific abilities and skills are required for a person to evidence competence within each competency phase. The preliminary index system of professional adaptability includes basic psychological abilities, comprehensive psychological abilities, operational competences, social-interpersonal competences and personality traits. Table 1 lists the dimensions and indicators, assessment methods and stages, the key psychological indicators closely related to core competency, work style competency, and abilities of coping with "black swan" and "grey rhino".

General information includes age, education background, interests, family status, effective social relations, major life changes, social pressure, etc.

Flight experience includes flight hours, license, type ratings, examination results.

Health status includes medical certificate, mental health.

Spatial abilities (dynamic) include flight path management with manual control and automatic management

Professionalism includes responsibility, self-discipline, motivation, etc.

Individual stability includes stress resistance, resilience, impulse control, etc.

Table 1. Professional adaptability assessment matrix of PLM

Professional adaptability of psychological competence		Assessment method	Assessment stage	Associated core competency	Associated work style competency	Abilities of coping with black swan & grey rhino
Dimension	Indicator					
Personal information	General information	Questionnaire, Interview, Docs	Professionalism Lifecycle		×	
	Flight experience				×	×
	Health status					
Basic psychological abilities	Information processing	PC-based psychometric tests	AB-initial			
	Spatial abilities (static)					
	Memory capacity					
	Technical comprehension					
	Reasoning					
	Logical ability					
	Long-term concentration					
	Eye-hand-foot coordination					
Comprehensive psychological abilities	Allocation of attention	Psychometric apparatus tests, PC-based psychometric tests	To be a F/O	×		
	Multi-tasking			×		
	Psychomotor abilities			×		
	Spatial abilities (dynamic)			×		
Operational competences	Problem solving and decision making	PC-based psychometric tests, FSTD-based tests, work samples psychometric tests	To be a F/O, PIC,TRI, DPE	××		××
	Workload management			××		×
	Situational awareness and management of information			××		×
Social and interpersonal competences	Communication	Standardized interviews, Group scenarios	To be a PIC, TRI, DPE	××		
	Leadership			××	×	
	Teamwork			××	×	
Personality traits	Professionalism	Standardized interviews, PC-based psychometric tests, FSTD-based tests	Professionalism Lifecycle		××	
	Individual stability				××	××
	Risk propensity				××	××
	Application of procedures and compliance with regulations			××	××	

Risk propensity includes hazardous attitudes, emotional stability, sensation seeking, etc.

The current assessment dimensions and indicators mainly refer to IATA's guidance material [11]. Scientific psychological assessment tools can play the role of screening and controlling increment at critical time points, such as to be a student pilot, entry transport airlines and key technology level promotion, and can also continuously identify the psychological characteristics and professional adaptability defects of existing pilots. The pilot psychological competency index system needs to be combined with the characteristics of the pilot's core competency and work style competency, and rely on the operation data for long-term tracking to continuously improve the assessment model. In the future, combined with the development of assessment method and collection of assessment data, the assessment index system of psychological competency will be revised and improved according to the reality and safety culture of civil aviation in China.

5 Problems and Prospects

There are still some problems and prospects for the Floating Iceberg Model of Pilots' Psychological Competence:

- The cognitive abilities and personality traits closely related to flight operation performance should be found to control the increment of entering the industry by collecting and tracking individual psychological data and flight data of pilots.
- The key flow factors of cognitive abilities and personality traits that are closely related to the professional competences and social competences as well as the flow channels between them should be explored to screen pilots already in training. At the same time, flow factors with plasticity should be improved through training.
- The flow factors within each dimension of professional adaptability that can affect the mental health status as well as the flow channels should be identified, so as to recognize the pilots who need to be focused. It is necessary to identify the key channels that working environment (emergency incidents) and organizational environment (safety culture) affect mental health, in order to carry out psychological intervention after emergency and the construction of safety culture.
- Mental health is a fluctuating mental state, which needs to be continuously monitored. Nowadays, many organizations have carried out Employee Assistance Program (EAP) and peer support for pilots, and we can also consider establishing a confidential mental health status reporting system in the future, so as to identify individual and group psychological distress as early as possible, and provide assistance in time.
- On the basis of initial assessment index system for the professional adaptability in PLM, we need to develop assessment instrument, collect and analyze data, gradually revise the index system and assessment methods, and finally build an assessment model which can be integrated into flight training.
- The construction of psychological competency for PLM needs the joint efforts of the whole civil aviation industry, including civil aviation authorities, operators, approved training organizations, research institutes and pilot representative bodies.

Acknowledgments. We appreciate the support to this work received from the National Natural Science Foundation of China (grant no. 32071063).

References

1. International Air Transport Association (IATA). Evidence-Based Training Implementation Guide. 1st edn. (2013). https://www.iata.org/contentassets/632cceb91d1f41d18cec52e375f38e73/ebt-implementation-guide.pdf
2. International Civil Aviation Organization (ICAO). Doc 9995, Manual of Evidence-based Training, First Edition (2013)
3. European Aviation Safety Agency (EASA). Implementation of evidence-based training (EBT) within the European regulatory framework, Explanatory Note to Decision 2015/027/R (2015). https://www.easa.europa.eu/sites/default/files/dfu/ED%20Decision%202015-027-R%20-%20Explanatory%20Note%20.pdf
4. Civil Aviation Administration of China (CAAC). Roadmap for the Construction and Implementation of the Airline Transport Pilots' Professionalism Lifecycle Management System in China (in Chinese) (2020). https://pilot.caac.gov.cn/includes/pdfjs-2.0.943/web/viewer.html?file=/servlet/FileUploadManagerServlet?method_40de8280-a6d9-498a-a9cd-4e9e71b24fd0
5. International Civil Aviation Organization (ICAO). Human Factors Training Manual, Doc 9683-AN/950 (1998)
6. World Health Organization (WHO). Mental health: strengthening our response, Fact Sheet, 30 March 2018. https://www.who.int/en/news-room/fact-sheets/detail/mental-health-strengthening-our-response
7. World Health Organization (WHO). The World health report: Mental Health: New Understanding, New Hope (2001). https://www.who.int/whr/2001/en/whr01_en.pdf?ua=1
8. International Civil Aviation Organization (ICAO). Electronic Bulletin, EB 2020/55. Promoting, maintaining and supporting mental well-being in aviation during the Covid-19 pandemic (2020). https://www.icao.int/Meetings/TRIP-Symposium-2018/Documents/EB_20_55_Mental%20Health.pdf
9. European Aviation Safety Agency (EASA). AMC1 CAT.GEN.MPA.175(b) Endangering Safety. Annex III to ED Decision 2018/012/R. https://www.easa.europa.eu/sites/default/files/dfu/Annex%20III%20to%20EDD%202018-012-R.pdf
10. Bor, R., Eriksen, C., Hubbard, T., King, R.E. (eds.): Pilot Selection: Psychological Principles and Practice. CRC Press, London (2019)
11. International Air Transport Association (IATA). Pilot Aptitude Testing, Guidance Material and Best Practices, 3rd edn. (2019). https://www.iata.org/contentassets/19f9168ecf584fc7b4af8d6d1e35c769/pilot-aptitude-testing-guide.pdf
12. Hoermann, H.J., Luo, X.: Empirical evaluation of a selection system for Chinese student pilots. Space Med. Med. Eng. **15**(1), 6–11 (2002). https://www.researchgate.net/publication/11401702_Empirical_evaluation_of_a_selection_system_for_Chinese_student_pilots
13. Hoermann, H.J., Damos, D.L.: The use of a perceptual speed test in civilian pilot selection. In: 20th International Symposium on Aviation Psychology, Dayton/OH, 7–10 May 2019, p. 391 (2019). https://www.researchgate.net/publication/332801230_The_use_of_a_perceptual_speed_test_in_civilian_pilot_selection
14. You, X., Yu, L.: The relationship between cognitive profile, field-independence and spatial orientation in flight (in Chinese). Acta Psychol. Sinica **32**(2), 158–225 (2000). https://doi.org/10.1007/s11769-000-0010-0
15. Ming, J., Can, Y., Haiyan, H., Ying, L., Quan, X.: The influence of trait mindfulness on incident involvement among Chinese airline pilots: the role of risk perception and flight experience. J. Saf. Res. **66**(1), 161–168 (2018). https://doi.org/10.1016/j.jsr.2018.07.005
16. Martinussen, M., Torjussen, T.: Pilot selection in the Norwegian air force: a validation and meta-analysis of the test battery. Int. J. Aviat. Psychol. **8**(1), 33–45 (1998). https://doi.org/10.1207/s15327108ijap0801_2

17. Li, S.: Study on the influence of psychological quality on performance for civil aviation pilot students (in Chinese). J. Transp. Inf. Saf. **4**, 37–43 (2016). https://doi.org/10.3963/j.issn1674-4861.2016.04.006
18. Hunter, D.R.: Measurement of hazardous attitudes among pilots. Int. J. Aviat. Psychol. **15**(1), 23–43 (2005). https://doi.org/10.1207/s15327108ijap1501_2
19. Wang, L., Zhang, J.: The effect of psychological risk elements on pilot flight operational performance. Hum. Factors Ergon. Manuf. Serv. Ind. 1–11 (2019). https://doi.org/10.1002/hfm.20816
20. Williams, P.G., Moroz, T.L.: Personality vulnerability to stress-related sleep disruption: pathways to adverse mental and physical health outcomes. Pers. Individ. Differ. **46**(5–6), 598–603 (2009). https://doi.org/10.1016/j.paid.2008.12.017
21. Lazarus, R.S., Folkman, S.: Stress, Appraisal, and Coping. Springer, New York (1984)
22. Stead, R., Shanahan, M.J., Neufeld, R.W.J.: "I'll go to therapy, eventually": procrastination, stress and mental health. Pers. Individ. Differ. **49**(3), 175–180 (2010). https://doi.org/10.1016/j.paid.2010.03.028
23. Rosenthal, S.A., Hooley, J.M.: Narcissism assessment in social–personality research: does the association between narcissism and psychological health result from a confound with self-esteem? J. Res. Pers. **44**(4), 453–465 (2010). https://doi.org/10.1016/j.jrp.2010.05.008
24. Ajzen, I., Timko, C., White, J.B.: Self-monitoring and the attitude–behavior relation. J. Pers. Soc. Psychol. **42**(3), 426–435 (1982). https://doi.org/10.1037/0022-3514.42.3.426
25. Seligman, M.E.P., Steen, T.A., Park, N., Peterson, C.: Positive psychology progress: empirical validation of interventions. Am. Psychol. **60**(5), 410–421 (2005). https://doi.org/10.1037/0003-066X.60.5.410
26. Davydov, D.M., Stewart, R., Ritchie, K., et al.: Resilience and mental health. Clin. Psychol. Rev. **30**(5), 479–495 (2010). https://doi.org/10.1016/j.cpr.2010.03.003
27. Nezhad, M.A.S., Besharat, M.A.: Relations of resilience and hardiness with sport achievement and mental health in a sample of athletes. Proc.-Soc. Behav. Sci. **5**(1), 757–763 (2010). https://doi.org/10.1016/j.sbspro.2010.07.180
28. Johnsen, B.H., Eid, J., Pallesen, S., et al.: Predicting transformational leadership in naval cadets: effects of personality hardiness and training. J. Appl. Soc. Psychol. **39**(9), 2213–2235 (2009). https://doi.org/10.1111/j.1559-1816.2009.00522.x
29. Maddi, S.R., Khoshaba, D.M.: Hardiness and mental health. J. Pers. Assess. **63**(2), 265 (1994). https://doi.org/10.1207/s15327752jpa6302_6
30. Horsburgh, V.A., Schermer, J.A., Veselka, L., et al.: A behavioural genetic study of mental toughness and personality. Pers. Individ. Differ. **46**(2), 100–105 (2009). https://doi.org/10.1016/j.paid.2008.09.009
31. Schutte, N.S., Malouff, J.F., Thorsteinsson, E.B.: A meta-analytic investigation of the relationship between emotional intelligence and health. Pers. Individ. Differ. **42**, 921–933 (2007). https://doi.org/10.1016/j.paid.2006.09.003
32. Ciarrochi, J., Deane, F.P., Anderson, S.: Emotional intelligence moderates the relationship between stress and mental health. Pers. Individ. Differ. **32**(2), 197–209 (2002). https://doi.org/10.1016/S0191-8869(01)00012-5
33. Guo, Y., Ji, M., You, X., Huang, J.: Protective effects of emotional intelligence and proactive coping on civil pilots' mental health. Aerosp. Med. Hum. Perform. **88**(9), 858–865 (2017). https://doi.org/10.3357/AMHP.4799.2017
34. Li, Y., Chen, H., Xin, X., et al.: The influence of mindfulness on mental state with regard to safety among civil pilots. J. Air Transp. Manage. **84** (2020). https://doi.org/10.1016/j.jairtraman.2020.101768
35. Salehinezhad, M.A.: Personality and mental health. Essent. Notes Psychiatry (2012). https://doi.org/10.13140/RG.2.1.1058.0003

Evaluation of Fatigue Induced During the Flight Tasks

Bo Shao, Zhen Wang(✉), and Shan Fu

Department of Automation, Shanghai Jiao Tong University, Shanghai, China
b2wz@sjtu.edu.cn

Abstract. The fatigue of the pilots in the flight tasks affects his cognition of flight data and the ability to operate the aircraft correctly. Due to the increasing number of daily flights, the probability of fatigue flying becomes higher, which is harmful to flight safety. Therefore, the research on how to detect pilot's fatigue during flight tasks is of great significance to ensure the safety of aviation. A simulation study with 11 groups of pilots was conducted to examine the correspondence between pilot's fatigue and their physiology data when they performed normal pilot flight training tasks. Result show that as fatigue accumulates, the participants' pupil diameter, heart rate, and duration of fixation on the area of interest are significantly different from the beginning of the experiment, we also proposed a comprehensive indicator to reflect the pilots' fatigue through polynomial analysis of these changing physiology parameter.

Keywords: Fatigue · Flight tasks · Physiology data

1 Introduction

According to the report from the Aviation Safety Network, a total of 20 airliners accidents occurred in 2019, resulting in 283 fatalities. Although the number of fatalities has decreased compared to 2018, the number of accidents has exceeded the five-year average number of airline accidents. Moreover, due to the particularity of air traffic, airline accidents will cause serious loss of life and property to the country and the individual. Many aircraft accidents are caused by pilot errors which refer to flight operations and maintenance procedures during flight tasks, and the fatigue is an important factor leading to pilot errors. In order to ensure flight safety, it is necessary to study the index of fatigue induced in flight tasks. The index of fatigue study could contribute to evaluate the pilots' physiology state and to help improve flight safety.

The evaluation of participants' fatigue level is hot research filed in ergonomics and human-computer interaction, the effective methods for evaluating fatigue can be divided into subjective methods, such as questionnaires (e.g., The Fatigue Assessment Scale [1]), and interviews (e.g., Meesters [2]), and objective methods that assess the fatigue status of participants by analyzing their physiological data

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 38–51, 2021.
https://doi.org/10.1007/978-3-030-77932-0_4

or behavioral performance. Compared with subjective evaluation, the advantage of objective evaluation is that it is not affected by subjective factors and has real-time performance. Many existing studies have shown that there is a strong correlation between human fatigue and physiological signals, such as electroencephalography (EEG) signals, electrocardiogram (ECG) signals, eye movement data, and heart rate. These physiological data can be used to assess the fatigue of participants during the experiment task.

The existing study of fatigue mainly focus on the Vehicle Driving tasks and Visual Display Terminal Tasks (VDT), they linked fatigue to changes in physiological data. Morales et al. [3] designed a driving experiment to prove that the EEG-metrics were sufficiently sensitive to reveal the different levels of fatigue of drivers. Because cascade is not voluntarily controlled [6], it is a suitable physiological indicator that has the potential to provide an objective marker of fatigue [10]. Previous research has found that in the laboratory environment, as the fatigue of the participants increases, the cascade velocities will change (e.g., Schmidt et al. [11], M. Prsa et al. [12]). Murat et al. [5] proposed an effectively method that uses both pupil diameter and focal accommodation to evaluated fatigue. Yet, experiments in VDT and Driving tasks usually include a few tedious video tasks or sleep deprivation to induce the fatigue of participants, which usually does not happen in a real operating environment, so the correctness of many research results can only be confirmed in the experimental environment. The correlation between fatigue and physiological indicator needs further research.

Research on aviation fatigue is more complicated, because the aircraft simulator is difficult to manufacture, and operating aircraft is a complex task that requires a high focus on flight data, interaction with air traffic control, and adapt to external contingencies [7–9]. The participants of previous studies on physiological data were less professional trained. How the physiological data of pilots, a population that was specially trained, will be influenced by fatigue have not been adequately studied. To address the challenges above, a flight simulator is required to be as realistic as possible, and the participants in the experiment are required to have flight experience to reflect the changes in physiological data of professionally trained people. Moreover, it should be noted that physiological data are affected by a variety of factors, such as task complexity, the workload of the task, and fatigue. The objective of the present research is to study the effect of fatigue on pilots' physiological data in the context of flying. Therefore, in addition to designing a realistic experimental task, it is also a challenge to eliminate the effects of other factors on participants' physiological data.

In this study, we used various equipment to continuously monitored participants' heart rate, respiratory rate, and eye movements, such as fixation duration, cascade, and coordinates of gaze points. We hypothesized that participants' fatigue would gradually increase with the accumulation of time in the experiment and that their physiological data would reflect this phenomenon. The specific details of the experimental and how we eliminate the influence of other factors on physiological data are in the next section. Furthermore, we base on physiological data proposed a multiple regression model that could output a prediction value as the indicator to evaluate the level of fatigue during flight tasks. To conclude, our contributions are as follows

- To determine the effect of fatigue on the physiological data in the context of flying, we analyzed the change of physiological data at various tasks and found that the fatigue had a significant effect on physiological data.
- We proposed a multiple regression model to evaluate the fatigue level during flight tasks and it yields satisfactory prediction.

2 Method

2.1 Participants

Participants were twenty certified male airline pilots, ages 30 to 50, flying hours ranging from 1000 to 18000 h. All volunteers were in good health and had normal visual acuity. The number of aircraft flight hours could affect physiological data and operational behavior. Therefore, according to their flying age, they were divided into ten crews, each of which includes a captain who has flying age of more than 10000 h and a co-pilot who has flying age of fewer than 10000 h (did not differ significantly between each group $p - values > 0.05$).

2.2 Apparatus

Flying Simulator. The experiment was conducted using a Level D full-flight simulator that simulates the cockpit of CRY-200. Before the beginning of the experiment, we adjust it according to the feedback from pilots in order to make the laboratory environment closer to the real aircraft cockpit. The entire experimental environment is shown in Fig. 1.

Eye Movement Tracker. Pilots' gaze behaviors were recorded by Tobii Glasses that is a head-mounted eye tracker with a 30-hertz sample rate developed by Tobii Technology AB, and a scene camera to change the local coordinates of gaze points from Tobii Glasses to the world coordinates.

Heart Rate Monitor. Since the accurate ECG monitor equipment will affect the pilots' normal operation, we did not monitor the ECG signal in this experiment. We use Zephyr Bio harness, a Multi-channel physiological measurement system developed by Zephyr Technology Crop, to monitor the heart rate at a sampling frequency of 1 Hz. The respiratory rate was recorded by a chest strap instrument.

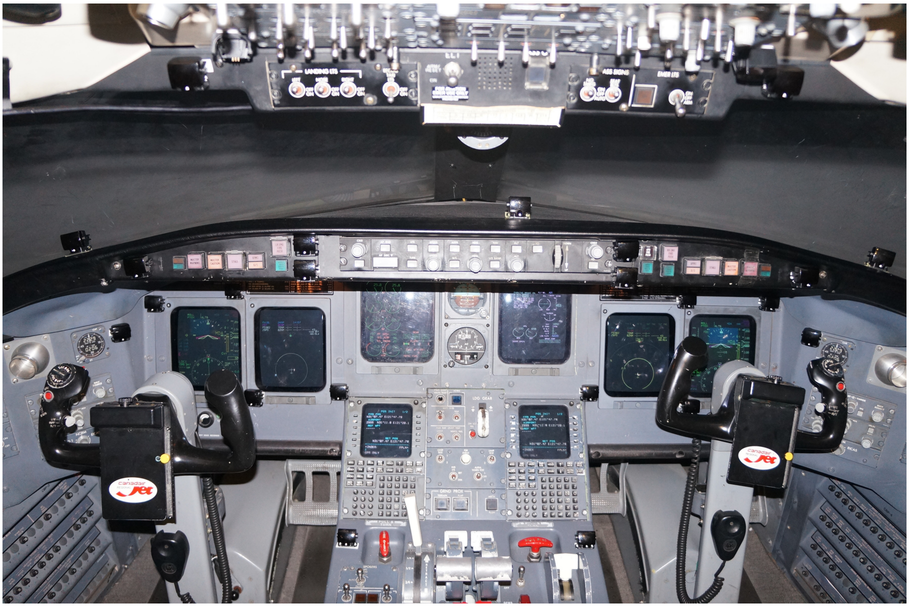


Fig. 1. The flight simulator used in our experiment. It is completely built according to the cockpit of the CRY-200 aircraft

2.3 Experimental Design

To address the question of how to induce fatigue from flight tasks, we considered the conclusion of previous research that a 2-h driving time was a common inducer of fatigue (Wijesuriya et al. [4]). Based on this, we designed an experiment to induce the fatigue of participants, in which each crew member was asked to complete 2 rounds of 2-h routine flight training, after each round of experimentation, the captain and co-pilot will be exchanged. Each experimental session consisted of fourteen flight tasks, all of which are common tasks during flight training, such as manual flight, generator failure, and fuel imbalance. Each task lasts about 10 to 20 min and there is a 5-min interval between the two tasks. During the experiment, the participants were asked to wear equipment to measure their physiological data, and there is a camera to record the operation behavior of the pilots. Eye movement (gaze points, cascade, fixation), pupil diameter, heart rate, and respiratory rate were measured to indicate the fatigue levels of participants.

2.4 Data Extraction

Physiological data are usually influenced by many factors, especially the difficulty of the task, so a direct analysis of changes in participants' physiological data during the experiment does not reflect the effect of fatigue on physiological data. In order to eliminate the impact of workload and task complexity, we extracted some sub-tasks that contained typical actions from video recordings, such as checking the altimeter gauge or controlling the stick. We consider the operation difficulty of these sub-tasks is the same, and then we compared the number of executions of the sub-tasks and selected the sub-tasks that the pilots would frequently perform in the 4-h experiment. We found that pilots have to adjust the Mode Control Panel (MCP), as shown in Fig. 2 frequently during flight training, so we analyzed the physiological data of pilots while operating the MCP, including heart rate, respiratory rate, and pupil diameter, to find how the fatigue would affect these data.

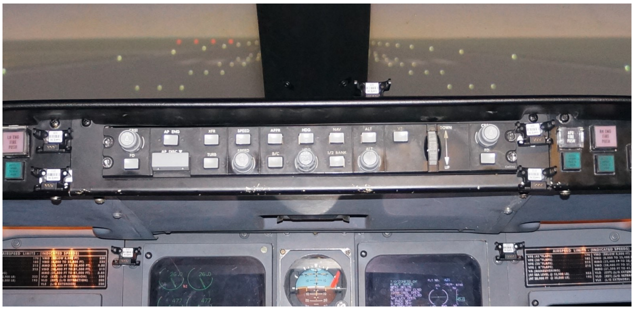


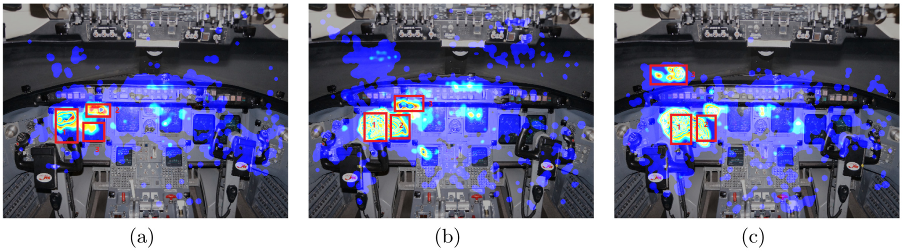
Fig. 2. Mode Control Panel (MCP) is the link between the pilots and the Autopilot flight Director System (AFDS)

However, the pilots' gaze is not always in the MCP area when operating the flying simulator. In order to analyze the change of the pilots' gaze points and to obtain the information about which part of the console that the pilots will focus on when they perform different tasks. We divided the console into several Areas of Interest (AOI) based on the statistical distribution of the pilots' gaze points in the experiment (we show this statistical distribution as a heat map, as shown in Fig. 3). Three AOIs are divided according to the heat map of each crew, one is the altimeter area that the pilot needs to observe at all flight tasks, and the other two AOIs are based on the brightness in the heat map. We choose the most obvious area and the second most obvious area as the second and third AOI.

The gaze points data captured by Tobii Glasses were local pixel position in its coordinate system. In order to obtain the world coordinate pixel position of gaze in the flight simulator, we calculated the transform matrix between these two coordinates system and transformed the gaze points from the original coordinate system to the world coordinate system of a scene camera.

$$\begin{pmatrix} W_x \; W_y \; 1 \end{pmatrix} = P \times \begin{pmatrix} X \; Y \; 1 \end{pmatrix} \tag{1}$$

where the W_x, W_y denotes the pixel position in the world coordinate, P denotes the transform matrix, and X, Y represents the pixel position in the Tobii coordinate.



(a) (b) (c)

Fig. 3. (a),(b),(c) are the heat map of three captain's gaze points, where the brighter the color, the more pilot's gaze points in the area. We divide the AOI according to the color levels. (Color figure online)

We classify eye movements based on the velocity of the direction of the eye. The velocity is most commonly given at visual levels per second ($°/s$). We determined that if the velocity of eye movement is below $30°/s$, the sample is classified as fixation. Therefore, with the above AOI and classified eye movement data, we can analyze the effect of fatigue on the eye data of pilots during the flight tasks. The average fixation duration of each crew focus on the different AOIs in experiment was calculated.

3 Experimental Results

In this section, measured data obtained in the experiment including heart rate, pupil diameter, and fixation duration of different AOIs are discussed. Since the participants would move their upper body in order to complete some training tasks during the experiment, the respiratory rate monitor will have poor contact with the participant's body resulting in a loss of respiratory data. Therefore, only a few numbers of participants' respiratory rate was fully recorded. We thought there a strong correlation between the heart rate and respiratory rate, so we did not discuss the effect of fatigue on the respiratory rate.

We performed ANOVA to analyze the difference of the physiological data. In the ANOVA analysis, the fatigue level accumulated over time is the independent variable, participant's heart rate, pupil diameter, and fixation duration are the dependent variables. The analysis program is run in Matlab under the Windows system.

3.1 Heart Rate Analyses

Each frame in the video recording of the experiment and the heart rate sampled by Zephyr Bio harness has a corresponding Coordinated Universal Time (UTC) timestamp. We recorded the start frame and the end frame of each time the pilot operates the MCP. By aligning the timestamp of the sample data with the timestamp of these two video frame, we obtained the heart rate data of the pilot while he performs the subtask, and we analyzed the changes in the pilot's heart rate with the accumulation of fatigue.

The boxplot of the heart rate of participants during one round of experiment as shown in Fig. 4. From the heart rate distribution in Fig. 4, we can have a result that the participants had a gradual decrease in heart rate with the accumulation of fatigue during the experiment. A one-way(fatigue) analysis of variance (ANOVA) performed for analyzed the effect of fatigue on heart rate. ANOVA results show that the mean heart rate was significantly affected by fatigue ($F(6, 217) = 11.27, P - value < .001$). In order to analyze more specifically the effect of fatigue on the participants, we analyzed the heart rate changes of each participant, and the heart rate graphs of two of the participants are shown in Fig. 5. In these two graphs, the y-axis represents the mean heart rate of participants operating MCP in flight tasks. The heart rate significantly decreases with the amount of time a participant was performed in the flight task. We performed ANOVA to analyze the effect of fatigue on the mean heart rate between each two crews. The ANOVA did not reveal significant effect of fatigue on different participants ($F(1, 14) = 0.35, P - value > 0.05$).

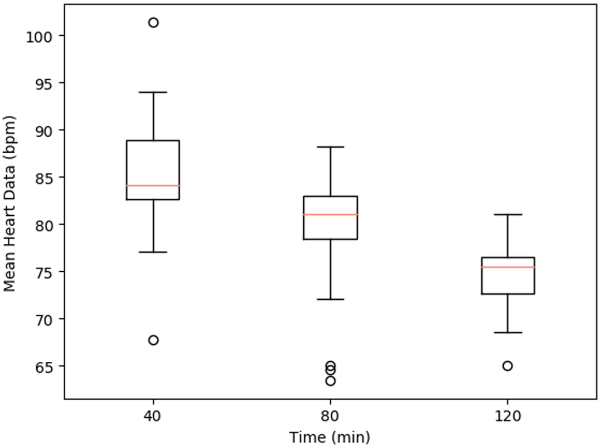


Fig. 4. The boxplot of the heart rate of participants. The y-aix represents heart rate (bpm), the x-aix represents flight training tasks performed by pilots in a round of experiments and there is a 40-min interval between every two tasks. Each data point represents the pilot's heart rate during the operation of MCP

3.2 Pupil Diameter Analyses

The process of pupil diameter extraction was the same as that of heart rate data extraction, which is also the data of participants during the operation of the

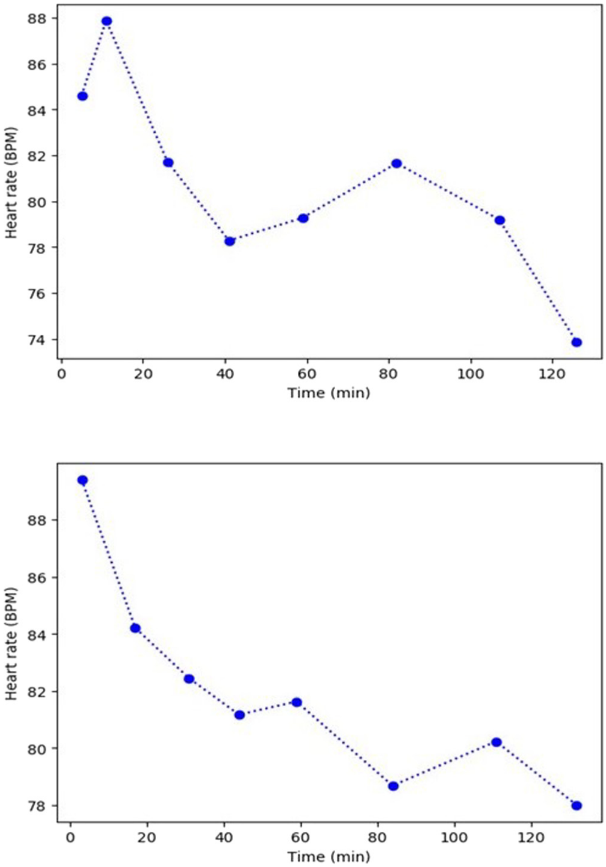


Fig. 5. The graph of two participants' mean heart rate

MCP. During this period the participant will look at the altimeter, and because of the light conditions of the cockpit and the fixed distance of the altimeter from the participants, the pupil diameter of the participants will not be interfered with by these two factors. The boxplot of pupil diameter of participants as shown in Fig. 6.

The ANOVA revealed a significant effect of the fatigue $(F(7, 361) = 6.71, P-value < 0.01)$, the pupil diameter decreases with the accumulation of fatigue, and the same trend are found in participants of different ages and flying ages. According to the results of the ANOVA, there was no significant difference in pupil diameter between each of two participants (Two of participants' mean pupil diameter are shown in Fig. 7, $F(1, 14) = 10.3, P - value > 0.05$).

As Fig. 6 and Fig. 7 showed, the pupil diameter decreases with time. It is well known that the pupil is not subjectively controlled, but is regulated by the automatic nervous system. The decreasing pupil diameter indicates that the

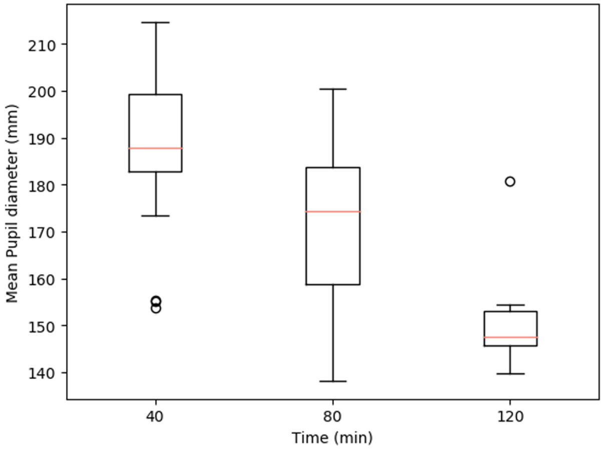


Fig. 6. The boxplot of pupil diameter of participants. The y-aix represents pupil diameter (mm).

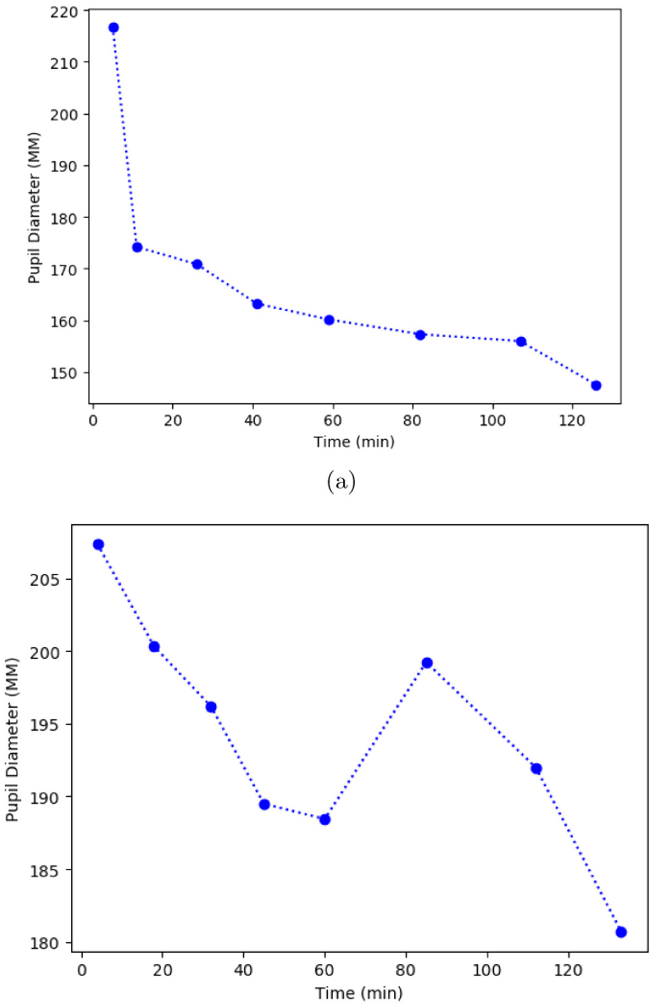
deceleration in activity of the autonomic nervous system, and the fatigue of the participants gradually increases.

3.3 Fixation Duration Analyses

We did not divide the AOI according to the location of aircraft cockpit instruments, because through the experimental recording video, we found that the participants do not pay attention to the aircraft windows and some other cockpit environment, so we considered showing the participant's gaze point distribution first and then selectively dividing the AOIs according to the distribution of gaze points. With this method, we can get the area where participants are more concerned throughout the experiment and analyze the effect of the fatigue levels on the fixation duration in the different AOIs. ANOVA was used to analyze whether the effect of fatigue on fixation duration was significant and analyze whether there were significant differences in the effects on each participant. As a result of a one-way ANOVA performed on these data, a significant effect of fatigue on mean fixation duration $(F(7, 136) = 2.51, P - value < 0.05)$, and we found that the fatigue has a main effect on the AOI2, which was the area of greatest interest to participants in the experiment. $(F(7, 40) = 2.3, P - value < 0.05)$. A one-way ANOVA performed on both AOI1 and AOI3 revealed no significant main effect of fatigue.

3.4 Multiple Regression Analysis

Through the analysis in Sect. 3.1, Sect. 3.2, and Sect. 3.3, it was demonstrated that fatigue had a significant effect on participants' pupil, heart rate, and fixation



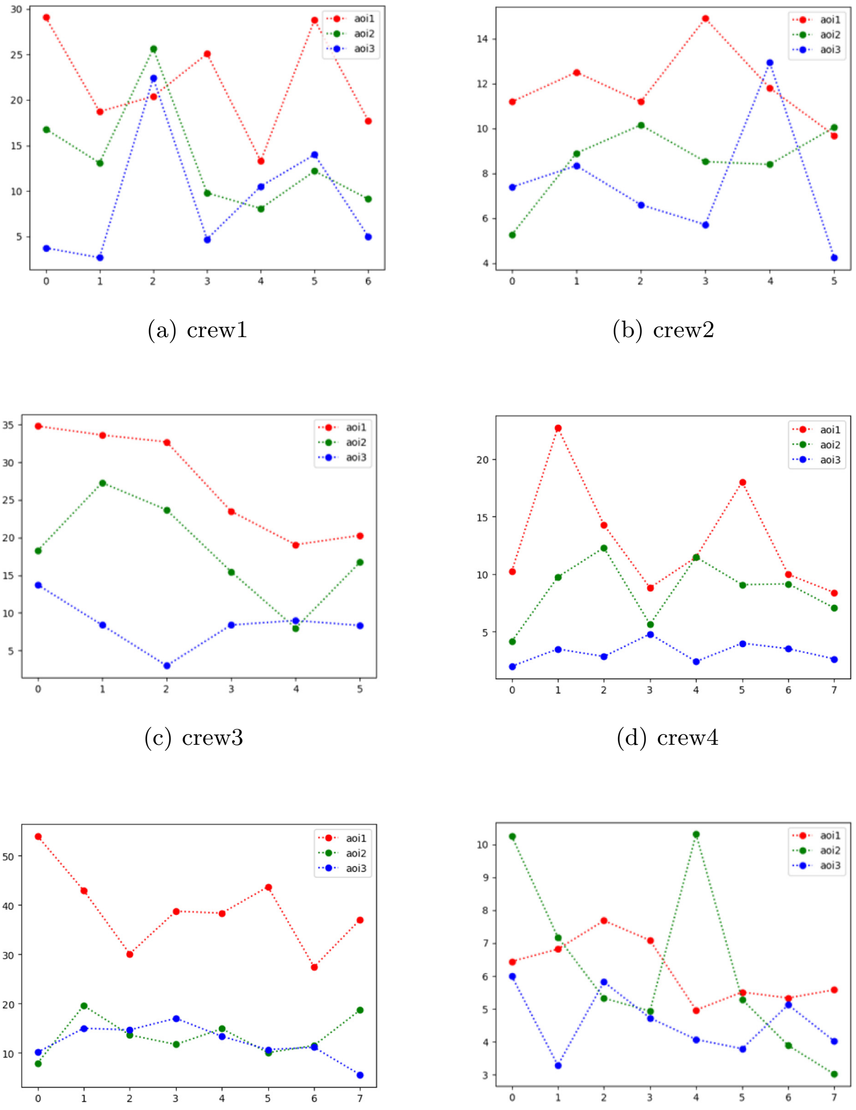
(a)

(b)

Fig. 7. The graph of two participants' mean pupil diameter

duration, so the physiological data also could reflect the changes in fatigue levels (Fig. 8).

Although the measured physiological data were affected by fatigue, they differed in the level of change and trend. We wanted to obtain a comprehensive indicator to directly reflect the fatigue level of the pilots during flight tasks. Therefor, we used multiple regression analysis to analyze these measured data. The fatigue was the dependent variable, and we consider it to be increasing with the amount of time. The physiological data of the participants corresponded to independent variables. Therefore, the heart rate, pupil diameter, and fixation



(a) crew1

(b) crew2

(c) crew3

(d) crew4

(e) crew5

(f) crew6

Fig. 8. The change of mean fixation duration in three different AOIs of six crews with time. The y-axis represents the mean fixation duration (Ms), and the x-axis note the number of flight tasks of a participant performed in one round of the experiment. The different color of the line represent different AOIs, which are divided int three areas according to the gaze points distribution. (Color figure online)

duration were entered into the multiple regression model. Moreover, according to the conclusion of Sect. 3.3, not all fixation duration are significantly correlated with fatigue, and only the fixation duration of participants in AOI2 (areas of greatest interest to participants) was affected by the fatigue. For the independent variable of fixation duration, only the fixation duration time of AOI2 was entered into the multiple regression model. Because the values and metrics of the physiological data were different for participant, we normalized these physiological data using the z-score normalization. The stepwise approach was used to select the significant independent variables in the regression model. The final regression equation was as follows:

$$RF = 2.67 * PL - 1.294 * HR + 0.8 * FD_{aoi2} + 5.5 \tag{2}$$

where the RF noted the rating score of fatigue, PL, HR correspond to the normalized pupil diameter and heart rate, the FD_{aoi2} represent the normalized fixation duration in the AOI2. The output of this model ranges from 0 to 10, with 0 indicated that the participants were not fatigue and 10 indicated that fatigue has reached a level that would affect the participants' operation. The square of multiple correlation coefficient was 0.56 (=0.75^2), and the results of ANOVA for multiple regression was $F(4, 28) = 11.36, P - value < 0.01$. The predicted fatigue score and the fitted curve as shown in Fig. 9. The result shows that the proposed index consisting of PL, HR, FD_{aoi2} yielded a satisfactory prediction to indicate the level of pilot fatigue during the flight tasks.

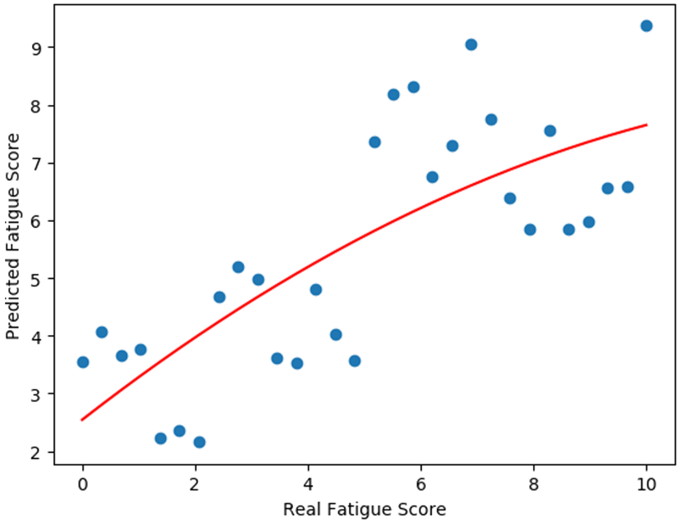


Fig. 9. The predicted fatigue score and fitted curve

Since we first normalized the physiological data, the magnitude of the participants' physiological data would not affect the correlation coefficient in Eq. 2, thus eliminating the effect of the difference in individual physiological base values. Therefore, the correlation coefficient is mainly related to the level of change of the physiological data and the range of the output, which makes this index of fatigue more applicable to various individuals.

4 Conclusion

In this paper, we designed a simulation experiment to study the effect of fatigue on physiological data, in which participants were asked to complete two 2-h rounds of the experiment. In order to simulate the real flying environment as much as possible, our experimental flight simulator was designed according to the real CRY-200 cockpit, and the tasks in the experiment are the tasks that pilots perform in their daily training. Moreover, since the pilots are specially trained population, their physiological data changes may be different from those of the general public, so for the accuracy of the experimental data, our participants were civilian aircraft pilots with flying age. We analyzed the effect of fatigue on participants' physiological data, and then proposed a composite physiological index to indicate the fatigue level.

The experimental results show that fatigue during flight tasks will affect the physiological data of the participants, and both the heart rate and pupil diameter will decrease as the fatigue level increases. However, for fixation duration, fatigue has different effects on the fixation duration of different AOIs. For example, the fatigue did not have a significant effect on the fixation duration of the altimeter are, while the fixation duration of the area that was participants paid the most attention decreased. According to the stepwise ANOVA, it was demonstrated that fatigue had a significant effect on physiological data.

Using physiological data that is significantly related to fatigue to entered to multiple regression model, we obtained a comprehensive index indication of the fatigue level of the participants while flying the flight simulator. The regression model can yield a satisfactory prediction to indicate the fatigue level. The results show that the fatigue level of pilots during the flight tasks could be effectively evaluated using the analysis method in this study. The equipment used in this study may not be available in the real flying context, and the extraction of physiological data relies on the operation in sub-tasks from the video recording. Further studies should investigate how to monitor fatigue of pilots in real time. In addition, future work should also investigate the physiological data that could continuously monitored during real flying context.

References

1. Michielsen, H.J., De Vries, J., Van Heck, G.L.: Psychometric qualities of a brief self-rated fatigue measure: the fatigue assessment scale. J. Psychosom. Res. **54**(4), 345–352 (2003)
2. Meesters, C., Appels, A.: An interview to measure vital exhaustion. I. Development and comparison with the Maastricht questionnaire. Psychol. Health **11**(4), 557–571 (1996)

3. Morales, J.M., Díaz-Piedra, C., Rieiro, H., et al.: Monitoring driver fatigue using a single-channel electroencephalographic device: a validation study by gaze-based, driving performance, and subjective data. Accid. Anal. Prev. **109**, 62–69 (2017)
4. Wijesuriya, N., Tran, Y., Craig, A.: The psychophysiological determinants of fatigue. Int. J. Psychophysiol. **63**(1), 77–86 (2007)
5. Murata, A., Uetake, A., Otsuka, M., et al.: Proposal of an index to evaluate visual fatigue induced during visual display terminal tasks. Int. J. Hum.-Comput. Interact. **13**(3), 305–321 (2001)
6. Leigh, R.J., Zee, D.S.: The Neurology of Eye Movements. OUP, Oxford (2015)
7. Gateau, T., Durantin, G., Lancelot, F., et al.: Real-time state estimation in a flight simulator using fNIRS. PloS ONE **10**(3), e0121279 (2015)
8. Causse, M., Dehais, F., Arexis, M., et al.: Cognitive aging and flight performances in general aviation pilots. Aging Neuropsychol. Cogn. **18**(5), 544–561 (2011)
9. Dehais, F., Behrend, J., Peysakhovich, V., et al.: Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: a behavioral and eye tracking study. Int. J. Aerosp. Psychol. **27**(1–2), 15–28 (2017)
10. Rowland, L.M., Thomas, M.L., Thorne, D.R., et al.: Oculomotor responses during partial and total sleep deprivation. Aviat. Space Environ. Med. **76**(7), C104–C113 (2005)
11. Schmidt, D., Abel, L.A., DellOsso, L.F., et al.: Saccadic velocity characteristics-intrinsic variability and fatigue. Aviat. Space Environ. Med. **50**(4), 393–395 (1979)
12. Prsa, M., Dicke, P.W., Thier, P.: The absence of eye muscle fatigue indicates that the nervous system compensates for non-motor disturbances of oculomotor function. J. Neurosci. **30**(47), 15834–15842 (2010)

The Relationship Between Cognitive Ability and Flight Driving Performance in Adolescent Pilot Cadets

Jun Tan[1,2] and Xianghong Sun[1,2](✉)

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China
sunxh@psych.ac.cn

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

Abstract. Cognitive ability test is an important part of pilot selection and is always considered as the best indicator to predict pilot training efficiency. China's pilot cadets are selected from the age of 18. At present, researchers have proposed whether this age can be appropriately advanced? Therefore, the research on the relationship between cognitive ability and flight performance in adolescents was started among 222 high school students in Junior Aviation School, in order to provide theoretical basis for the daily training of high school pilot cadets. In Study 1, 134 students completed 11 cognitive tests on Lenovo Myoga Tab 3 Plus tablets and experienced flight driving for the first time. Their flight driving performance were evaluated by their instructors. Pearson correlation results showed that spatial cognitive ability ($r = .244$) and cognitive flexibility ($r = -.186$) were significantly correlated with flight driving performance. Study 2 explored the relationship between executive function and flight driving performance. Inhibitory control, working memory and cognitive flexibility were tested as the index of executive functions with E-Prime 2.0 software (PST Inc.). The results of study 2 were similar to those of study 1. Spatial working memory ($r = -.185$) and cognitive flexibility ($r = .192$) were significantly correlated with first time flight driving performance. Our results support the idea that general cognitive ability is an important predictor of success in pilot training, even in early adolescence. In the future, more studies are needed to explore the plasticity of cognitive flexibility and spatial working memory in adolescents, and whether the improvement of cognitive ability is beneficial to flight driving.

Keywords: Cognitive ability · Flight performance · Pilot candidate · Adolescence

1 Introduction

With the development of economic globalization and the increase of international affairs, airplanes have become the first choice for people to travel. This makes Commercial aviation one of the world's largest industries. Pilot selection received a large amount of concern because a candidate who fails would mean a higher cost than what is invested for a more effective selection method. The notion that specific aspects of cognition play

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 52–64, 2021.
https://doi.org/10.1007/978-3-030-77932-0_5

a crucial role in aircraft accidents suggests the potential benefit of implementing more efficient cognitive screening procedures for pilots. Airlines also be required by aviation regulatory authorities to pay more attention to the psychological wellbeing of pilots and to the selection of them, to ensure a good ‘fit’ between the pilot and position. Mental health and stability, personality and cognitive performance are increasingly relevant in the pilot assessment and recruitment process. In addition, given the high cost to train as a pilot, precise selection and efficient training are very meaningful.

In the pilot selection history, Germany developed special medical standards for pilot in 1910 followed by the United States in 1912. In order to increase the effective of selection method, knowledge, cognitive ability, skills, and other individual factors have been studied by researchers. Italy was the first country to launch pilot selection research program. They concluded that attention, resistance to emotional stimuli, and psychomotor activity are qualities of good pilot should have. A good airplane pilots is one who to a sufficient speed of perception and to a notable degree of extension and distribution of attention, adds constancy, precision, coordinating ability of the psycho-motor activity and who possesses a sufficient inhibitory power emotive reactions not to be disturbed in the above functions on account of emotional stimulus [1]. The U.S. Army Air Services adopted a test battery which consisted of 23 tests that assessed a variety of medical, physical, cognitive, psychological, and psychomotor attributes [2]. Subsequently, Stratton, McComas [3] investigated the predictive utility of those test battery and found that the most promising tests for determining aviation ability were judgment of relative speed of moving objects, judgment of parabolic curves, and complex reaction time.

After the world war I, endurance, resistance to disorientation, and the speed and accuracy of responses to visual and auditory stimuli have the attention of researchers. Influenced by multiple aptitude theory, the research of pilot selection focused on intelligence, mechanical comprehension, psychomotor, personality, and spatial ability in world war II. After world war II, most countries focused on the development of new forms of multiple aptitude tests. Paper-and-pencil and apparatus-based testing were the main test methods during this period. Since 1970, studies of multiple aptitudes and psychomotor abilities have been prevalent [4]. In 1982, the Euro-NATO Aircrew Selection Working Group (ACSWG) was established to document information about pilot candidate selection procedures and recommend best practices. They found that the highest rated attributes for pilot were situation awareness, memorization, achievement motivation, reasoning, perceptual speed, time sharing, agreeableness, selective attention, response orientation, divided attention, emotional stability, and psychomotor coordination. In addition to traditional multiple aptitude batteries, some counties examined the utility of simulator-based apparatus tests that used a job sample approach in the 1970s. Since the 1990s, computer-based testing of pilot aptitude has become commonplace. The UK Royal Air Force identified five aircrew-ability domains: attention capacity, mental speed, psychomotor, reasoning, and spatial. Though there have been advances in psychological theory and measurement methods since world war I, the constructs being measured have remained very stable. These include medical/physical fitness, cognitive ability, academic achievement, life experiences (including aviation experience/knowledge), psychomotor ability, and personality. It is likely these constructs will continue to be important indictors of pilot training aptitude. Hilton [5] identified three theoretical aspects: psychomotor skill/quickness, intelligence/aptitude, and personality/character. As we can see from the pilot selection history, Cognitive abilities have been an important part in military pilot aptitude selection test batteries since World war I, although measurement methods and instruments vary cross many counties and different period.

Before 2015, pilot selection process commenced at the age of 18 in china. But adolescence is a sensitive period of cognitive abilities development [6]. There is greater plasticity in adolescence than adult period [7–10]. It means that training of cognitive abilities associated with flight skill in adolescence can gain greater benefit than that in adolescence. Advancing the time of screening and training process of pilot cadets seems meaningful. A large number of researches have focused on the relationship between specific cognitive ability and pilot driving performance in adult pilot [11–13]. So far, there are few researches concerning the relationship between cognitive abilities and flight skills in adolescent pilot cadets.

In the present study, considering that cognitive abilities have significant correlation with flight skill in the adult pilot group. We want to explore the relationship between cognitive abilities with flight skills in adolescent pilot cadets. There are two studies to explore this issue. A wide range of cognitive abilities including attention, spatial cognitive ability and executive function, were contained in study 1. Study 2 focused on the relationship between executive function with flight skills.

2 Study 1: The Relationship Between Cognitive Abilities and Flight Driving Performance in Adolescent Cadets

2.1 Methods

Participants. Participants were 143 high school pilot cadets (all male, ethnic Chinese, sophomore) from four aviation school in china. Participants were all righthanded. None of the participants had history of neurological or psychiatric illness.

Materials. The test battery including 10 cognitive tasks: processing speed task, go/nogo task, cognitive flexibility task, depth perception task, mental rotation task, direction judgment task, digital working memory span task, spatial working memory task, attention span task, 3-back task. In go/nogo task, participants were instructed to respond by a button on pad screen when a pure green circle appeared (Go trials) and withhold responding when a green circle with squares inside appeared (No-go trails). There are 45 Go trials and 15 No-go trials. In the 3-back task, participants compared the current letter with the letter presented two trials earlier. If both letters were the same, this was regarded as a match trial. If both letters differed, it was regarded as mismatch trial. Participants were asked to press the button for a match trial and withhold responding for a mismatch trial. There are 30 mismatch trials and 10 match trials. In the depth perception task, participants were presented with the destination and then destination was hidden by clouds. The participants press button if the airplane icon reached the destination according to the flying speed of airplane. In the cognitive flexibility task, participants were asked to click the digitals and letters randomly distributed on the screen in the order of 1A2B3C…. The total time completing the task were recorded. In the spatial working memory, participants needed to remember the flickering squares and point out all of them in the correct order to pass a trial. In the digital working memory task, participants were asked to solve an addition or subtraction equations of single digits one by one and were asked to remember the minuend digit. When all equations presented, participants need to write down all the minuends in presented order. In the mental rotation task, participants were presented with two rotated three-dimensional figures and were asked to judge whether the two figures are the same. In the attention span task, participants were presented with some white dots (4–9) on the screen in 200 ms. After the white dots disappeared, participants tell the number of dots by click digital button. In the processing speed task, participants were asked to click the digital (1–25) as order randomly distributed in the 5 × 5 squares. All tasks mentioned above were presented on LENOVO™ YOGA TAB 3 PLUS pad. The tests adopt the form of group measurement.

The flight driving performance scores are calculated from instructor pilot rating of cadets flying performance organized by the Air Force. The cadet will be accompanied and rated by the instructor during flight driving. The flight driving performance scores are composed of three indices, including 6th sortie score, instructor rating score and combined score. The score theoretically can range from 0 to 100. The higher the scores, the better the driving performance.

Experimental Design. Relevant research was conducted in study 1. There are 10 cognitive tasks in study 1 concerning attention, spatial cognitive ability and executive functions. The main dependent variable indicators of those tasks include accuracy and response time. Flight driving performance was obtained through a real military aircraft driving with aviation instructor besides. Instructor rates the driving performance of adolescent pilot candidate.

Data Analysis. All data were analyzed with SPSS 23. Pearson's correlation analysis were conducted to explore the relationships between 10 cognitive abilities with three indictors of plight driving performance. According to the Kolmogorov-Smimov test, all the indicators except for the 3-back task and the cognitive flexibility task are not normally distributed. The logarithmic transformation was used to normalize the original data, and then the bivariate Pearson correlation test was conducted.

2.2 Results

The descriptive results for the 10 cognitive tasks and flight driving performance are displayed at Table 1.

Table 2 summarizes the correlational analyses results. seven of the 36 correlations between the cognitive abilities and the flight driving performance were statistically significant. Direction judgement task was positively correlate with three flight driving performance scores. Spatial working memory task was positively correlate with two flight driving performance scores. Cognitive flexibility and mental rotation task were positively correlate with one flight driving performance scores. Spatial working memory task, direction judgement task, and mental rotation task measure the ability to maintain and manipulate spatial-related information in mind. Cohen characterized correlations of .10 as small, .30 as medium, and .50 or greater as large. All of the observed correlation between cognitive ability tasks with flight driving performance range small to medium.

Table 1. The descriptive statistics of 10 cognitive tasks

	Subject	Min	Max	Mean	SD	Skewness	Kurtosis
GO/NOGO, Go acc	143	.91	1	.99	.01	−6.66	57.17
GO/NOGO, Nogo ACC	143	.67	1	.93	.08	−1.3	1.3
3-back ACC	143	.39	.95	.68	.12	−.1	−.68
3-back RT (ms)	143	477.50	1615.00	998.44	196.04	.38	.67
Depth perception task (pixel)	143	38.86	281.37	121.06	44.47	1.11	1.48
Direction judgment task ACC	143	.00	1	.69	.28	−1.18	.27
Cognitive flexibility RT (s)	143	31.75	71.43	47.99	8.25	.59	−.12
Spatial working memory ACC	143	.10	1	.74	.20	−.82	.29
Digital working memory span	143	0	8	6.2	2.36	−1.54	1.50
Mental rotation ACC	143	.48	.92	.74	.09	−.71	.84
Attention span	136	4	9	7.41	1.69	−.93	−.45
Processing speed (s)	111	23.89	50.18	34.15	5.95	.53	−.35
6th sortie score	143	40	96	78.63	10.46	−.82	.73
Expert score	143	46	95	77.75	11.72	−.89	.18
Combined score	143	53.97	93.72	78.38	9.33	−.78	.04

Both processing speed task and attention span task show no significant correlation with flight driving performance. There are also no significant correlations between go/nogo task, 3-back task, and digit working memory span task with any indictors of flight driving performance.

The correlation coefficients among 6th sortie score, expert score, and combined score were large. This indicated that the three indictors for measuring flight driving performance have a high internal consistency.

2.3 Discussion

Spatial cognitive ability and cognitive flexibility show significant correlation with flight driving performance, consistent with previous findings found in adult pilot group [14–16]. It can be concluded that spatial ability and cognitive flexibility are important to drive airplane for adolescent pilot cadets.

No significant correlation between attention with three outcomes of pilot performance be found in our study. This result consistent with the conclusion concluded by Almamari and Traynor [17] that controlled attention (.10) had the lowest mean correlations, less than general ability (.18), acquired knowledge (.19) and work sample (.34).

Table 2. Correlation matrix of cognitive tasks and flight driving performance

	PS	AS	DP	DJ	SWM	MR	GNG Go	GNG NG	3-back ACC	3-back RT	DWMS	CF	6th sortie	Expert	Combined
PS	1	−.244*	−.014	−.180	−.333**	−.241*	−.154	−.154	−.189*	.175	−.107	.656**	−.103	−.148	−.123
AS	−.244*	1	−.039	.062	.139	.220*	.168	.089	.055	−.089	−.036	−.230**	.106	.157	.132
DP	−.014	−.039	1	−.164	.029	.029	−.182*	−.211*	.110	.051	−.015	.081	−.106	−.164	−.158
DJ	−.180	.062	−.164	1	.234**	.213*	−.059	.069	.093	.076	.066	−.270**	.186*	.181*	.195*
SWM	−.333**	.139	.029	.234**	1	.093	−.105	.097	.082	.075	.071	−.424**	.147	.244**	.223**
MR	−.241*	.220*	.029	.213*	.093	1	−.070	.002	−.051	−.027	−.017	−.196*	.122	.187*	.153
GNG Go	−.154	.168	−.182*	−.059	−.105	−.070	1	.087	.001	.016	−.040	−.011	−.094	−.079	−.085
GNG NG	−.154	.089	−.211*	.069	.097	.002	.087	1	.176*	.037	−.097	−.133	.061	.057	.068
3-back ACC	−.189*	.055	.110	.093	.082	−.051	.001	.176*	1	.284**	.123	−.107	.073	.071	.062
3-back RT	.175	−.089	.051	.076	.075	−.027	.016	.037	.284**	1	−.101	.142	.095	.115	.101
DWMS	−.107	−.036	−.015	.066	.071	−.017	−.040	−.097	.123	−.101	1	−.164	.070	.129	.124
CF	.656**	−.230**	.081	−.270**	−.424**	−.196*	−.011	−.133	−.107	.142	−.164	1	−.080	−.186*	−.153
6th sortie	−.103	.106	−.106	**.186***	.147+	.122	−.094	.061	.073	.095	.070	−.080	1	.835**	.911**
Expert	−.148	.157+	−.164+	**.181***	**.244****	**.187***	−.079	.057	.071	.115	.129	**−.186***	.835**	1	.975**
Combined	−.123	.132	−.158+	**.195***	**.223****	.153+	−.085	.068	.062	.101	.124	−.153+	.911**	.975**	1

1. PS = processing speed; AS = attention span; DP = depth perception; DJ = direction judgment; SWM = spatial working memory; MR = mental rotation; GNG Go = GO/NOGO Go acc; GNG NG = GO/NOGO NOGO acc; DWMS = digital working memory span; CF = cognitive flexibility.

Although perceptual speed be regarded as a significant predictor of aviator flying performance [16], we didn't found the significant correlation between processing speed task with flight driving performance in adolescent pilot cadets group.

Go/nogo task measures the ability to inhibit prepotent response which called response inhibition ability. 3-back task and digit working memory task measure the ability to maintain and manipulate information in mind, and also reflect the cognitive resource level. Pilots have higher level on cognitive resource than control group [18, 19]. But we didn't found significant correlations between response inhibition ability and working memory ability with flight driving performance in adolescent pilot cadets group. Differing from results of adult pilot group, it may indicate that inhibition response and working memory play a less important role while adolescent pilot cadets drive airplane.

Executive functions refer to a family of top-down mental processes needed when you have to concentrate and pay attention, when going on automatic or relying on instinct or intuition would be insufficient [20]. There is general agreement that there are three core EFs: inhibition (response inhibition and cognitive inhibition), working memory and cognitive flexibility [21]. Response inhibition and semantic working memory didn't show significant correlation with flight driving performance, while spatial working memory and cognitive flexibility show significant correlation with flight driving performance in study 1. Study 2 was conducted to confirmed the findings found in study 1.

3 Study 2: The Relationship Between Executive Function and Flight Driving Performance in Adolescent Pilot Cadets

3.1 Methods

Participants. 88 adolescent pilot candidates (all mate, ethnic Chinese, sophomore, Mage = 17.2) from high school aviation school in china were recruited in study 2. Participants were all righthanded. None of the participants had history of neurological or psychiatric illness.

Materials.

Demographic. A self-report questionnaire was used to obtain information on age, and other demographic descriptors.

GO/NOGO Task. In the GO/NOGO task, participants had to respond by pressing "A" or "L button when a white circle and arrow was presented and inhibit responding when a orange circle was presented. In the Go trial, participants press key "A" on the keyboard when the arrow is pointing to left, and press key "L" when the arrow is pointing to the right. There are 60 Go trials and 20 No-go trials.

Flanker Task. In the Flanker task, participants viewed stimuli presented one at a time surrounded by either distracting or facilitating items. Distracting items are typically associated with an opposite response ("incongruent" = pointing in opposite direction to target stimulus), whereas facilitating items are typically associated with the same response as the target stimulus ("congruent" = pointing in the same direction as the target stimulus). Participants press key "A" on the keyboard when the arrow is pointing to left, and press key "L" when the arrow is pointing to the right.

2-Back Task. In the 2-back task, participants compared the current digit with the digit presented one trials earlier. If both digit were the same, this was regarded as a match trial. If both digit differed, it was regarded as mismatch trial. Participants were asked to press the "Enter" button for a match trial and withhold responding for a mismatch trial. There are 37 mismatch trials and 17 match trials.

Spatial Working Memory Task. We adopted a computerized version of the Corsi Block-Tapping task, which proposed by Corsi in 1972 and has been widely used by many researchers to assess spatial working memory [22]. In every trial, an array of 25 squares in five columns and five rows appeared in the center of the screen, some of which then flickered one by one. The participants needed to remember the flickering squares and point out all of them in the correct order to pass a trial. The first session of three trials presented three flickering squares. If the participant passed two or three trials, he would pass the current session and enter the next with one more flickering square until he could not pass at least two trials in a session. The score was calculated by the number of squares in the last session passed. The order of the flickering squares was random.

Switch Task. On each trial, participants were presented with a digit located either on the left or right of the screen. They were told to perform Larger/Smaller task when digit was located on the left of the screen and to perform Odd/Even task when digit was located on the right of the screen. Performing Larger/Smaller task, press key "A" for digit less than 5 and press key "L" for digit more than 5. Performing Odd/Even task, press key "A" for digit was odd and press key "L" for digit was even.

Flight Driving Performance. Check-flight grades in flight driving tests were recorded. Instructor pilots assess candidate flight performance during each sortie using a scale ranging from A to E. A is the best and E is worst.

All of those executive function tasks were presented with the software Eprime 2.0 (PST Inc.). Participants were required to response as fast and accuracy as possible when performing tasks except for Corsi task.

Experimental Design. A desktop LENOVO computer with 23-inch monitor was used to administer the tasks used in this study. After completion of demographic information, executive function tasks were introduced. The administration order of tasks was GO/NOGO task, Flanker task, digital 2-back task, Corsi task and task-switch. There are exercise block before each test. The accuracy of exercise block of GO/NOGO, Flanker, 2-back task and task-switch were 85%, 85%, 85%, 90, respectively. A set of headphones were given to each participant in this study to reduce mutual interference.

Participants received the flight driving test in the summer of 2020 and completed the executive function tests in November of 2020.

Data Analysis. Statistical analyses were performed using SPSS 23. We performed a Kendall's correlation analysis to investigate whether executive functions correlated with the flight driving performance since flight driving performance were ranked data.

3.2 Results

The number of A, B, C, D, E is 21, 38, 21, 4, and 4, respectively.

The descriptive results of executive function tasks are shown in Table 3.

Table 3. Descriptive statistics of executive functions task

	N.	Mean	SD	Min	Max	Skewness	Kurtosis
Reaction Time (ms)	88	290.17	35.00	230.25	451.17	1.88	6.43
GNG Nogo ACC	88	0.95	0.05	0.80	1.00	−1.03	0.61
GNG GO ACC	88	0.98	0.02	0.88	1.00	−1.70	3.50
GNG GO RT (ms)	88	392.43	31.99	315.50	486.77	0.37	0.56
Flanker Effect ACC	88	0.10	0.08	−0.05	0.35	0.55	−0.06
Flanker Effect RT (ms)	88	67.81	29.23	−5.76	144.92	0.04	0.16
2-back Total ACC	88	0.93	0.06	0.70	1.00	−1.57	2.61
2-back False Alarm	88	0.04	0.05	0.00	0.24	1.92	3.86
2-back Hit	88	0.89	0.13	0.24	1.00	−2.00	5.91
2-back Hit RT (ms)	88	561.73	125.62	374.14	937.21	0.61	−0.19
Corsi	88	6.03	0.88	4.50	7.50	0.05	−0.69
Switch Cost ACC	88	0.01	0.08	−0.31	0.19	−0.62	2.22
Switch Cost RT (ms)	88	159.96	167.64	−250.58	653.52	0.36	0.20

Kendall's correlation analysis results show that Corsi task ($r = -.185$, $p < 0.05$) and reaction time cost of task-switch($r = .192$, $p < 0.05$) were significant correlated with flight driving performance. No significant correlations were found between simple reaction time, GO/NOGO task, Flanker task, 2-back task with flight driving performance. See Table 4 for detail.

Nogo trial accuracy of go/nogo task was significantly correlated with flanker effect accuracy ($r = 0.202$, $p < 0.05$). Corsi was significantly correlated with total accuracy ($r = 0.182$, $p < 0.05$) and hit accuracy ($r = 0.2$, $p < 0.05$) of 2-back task. No correlations were found between task-witching task with go/nogo task, flanker task, 2-back task, and Corsi task.

3.3 Discussion

In study 2, we explored the relationship between executive function with flight driving performance in order to confirm the results found in study 1 that spatial cognitive ability and cognitive flexibility were significantly correlated with flight driving performance. We focused on executive function because cognitive abilities significantly related to flight driving skill were included in executive function. We reselected the paradigms and participants to enhance the validity of our conclusions.

Table 4. Correlation matrix of executive function tasks and flight driving performance

	FDP	SRT	GNA	GGA	GGR	FEA	FER	2-back TA	2-back FA	2-back HA	2-back HR	Corsi	SCA	SCR
FDP	–													
SRT	0.021	–												
GNA	−0.089	0.028	–											
GGA	−0.015	0.046	0.26**	–										
GGR	0.084	0.212**	0.009	0.082	–									
FEA	−0.027	−0.011	−0.202*	−0.11	−0.234**	–								
FER	−0.026	−0.047	0.047	0.127	0.057	0.18*	–							
2-back TA	0.049	−0.115	0	−0.034	−0.127	0.167*	0.027	–						
2-back FA	−0.06	0.095	−0.013	−0.078	0.092	−0.151	−0.058	−0.699***	–					
2-back HA	0.026	−0.15	−0.084	−0.13	−0.186*	0.181*	0.018	0.753***	−0.341***	–				
2-back HR	−0.129	0.019	0.022	−0.082	0.075	−0.034	0.057	−0.065	0.103	−0.071	–			
Corsi	**−0.185***	−0.001	−0.053	0.127	0.006	0.023	0.045	0.182*	−0.15	0.2*	−0.082	–		
SCA	−0.044	0.077	0.013	0.123	−0.003	0.109	0.11	−0.012	−0.003	0.006	−0.045	0.014	–	
SCR	**0.192***	−0.063	0.044	0.019	−0.061	−0.042	−0.036	0.008	−0.006	−0.016	0.088	−0.03	0.065	–

1. FDP = flight driving performance; SRT = simple reaction time; GNA = GO/NOGO nogo acc; GGA = GO/NOGO go acc; GGR = GO/NOGO go rt; FEA = flanker effect acc; FER = flanker effect rt; TA = total acc; FA = false alarm; HA = hit acc; HR = hit rt; SCA = switch cost acc; SCR = switch cost rt.

2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In study 2, we chosen go/nogo task, flanker task, digit 2-back task, Corsi task, and task-witching task to measure response inhibition, cognitive inhibition, semantic working memory, spatial working memory, and cognitive flexibility, respectively. The correlation analysis showed that spatial cognitive ability and cognitive flexibility were significantly correlated with flight driving performance, which were consistent with the results of study 1.

Response inhibition significantly correlated with cognitive inhibition. Semantic working memory significantly correlated with spatial working memory. Semantic working memory significantly correlated with response inhibition and cognitive inhibition. The correlations among executive functions indicated that inhibition, working memory and cognitive flexibility are separated, consistent with the opinion of Miyake, Friedman [21].

4 Discussion

Our main objective was to investigate whether the notion that the relationship between cognitive ability and flight driving performance in adult pilot apply to adolescent pilot cadets. We found that spatial cognitive ability and cognitive flexibility are significantly correlated with flight driving performance in adolescent pilot cadets in study 1. Study 2 confirmed this result with different test tasks in another adolescent pilot cadets group. Spatial ability is one of the central competences of airline pilots in adult group [23, 24]. This notion is confirmed in adolescent cadets by our two experiments.

Although Carretta [25] confirmed that attention is among factors that have consistently shown a relation to flying performance and King, Carretta [12] proved that tests of cognitive functioning, which include attention, can predict the training outcomes. Rosario Cannavò [26] also compared attention factors directly between 50 experienced pilot and 50 untrained people. The results showed that tasks performances of pilot group are better than that of control group, including simple reaction time, multiple search, color-word interference, ground interference, divided attention, and global vision. We didn't found the significant direct correlation between attention with flight driving performance in study 1. It seems suggested that attention was less relevant to flight skill in adolescent pilot cadets.

Inhibition control and working memory also be regarded as two important aspects, given lots of information, such as monitoring cockpit information and receiving instructions from Air Traffic Control, are needed to deal with during flight driving. The facts that pilots have a higher working memory span and inhibition control ability than control group were confirm by Rosario Cannavò [26]. But the results of our studies showed no significant correlations between working memory and inhibition control with flight driving performance. It seems that working memory and inhibition control plays a less important role when flight driving in adolescent cadets. With the increasing of flight driving experience, the role of working memory and inhibition control are becoming more and more important.

From the perspective of compiling a cognitive ability test battery to screen pilot candidate, it's not that the more cognitive abilities included in test battery is the best [17]. Performing a wide range of ability tests during the selection process does not necessarily improve the predictions or guarantee a better selection model. Almamari and Traynor [17] confirmed that different test battery categories yielded different mean validities, indicating that some abilities play a greater role in pilot performance than others. Optimal balance between the quality and quantity of the tests included in the battery is what really matters. Bringing the cognitive abilities closely relevant to pilot skill into test battery can have the better predictive validity. We recommended that more focus should be given to spatial cognitive ability (e.g., spatial working memory) and cognitive flexibility when compiling pilot selection test battery.

The training of pilot is particularly difficult and expensive. For this reason, the selection processes are of fundamental importance. The aim of this research is to guide the development of an optimized cognitive screening tool well-suited for the successful selection of pilot candidates. Despite its potential utility, spatial working memory and cognitive flexibility testing is relatively uncommon in pilot candidate selection. Practically, use of spatial working memory and cognitive flexibility testing in selection of adolescent pilot cadets has the potential to reduce costs associated with pilot training and improve the accuracy rate of excellent pilot. We contributed to establish that spatial working memory and cognitive flexibility are reliable cognitive abilities to predict pilot driving performance.

Of course, more researches are needed to confirm the relationship between cognitive ability and pilot driving performance founded in adolescent pilot performance.

References

1. Dockeray, F.C., Isaacs, S.: Psychological research in aviation in Italy, France, England, and the American expeditionary forces. J. Comp. Psychol. **1**(2), 115–148 (1921)
2. Yerkes, R.M.: Report of the psychology committee of the national research council. Psychol. Rev. **26**(2), 83–149 (1919)
3. Stratton, G.M., et al.: Psychological tests for selecting aviators. J. Exp. Psychol. **3**, 405–423 (1920)
4. Hunter, D.R., Burke, E.F.: Pilot selection: meta-analysis of the research findings. In: 19th Conference of the Western European Association for Aviation-Psychology. WEAAP, Nice (1991)
5. Hilton, T.F., Dolgin, D., Pilot selection in the military of the free world. In: Handbook of Military Psychology. Wiley, New York (1991)
6. Thompson, A., Steinbeis, N.: Sensitive periods in executive function development. Curr. Opin. Behav. Sci. **36**, 98–105 (2020)
7. Mak, C., et al.: Efficacy of mindfulness-based interventions for attention and executive function in children and adolescents-a systematic review. Mindfulness **9**(1), 59–78 (2018)
8. Rolle, C.E., et al.: Enhancing spatial attention and working memory in younger and older adults. J. Cogn. Neurosci. **29**(9), 1483–1497 (2017)
9. Karbach, J.: Plasticity of executive functions in childhood and adolescence: effects of cognitive training interventions. Rev. Argentina Ciencias Comport. **7**(1), 64–70 (2015)
10. Wass, S.V., Scerif, G., Johnson, M.H.: Training attentional control and working memory - is younger, better? Dev. Rev. **32**(4), 360–387 (2012)
11. Carretta, T.R., et al.: Consistency of the relations of cognitive ability and personality traits to pilot training performance. Int. J. Aviat. Psychol. **24**(4), 247–264 (2014)
12. King, R.E., et al.: Standard cognitive psychological tests predict military pilot training outcomes. Aviat. Psychol. Appl. Hum. Factors **3**(1), 28 (2013)

13. Causse, M., Dehais, F., Pastor, J.: Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. Int. J. Aviat. Psychol. **21**(3), 217–234 (2011)
14. Causse, M., Chua, Z., Matton, N.: Performance and brain activity during a spatial working memory task: application to pilot candidate selection. In: Ayaz, H. (ed.) AHFE 2019. AISC, vol. 953, pp. 45–55. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-20473-0_5
15. Caponecchia, C., Zheng, W.Y., Regan, M.A.: Selecting trainee pilots: predictive validity of the WOMBAT situational awareness pilot selection test. Appl. Ergonomics. **73**, 100–107 (2018)
16. Johnson, J.F., et al.: Predictive validity of spatial ability and perceptual speed tests for aviator training. Int. J. Aerosp. Psychol. **27**(3–4), 109–120 (2017)
17. Almamari, K., Traynor, A.: Multiple test batteries as predictors for pilot performance: a meta-analytic investigation. Int. J. Sel. Assess. **27**(4), 337–356 (2019)
18. Verde, P., et al.: Domain-specific interference tests on navigational working memory in military pilots. Aerosp. Med. Hum. Perform. **87**(6), 528–533 (2016)
19. Verde, P., et al.: Gender differences in navigational memory: pilots vs. nonpilots. Aerosp. Med. Hum. Perform. **86**(2), 103–111 (2015)
20. Diamond, A.: Executive functions. In: Fiske, S.T. (ed.) Annual Review of Psychology, vol. 64, pp. 135–168 (2013)
21. Miyake, A., et al.: The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cogn. Psychol. **41**(1), 49–100 (2000)
22. Wang, H., et al.: Working memory: a criterion of potential practicality for pilot candidate selection. Int. J. Aerosp. Psychol. **28**(3–4), 64–75 (2019)
23. Verde, P., et al.: Spatial orientation and directional judgments in pilots vs. nonpilots. Aerosp. Med. Hum. Perform. **89**(10), 857–862 (2018)
24. Sladky, R., et al.: Neurobiological differences in mental rotation and instrument interpretation in airline pilots. Sci. Rep. **6**, 1–6 (2016)
25. Carretta, T.R.: Pilot candidate selection method: still an effective predictor of US air force pilot training performance. Aviat. Psychol. Appl. Hum. Factors **1**(1), 3–8 (2011)
26. Cannavo, R., Conti, D., Di Nuovo, A.: Computer-aided assessment of aviation pilots attention: design of an integrated test and its empirical validation. Appl. Comput. Inf. **12**(1), 16–26 (2016)

An Improved Washout Algorithm for UPRT Scenario

Wei Tang, Zhen Wang, and Shan Fu(✉)

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, People's Republic of China
sfu@sjtu.edu.cn

Abstract. In recent years, with the continuous upgrading of technology, flight accidents caused by aircraft failures have been greatly reduced. It has become the most important cause of major aviation accidents that pilots operate aircraft into complex state and fail to recover. It is very important to train pilots with UPRT, so it is urgent to improve the UPRT performance of existing flight simulators. This paper proposes an improved washout algorithm for UPRT scenario. The key point of the algorithm is to combine the advantages of the washout algorithm based on model predictive control and the classical washout algorithm. The washout algorithm based on model predictive control (MPC washout algorithm) can simulate large amplitude and low frequency motion better, and has greater space utilization, while the classical washout algorithm is efficient and simple, and has better simulation effect for small amplitude and high frequency motion. By setting the exponential weighted fusion algorithm to fuse the output of the two algorithms, it can have better dynamic simulation effect in large and small motion, and is more suitable for UPRT scenario. In addition, it also saves the time of designing and debugging algorithms or parameters separately for different missions.

Keywords: Classical washout algorithm · UPRT · MPC · Exponential weighted fusion algorithm

1 Introduction

At present, in the field of civil aviation, the progress of aircraft manufacturing technology has effectively reduced the aircraft accident rate. In the past decade, loss of control in flight (LOC-I) has replaced controlled flight to the ground (CFIT) as the number one cause of serious flight accidents [1]. In the field of aviation, UPRT refers to complex state prevention and recovery training. The FAA announced that all new flight simulators in the United States must have UPRT capability from March 12, 2019 [2].

After flight simulation training in UPRT scenario, pilots can greatly improve their flight experience and ability of preventing and identifying complex states and changing them in complex states. In UPRT scenario, dynamic simulation is the most important part in the process of simulator design. Because the complex state in UPRT scenario generally has large amplitude and high frequency motion, it is necessary to improve

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 65–79, 2021.
https://doi.org/10.1007/978-3-030-77932-0_6

the motion simulation algorithm to achieve large amplitude and high frequency realistic motion simulation.

The motion simulation algorithm of simulator is generally called washout algorithm, which is also called motion cuing algorithm (MCA). It is widely used in the motion simulation design of mechanical and electrical products such as aviation simulator, vehicle simulator, navigation simulator and motion cinema.

Washout algorithm can simulate the movement of pilots in the limited space. When the simulator provides the pilots with motion sense, it is restricted by the available motion space. In order to enable pilots to sense motion without exceeding the physical limits of the motion system, various motion washout filters have been used in the past. A good tradeoff between motion perception and displacement requires minimizing the perceptual motion error and keeping the platform within the physical boundary [3–6].

The main organ of human perception of movement is vestibular system, which is mainly composed of otolith and semicircular canal, in which the otolith mainly perceives linear acceleration, and the semicircular canal mainly perceives angular velocity of rotation [7, 8]. The washout algorithm outputs the linear acceleration of otolith perception and angular velocity of semicircular canal perception in vestibular system of human brain by filtering and integrating, and controls the motion platform to generate motion by the output signal. Finally, the motion platform slowly returns to the initial position at a speed lower than human perception threshold, and prepares for the next motion simulation.

The first one to propose washout algorithm is Schmidt [9, 10], an engineer of NASA. At this time, washout algorithm includes high pass filter module and low pass filter module. Schmidt and Conrad's work were later revised and further studied by Reid and Nahon of the Institute of Aeronautics and Astronautics (UTIAS) of the University of Toronto, adding the tilt coordination module, resulting in the generation of classical washout algorithm [11–13]. The classical washout algorithm has the advantages of simple structure, fast and effective. In 1975, in order to solve the problem of fixed parameters and low space utilization of the classical washout algorithm, Parrish et al. proposed a Coordinated Adaptive washout algorithm [14]. Adaptive washout algorithm can adjust algorithm parameters according to different inputs. In 1982, Sivan et al. applied optimal control theory to washout algorithm and proposed optimal washout algorithm [15]. The optimal washout algorithm introduces the vestibular system of human to directly optimize the calculation of the driver's perception error, which is more accurate. The classical washout algorithm has the advantages of fast running speed, simple structure and suitable for adjustment, but it also has some defects such as insufficient utilization of motion space and fixed parameters. Adaptive washout algorithm can make up for the problem of fixed parameters, but it takes a long time to run and is difficult to debug. The optimal washout algorithm introduces the driver's vestibular model, which has the highest fidelity, but the complexity is relatively increased, and the constraints are not taken into account.

The above three algorithms are the most important three kinds of washout algorithms. Now most simulators use one of the three algorithms, or make some minor modifications. Until 2003, Song et al. proposed a classical washout filter based on fuzzy tuning for motion simulator to overcome the problem of fixed parameters and adaptation to different signals [16]. After that, Asadi, Mohammadi and others have done some research on fuzzy adaptive washout algorithm [17, 18].

In 2009, Dagdelen et al. proposed a washout algorithm based on model predictive control (MPC washout algorithm), which also attracted some attentions [19]. Garrett, Mohammadi and Salisbury have made further research on MPC washout algorithm [20–24]. MPC washout algorithm takes advantage of online real-time optimization of model predictive control and considering constraints. Therefore, MPC washout algorithm can make better use of platform space and achieve better dynamic simulation effect.

This paper proposes an improved washout algorithm for UPRT Scenario. By fusing the output of MPC washout algorithm and classical washout algorithm, the improved washout algorithm proposed in this paper can not only adapt to high frequency and low amplitude motion simulation, but also simulate low frequency and large amplitude motion. The MPC algorithm can deal with the multivariable nonlinear system and is suitable for solving dynamic simulation problems. MPC washout algorithm can take constraints into account, so it can make more efficient use of the motion space of the motion platform. For large amplitude and low frequency motion, it has better simulation fidelity than the classical washout algorithm. Although the parameters of the classical wash out algorithm are fixed and the space utilization is low, the classical washout algorithm has the advantages of fast, simple and stable. MPC washout algorithm has better effect for high frequency and small amplitude motion simulation. The innovation of this paper is that by setting the exponential fusion weight function and fusing the output of classical washout algorithm and MPC washout algorithm, a more balanced dynamic simulation output is obtained.

2 Method

2.1 Vestibular System

Vestibular system accounts for about 30% of people's perception of movement. Vestibular system is composed of semicircular canals and otoliths [7, 8]. The otolith and semicircular canal can sense the linear acceleration and angular velocity above a certain threshold d_{TH}, δ_{TH} respectively. So, the linear acceleration sensed by otolith is the vector sum of external force acceleration and negative gravity acceleration, also known as specific force:

$$f = a - g \tag{1}$$

where: f – specific force; a – actual acceleration; g – acceleration of gravity.

Otolith organ transfer function:

$$\frac{f}{f^{'}} = \frac{k_h(\tau_A s + 1)}{(\tau_L s + 1)(\tau_S s + 1)} \tag{2}$$

where: $f^{'}$ – input specific force; f – specific force after perception; k_h – sensing gain coefficient; τ_A, τ_L, τ_S – physical parameters.

There semicircular canals transfer function models:

$$\frac{w}{w'} = \frac{T_L T_A s^2}{(T_L s + 1)(T_S s + 1)(T_A s + 1)} \tag{3}$$

where: w'− the input angular velocity of the semicircular canal organ; w− angular velocity of semicircular canal organs after perception; T_L, T_S, T_A− physical parameters of otolith model.

2.2 Classical Washout Algorithm

The classical washout algorithm is mainly composed of three parts: linear acceleration high pass filter, angular velocity high pass filter and linear acceleration low pass filter [10] as follow:

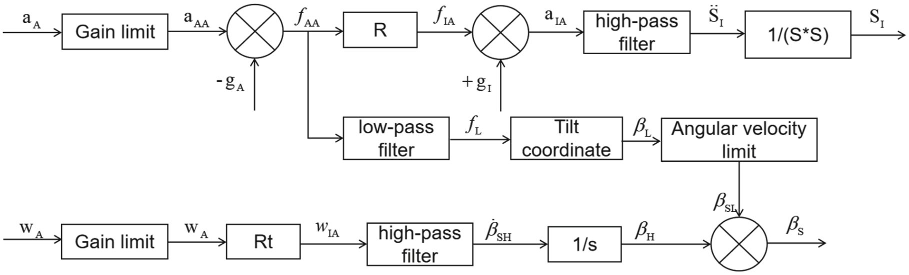


Fig. 1. Schematic block diagram of classical washout algorithm

In Fig. 1, to adapt to different platform space, the gain limit module is added. The input acceleration and the reverse gravity acceleration are added to obtain the specific force vector. Then remove the acceleration of gravity in the fixed coordinate system.

It processed by a high pass filter. At last, the high-frequency linear displacement component is obtained by secondary integration.

The high pass filter is a third-order filter:

$$H_{ah} = \frac{a_h}{a} = \frac{s^3}{(s + w_m)(s^2 + 2w_{ah}\xi_{ah}s + w_{ah}^2)} \tag{4}$$

where: w_{ah}, w_m is the natural cut-off frequency, ξ_{ah} is the filter damping.

The acceleration low-pass filter has only the longitudinal and lateral second-order transfer functions:

$$H_{alx} = \frac{f_{Lx}}{f_x} = \frac{w_{alx}^2}{(s^2 + 2w_{alx}\xi_{al}s + w_{alx}^2)} \tag{5}$$

$$H_{aly} = \frac{f_{Ly}}{f_y} = \frac{w_{aly}^2}{(s^2 + 2w_{aly}\xi_{aly}s + w_{aly}^2)} \tag{6}$$

where, w_{alx}, w_{aly} is the natural cut-off frequency. ξ_{alx}, ξ_{aly} is the longitudinal and lateral acceleration.

Low pass acceleration channel is also called tilt coordination channel. The components of gravity acceleration in lateral and longitudinal directions are used to simulate the low-frequency sustained acceleration.

$$\theta_{sL} = sin^{-1}\frac{f_{Lx}}{g} \tag{7}$$

$$\varphi_{sL} = sin^{-1}\frac{f_{Ly}}{g} \tag{8}$$

where, $\theta_{sL}, \varphi_{sL}$ is the angular displacement; f_{Lx}, f_{Ly} is the acceleration.

The high pass filter module are as follows:

$$H_{wh} = \frac{w_h}{w} = \frac{s^2}{(s^2 + 2w_{wh}\xi_{wh}s + w_{wh}^2)} \tag{9}$$

where, w_{wh} is the natural cut-off frequency; ξ_{wh} it is the damping coefficient.

We get the target angular displacement β_S:

$$\beta_s = \beta_{sH} + \beta_L \tag{10}$$

2.3 Washout Algorithm Based on Model Predictive Control (MPC Washout Algorithm)

For UPRT scenario, the combination of model predictive control and washout algorithm can make the motion simulation process have more space utilization rate, the control of each variable is more accurate. The MPC washout algorithm is generally divided into four modes: longitudinal mode (pitch angular velocity and longitudinal acceleration), lateral mode, up and down mode and yaw mode. In this paper, we use the vestibular system as a prediction model (Fig. 2):

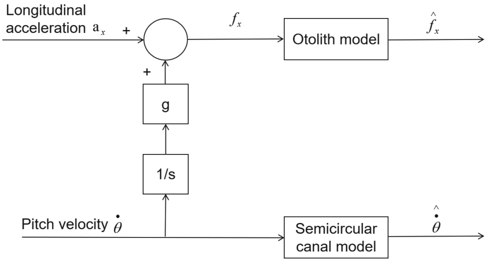


Fig. 2. Block diagram of vestibular system in longitudinal mode

The specific forces perceived by the vestibular system:

$$f_x^{'} = a_x cos\theta + gsin\theta \cong a_x + g\theta \tag{11}$$

Transfer function based on otolith model, and differential equation:

$$\frac{f}{f'} = \frac{k_h(\tau_A s + 1)}{(\tau_L s + 1)(\tau_S s + 1)} \tag{12}$$

$$f_x = \frac{k_h(\tau_A s + 1)}{(\tau_L s + 1)(\tau_S s + 1)} * (a_x + g\frac{1}{s}\dot{\theta}) \tag{13}$$

Similarly, the transfer function of the semicircular canal:

$$\hat{\dot{\theta}} = \frac{T_L T_A s^2}{(T_L s + 1)(T_S s + 1)(T_A s + 1)}\dot{\theta} \tag{14}$$

Combine the two equations, for the state $u = \begin{bmatrix} a_x & \dot{\theta} \end{bmatrix}^T$, we can derive:

$$\dot{x_v} = A_v x_v + B_v u \tag{15}$$

$$\widehat{y_v} = C_v x_v + D_v u \tag{16}$$

A_v, B_v, C_v, D_v were fusion model state space equation parameters of vestibular system. It was used as prediction system of the MPC washing algorithm. The block diagram of the whole MPC washing algorithm is as follows (Fig. 3):

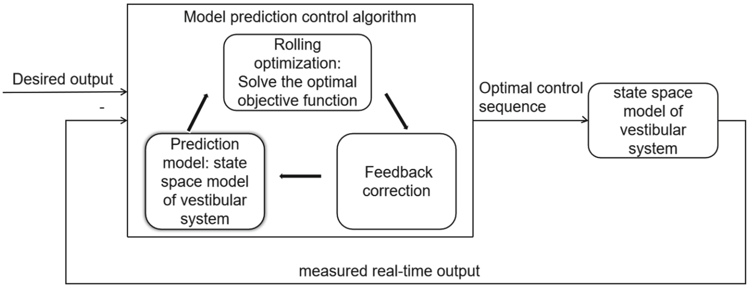

Model prediction control algorithm
Rolling optimization: Solve the optimal objective function
Desired output
-
Prediction model: state space model of vestibular system
Feedback correction
Optimal control sequence
state space model of vestibular system
measured real-time output

Fig. 3. MPC washout algorithm schematic module diagram

2.4 Exponential Weighted Fusion Algorithm

Due to the fixed parameters of the classical washout algorithm, it is not flexible enough, so its space utilization is very limited. The abnormal complex state in UPRT scenario is usually the state of large angular velocity and acceleration and high frequency. In this case, MPC washout algorithm can better simulate the motion. In addition, for different flight training tasks, such as take-off, climb, landing these basic tasks, as well as encounter stall, airflow, storm and other complex situations, using a single classical wash out algorithm obviously could not meet the training requirements. For example, if the washout algorithm in normal cruise is used to simulate the motion of stall flight, the fidelity of simulation will be greatly reduced. Furthermore, it is time-consuming and energy-consuming to reset a set of parameters or algorithms for other scenarios. The exponential weighted fusion algorithm proposed in this paper is to fuse the outputs of the classical and MPC washout algorithm. To a certain extent, the two complementary washout algorithms are combined to make it more balanced to simulate the movement of more complex states.

The principle of the improved algorithm is: when the input value range is small, the weight of the output value of the classical washout algorithm is expected to be greater; when the input value range is large, the weight of the output value of MPC washout algorithm is expected to be greater. The weights of the output values of the classical washout algorithm and the MPC washout algorithm vary between [0 1], and the sum of the weights is 1, so as to achieve the goal of balancing the dynamic simulation of multiple scenes. In addition, the exponential weighted fusion algorithm has almost the same effect as the classical and MPC washout algorithm for large and small amplitude motion. The weight change function of exponential weighted fusion algorithm is as follows:

$$x = x_0 * e^{\beta|u|} \tag{17}$$

$$\beta = \frac{1}{5} * \log_e(\frac{x_f}{x_0}) \tag{18}$$

$$Class\_Weight = |2 - x| \tag{19}$$

$$MPC\_Weight = |x - 1| \tag{20}$$

where: $x_f = 2, x_0 = 1, u \in [01]$.

For the two washout algorithms, the constraints are set in advance, and the maximum input value is no more than 5. Draw the graph of exponential weight distribution function, the independent variable is the input value, and the input value changes between 0–5.

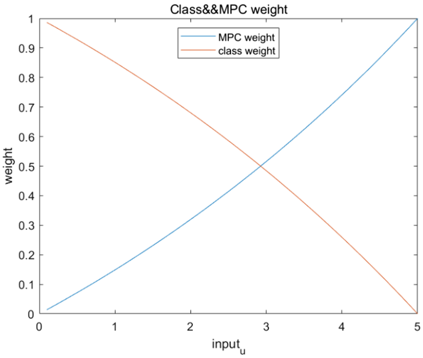


Fig. 4. Class&&MPC weight (Color figure online)

The blue line in the Fig. 4 represents the weight of the output value of MPC algorithm, as the input-u increases from 0 to 5, the weight of the output value of MPC algorithm increases from the lowest 0 to the highest 1; the red line represents the weight of the output value of classical washout algorithm, and with the increase of the input amplitude, the weight value gradually decreases from the maximum 1 to the minimum 0. The weight curves of classical washout algorithm and MPC washout algorithm meet the design requirements.

3 Experimental Design and Result Analysis

The experiment in this paper is a simulation experiment. The simulation program of washout algorithm is based on MATLAB r2020a Simulink. In order to ensure the simplicity of the experiment, this paper analyzes the longitudinal acceleration and pitching angular velocity data of simulated flight in longitudinal mode. In addition, the flight data is obtained from the motion variable data of simulated flight in flightgear2.8.0.5 flight software.

In main process of this experiment, first we set up different flight tasks and simulate flying Boeing 777 in flightgear2.8.0.5 software. In the flight process, the data of six degrees of freedom in the flight process is used as the input source data through network communication. Then the human vestibular system model perceives the output, and the perceived output is filtered by MPC and classical washout algorithm respectively to get the final output. The specific steps of the experiment are as follows:

1. Write C++ program, set socket client, set UDP network communication protocol interface and IP address, set UDP communication protocol and configuration file of output data of flightgear2.8.0.5 software.
2. Establish three block diagrams of Simulink washout algorithm. Firstly, the program block diagram of the classical washout algorithm is designed, and the input signal is transferred into the program block diagram of the classical washout algorithm after passing through the vestibular system. Then the simulation block diagram of MPC washout algorithm is designed. Finally, the exponential weighted fusion algorithm is designed to fuse the output values of classical and MPC washout algorithm program.
3. Simulate four flight missions in flightgear2.8.0.5 software:

   a) Take off, climb to the designated altitude, level off.
   b) Cruise flight, take a larger angle of attack flight.
   c) Increase the angle of attack to stall and return to level flight after several seconds.
   a) Take off in crosswind to the specified altitude, then keep level.

Each flight process lasts about 100 s−200 s, and the motion data of six degrees of freedom in the flight process is $a_x, a_y, a_z, w_\varphi, w_\theta, w_\psi$. These six values represent the linear acceleration along the x, y, z axis and the angular velocity around x, y, z axis of the simulator respectively.

Finally, we get the curves of longitudinal acceleration and pitching angular velocity measured by the classical washout algorithm and MPC washout algorithm, the output value after exponential weighted fusion, the expected value curve, and the original data not perceived by vestibular system. A total of five curves are drawn in the same graph.

The way of drawing pictures is the same, here we will explain the basic format and meaning of pictures. There are four pairs of graphs, each of which represents the longitudinal acceleration and pitching acceleration curves of a flight mission. The horizontal axis of all figures represents the time, which starts from 0s. In each pair of graphs, the longitudinal axis of the acceleration curve represents the acceleration in m/s^2, and the longitudinal axis of the angular velocity curve represents the angular velocity in rad/s. In each of the four pairs of experimental result graphs, there are five curves with different colors, which represent respectively:

1. Light green curve: raw linear acceleration or angular velocity data
2. Purple curve: the linear acceleration or angular velocity output value of the original data perceived by the vestibular system of human brain
3. Yellow curve: output value of linear acceleration or angular velocity processed by MPC wash out algorithm program
4. Light yellow curve: linear acceleration or angular velocity output value processed by exponential weighted fusion algorithm program
5. Brown curve: the output value of linear acceleration or angular velocity processed by classical wash out algorithm program (Figs. 5 and 6).

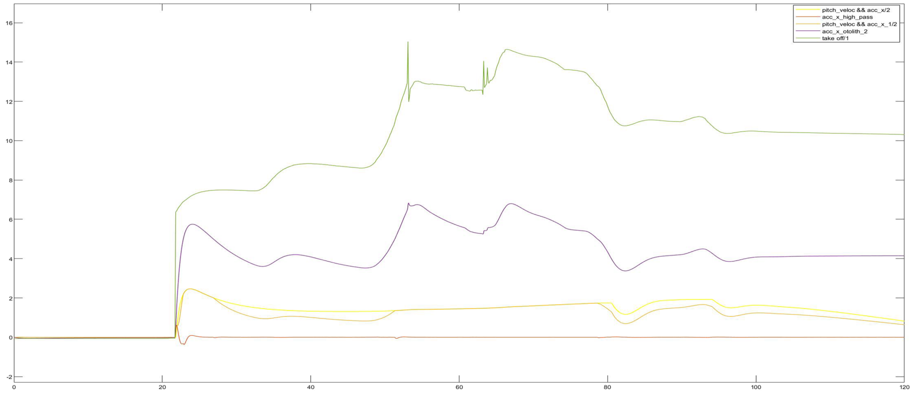


Fig. 5. Take off—longitudinal acceleration curve

The above two figures are the change curves of longitudinal acceleration and pitching angular velocity during normal takeoff. The acceleration increases from $0\,\mathrm{m/s^2}$ *to* $14.5\,\mathrm{m/s^2}$, then decreases, tends to be stable and reaches the level flight state.

From the dynamic simulation of acceleration, the simulation effect of the classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively better, and the MPC washout algorithm is the best. From the dynamic simulation of angular velocity, when the amplitude is less than $0.5\,\mathrm{rad/s}$, such as $t = 22\,\mathrm{s}{-}25\,\mathrm{s}$, the classical washout algorithm is better; the exponential weighted fusion algorithm is relatively poor, and the MPC washout algorithm is the worst. When the amplitude is greater than $0.5\,\mathrm{rad/s}$, for example, when $t = 45\,\mathrm{s}{-}50\,\mathrm{s}$, the simulation effect of the

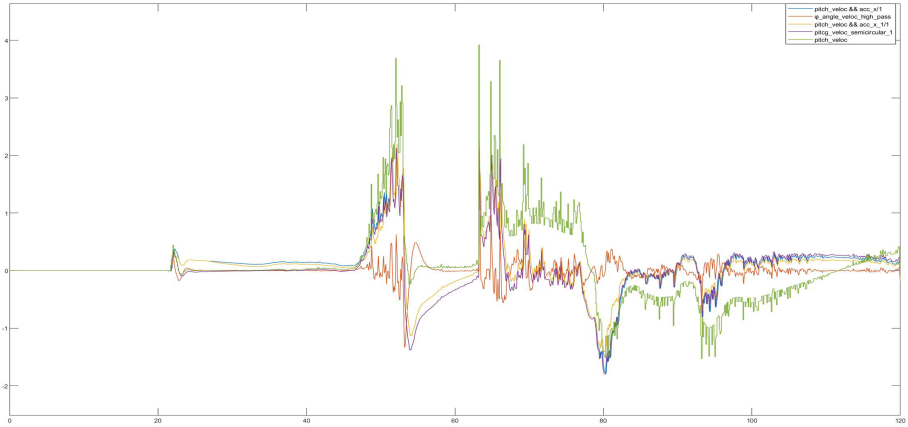


Fig. 6. Take off—pitch velocity curve

classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively better, and the MPC washout algorithm is the best (Figs. 7 and 8).

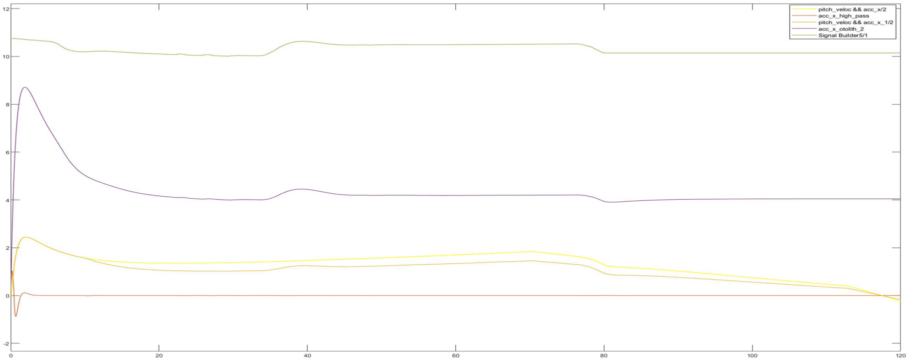


Fig. 7. Cruising flight—longitudinal acceleration curve

The above two figures are the curves of longitudinal acceleration and pitching angular velocity during simulated cruise flight. The acceleration is basically maintained between $10\,\mathrm{m/s^2}-12\,\mathrm{m/s^2}$, which is a level flight state. The pitching angular velocity is affected by the mis-operation at the beginning, reaching $1.6\,\mathrm{rad/s}$, and fluctuates up and down in the range between $0.2\,\mathrm{rad/s}-0.2\,\mathrm{rad/s}$ after it keeps stable.

From the dynamic simulation of acceleration, similar to the normal take-off state, the simulation effect of the classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively better, and the MPC washout algorithm is the best. From the dynamic simulation of angular velocity, the whole is in a small range, because the plane is in level flight, the small fluctuation may also be caused by error and noise. Due to the influence of the perception threshold of vestibular system, the washout

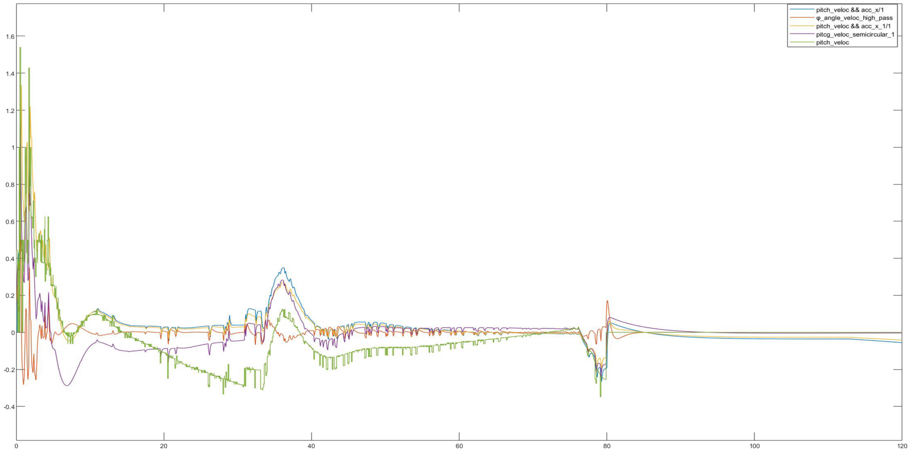


Fig. 8. Cruising flight—pitch velocity curve.

algorithm has little effect, and the demand for dynamic simulation is insufficient (Figs. 9 and 10).

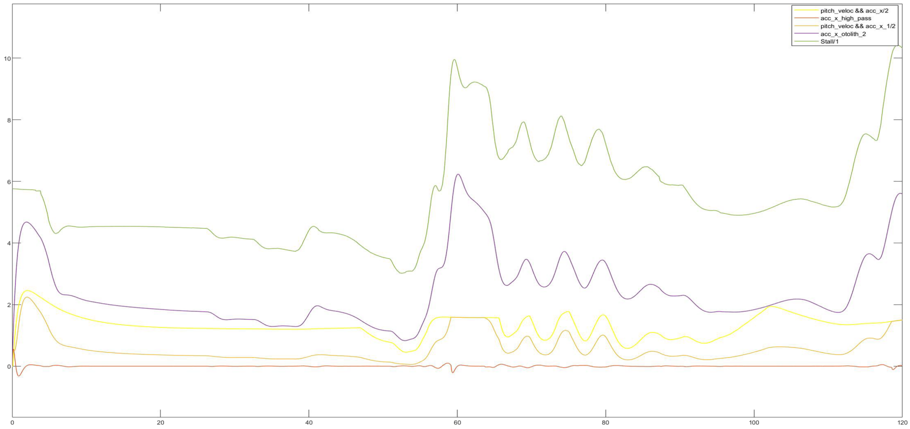


Fig. 9. Excessive Angle of attack stall—longitudinal acceleration curve

The above two figures are the curves of longitudinal acceleration and pitch angular velocity during the simulation of stall caused by high attack angle flight. In the process of stall, the acceleration is changing rapidly, with a wide range of $4\,\mathrm{m/s^2}{-}10\,\mathrm{m/s^2}$. The maximum and minimum difference of pitch angular velocity is $13\,\mathrm{rad/s}$, and there is a direction transition from positive to negative, which is very dangerous.

From the dynamic simulation of acceleration, similar to the previous situation, the simulation effect of the classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively better; the MPC washout algorithm is the best. From the dynamic simulation of angular velocity, when the amplitude is less than

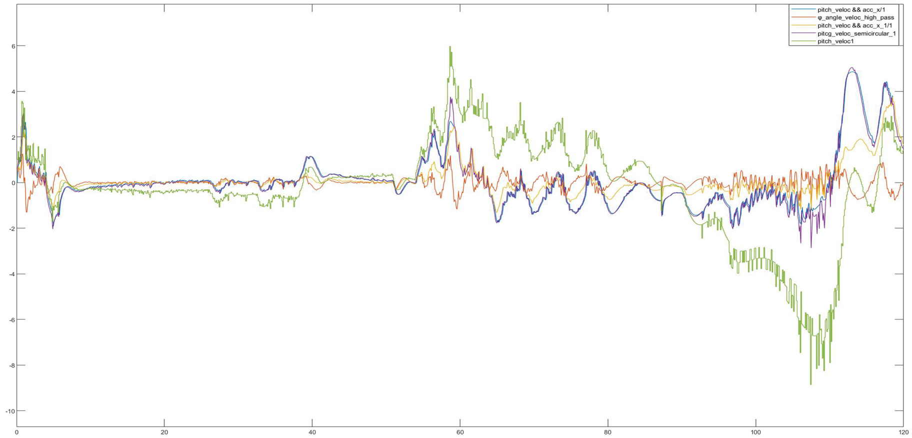


Fig. 10. Excessive Angle of attack stall—pitch velocity curve

0.5 rad/s, for example $t = 80\,\text{s} - 90\,\text{s}$, the classical washout algorithm is better; the exponential weighted fusion algorithm is relatively poor, and the MPC washout algorithm is the worst. When the amplitude is greater than 0.5 rad/s, for example $t = 50 - 60\,\text{s}$, the simulation effect of the classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively better; the MPC washout algorithm is the best (Figs. 11 and 12).

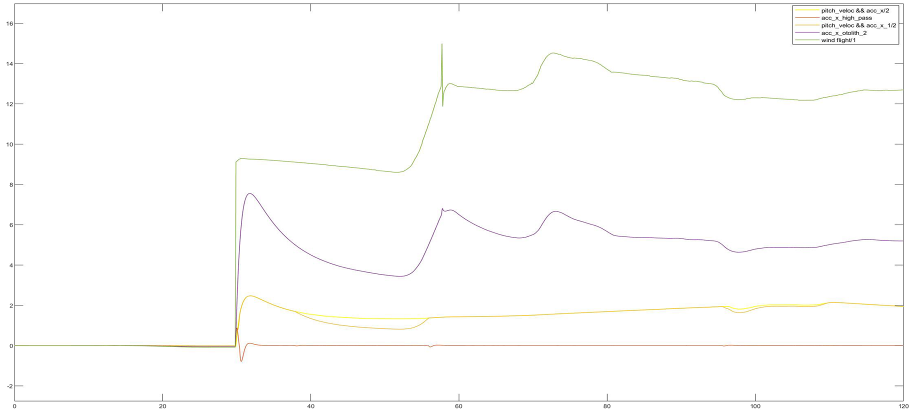


Fig. 11. Cross wind take-off—longitudinal acceleration curve.

The above two figures are the curves of longitudinal acceleration and pitching angular velocity during simulated crosswind takeoff mission. From the beginning of take-off, the acceleration increases from $0\,\text{m/s}^2 - 15\,\text{m/s}^2$, and then decreases gradually, but the acceleration curve is very steep, which is not as smooth as that of normal takeoff. The maximum pitching angular velocity of crosswind takeoff is about 3.2 rad/s, and the

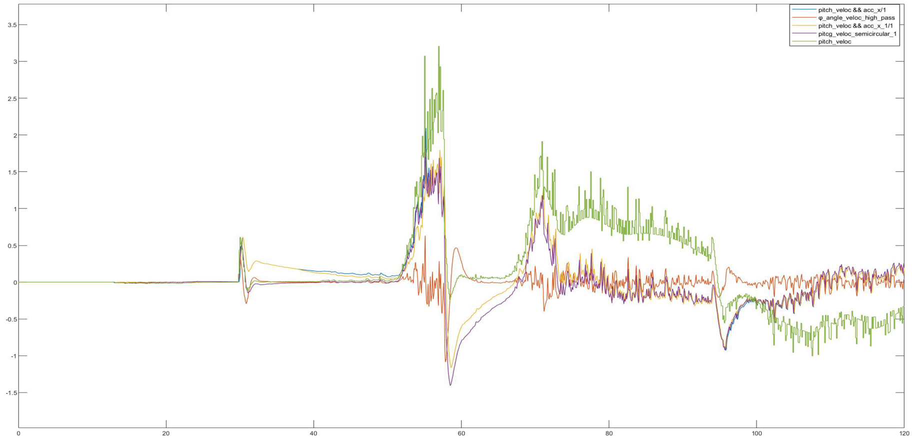


Fig. 12. Cross wind take-off—pitch velocity curve

minimum pitching angular velocity is about $-1\,\mathrm{rad/s}$. On the whole, the change range is quite large, and similar to that state of normal takeoff.

From the dynamic simulation of acceleration, similar to the previous situation, the simulation effect of the classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively good, and the MPC washout algorithm is the best. From the dynamic simulation of angular velocity, when the amplitude is less than 0.5rad/s, for example $t = 28\,\mathrm{s}{-}32\,\mathrm{s}$, the classical washout algorithm is better; the exponential weighted fusion algorithm is relatively poor, and the MPC washout algorithm is the worst. When the amplitude is greater than 0.5rad/s, such as $t = 52\,\mathrm{s}{-}58\,\mathrm{s}$, the simulation effect of the classical washout algorithm is the worst; the exponential weighted fusion algorithm is relatively good, and the MPC washout algorithm is the best.

The above four experiments designed washout algorithm to deal with acceleration and angular velocity signals respectively for ordinary flight state and complex flight state. Compared with three different washout algorithms, it shows a trend in four kinds of missions. Under the low frequency and large range of dynamic data, the simulation accuracy of MPC washout algorithm is higher, followed by the accuracy of exponential weight fusion algorithm, and the effect of classical washout algorithm is relatively poor. For high frequency and small amplitude motion data, the simulation effect of the classical washout algorithm is relatively good, and the following degree is high. While the accuracy of the exponential weight fusion algorithm is second, and that of the MPC washout algorithm is relatively poor. In general, the exponential weight fusion algorithm is the best for the balance between the complex state and the ordinary state, and can better simulate the dynamic data of low frequency and high frequency.

4 Conclusion

Based on the actual requirements of civil aviation field for flight simulator with UPRT function, this paper proposes an improved washout algorithm for UPRT scenario. In this paper, the exponential weighted fusion algorithm is used to combine MPC washout algorithm with classical washout algorithm, which makes the improved washout algorithm have better adaptability to UPRT scenario.

This paper first introduces the vestibular system, introduces the perception process of otolith and semicircular canal to linear acceleration and angular velocity, and gives the transfer function model and parameter table of otolith and semicircular canal respectively. Then, this paper introduces the classical washout algorithm framework, which is mainly composed of three channels. The operation principle, limiting and saturation are explained. The classical washout algorithm has the advantages of fast running speed, simple structure and easy design, but also due to the fixed parameter. Then, this paper introduces the principle and implementation process of MPC washout algorithm. MPC washout algorithm is suitable for multivariable nonlinear systems, and can take the input and output constraints into account. Finally, an improved washout algorithm is introduced, which combines MPC washout algorithm and classical washout algorithm by using exponential weighted fusion algorithm. The improved algorithm has better simulation effect in the UPRT environment, and improves the adaptability to different training tasks. To a certain extent, it can avoid the workload of designing and debugging parameter sets or algorithms for different tasks.

The experimental results also verify the adaptability and balance of the improved washout algorithm in small high frequency and large low frequency scenes. The improved washout algorithm can better simulate the motion in UPRT scenario.

There are still many problems need to be further studied, such as the rules for setting the constraints of the fusion algorithm and the principles for setting the range of index weight. For the experimental process, the introduction of the physical model of the simulator Stewart model should be more convincing. Finally, the measurement of realistic effect of pilots trained by washout algorithm is also a topic worthy of study [25]. It is also hoped that the dynamic simulation experiment of stall flight can be carried out in the actual simulator in the future, and the ergonomics evaluation of pilots' training results can be carried out, which is the place that needs to be improved in the future.

References

1. Boeing, Statistical Summary of Commercial Jet Airplane Accidents Worldwide Operations, 1959–2017
2. Rosenkrans, W.: Brave new world: FAA simulator requirements enable acceptably realistic recoveries from full stalls and upsets in commercial jets. AeoSafety World, July-August 2016
3. Chang, Y.-H., Liao, C.-S., Chieng, W.-H.: Optimal motion cueing for 5-DOF motion simulations via a 3-DOF motion simulator. Control Eng. Pract. **17**(1), 170–184 (2009)
4. Aminzadeh, M., Mahmoodi, A., Sabzehparvar, M.: Optimal motion cueing algorithm using motion system kinematics. Eur. J. Control **18**(4), 363–375 (2012)
5. Lee, W.-S., Kim, J.-H., Cho, J.-H.: A driving simulator as a virtual reality tool. In: Proceedings of the 1998 IEEE International Conference on Robotics and Automation, vol. 1, pp. 71–76. IEEE (1998)
6. Telban, R.J., Cardullo, F.M.: Motion cueing algorithm development: human-centered linear and nonlinear approaches. Technical report, NASA/CR-2005–213747 (2005)
7. Asadi, H., Mohamed, S., Lim, C.P., Nahavandi, S., Nalivaiko, E.: Semicircular canal modeling in human perception. Rev. Neurosci. **28**(5), 537–549 (2017)

8. Asadi, H., Mohamed, S., Lim, C.P., Nahavandi, S.: A review on otolith models in human perception. Behav. Brain Res. **309**, 67–76 (2016)
9. Conrad, B., Schmidt, S.F.: A study of techniques for calculating motion drive signals for flight simulators. NASA CR-114345 (1971)
10. Conrad, B., Schmidt, S., Douvillier, J.: Washout circuit design for multi-degrees of freedom moving base simulators. In: Proceedings of the AiAA Visual and Motion Simulation Conference, Palo Alto (CA), 10 September 1973, vol. 12 (1973)
11. Reid, L., Nahon, M.A.: Flight simulation motion-base drive algorithms: part 1. Developing and testing equations. University of Toronto, Technical report (1985)
12. Reid, L., Nahon, M.A.: Flight Flight simulation motion-base drive algorithms: part 2. Selecting the system parameters. University of Toronto, Technical report (1986)
13. Reid, L., Nahon, M.: Flight simulation motion-base drive algorithms. Part 3: Pilot evaluations (1986)
14. Parrish, R.V., Dieudonne, J.E., Martin Jr., D.J.: Coordinated adaptive washout for motion simulators. J. Aircr. **12**(1), 44–50 (1975)
15. Sivan, R., Ish-Shalom, J., Huang, J.-K.: An optimal control approach to the design of moving flflight simulators. IEEE Trans. Syst. Man Cybern. **12**(6), 818–827 (1982)
16. Song, J.-B., Jung, U.-J., Ko, H.-D.: Washout algorithm with fuzzy-based tuning for a motion simulator. KSME Int. J. **17**(2), 221–229 (2003)
17. Asadi, H., Mohammadi, A., Mohamed, S., Nahavandi, S.: Adaptive translational cueing motion algorithm using fuzzy based tilt coordination. In: Loo, C.K., Yap, K.S., Wong, K.W., Beng Jin, A.T., Huang, K. (eds.) ICONIP 2014. LNCS, vol. 8836, pp. 474–482. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12643-2_58
18. Asadi, H., Mohammadi, A., Mohamed, S., Rahim Zadeh, D., Nahavandi, S.: Adaptive washout algorithm based fuzzy tuning for improving human perception. In: Loo, C.K., Yap, K.S., Wong, K.W., Beng Jin, A.T., Huang, K. (eds.) ICONIP 2014. LNCS, vol. 8836, pp. 483–492. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12643-2_59
19. Dagdelen, M., Reymond, G., Kemeny, A., Bordier, M., Maizi, N.: Model-based predictive motion cueing strategy for vehicle driv-ing simulators. Control Eng. Pract. **17**(19), 995–1003 (2009)
20. Garrett, N.J.I., Best, M.C.: Model predictive driving simulator motion cueing algorithm with actuator-based constraints. Veh Syst Dyn **51**(8), 1151–1172 (2013). https://doi.org/10.1080/00423114.2013.783219
21. Mohammadi, A., Asadi, H., Mohamed, S., Nelson, K., Nahavandi, S.: MPC-based motion cueing algorithm with short prediction hori-zon using exponential weighting. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE (2016)
22. Salisbury, I.G., Limebeer, D.J.: Optimal motion cueing for race cars. IEEE Trans. Control Syst. Technol. **24**(1), 200–215 (2016)
23. Mohammadi, A., Asadi, H., Mohamed, S., Nelson, K., Nahavandi, S.: Multiobjective and interactive genetic algorithms for weight tuning of a model predictive control-based motion cueing algorithm. IEEE Trans. Cybern. **49**(9), 3471–3481 (2018)
24. Qazani, M.R.C., Asadi, H., Khoo, S., Nahavandi, S.: A linear time-varying model predictive control-based motion cueing algorithm for hexapod simulation-based motion platform. IEEE Trans. Syst. Man Cybern.: Syst. (2019)
25. Pool, D.M., Zaal, P.M.T., Damveld, H., van Paassen, M. M., Mulder, M.: Evaluating simulator motion fidelity using in-flight and simulator measurements of roll tracking behavior. In: AIAA Modeling and Simulation Technologies Conference, August 2012

Emotional Stressor on Human Errors in Flight: A Heart Rate Variance Examination

Shuo Wang(✉), Lei Wang, and Shu Li

Flight Academy, Civil Aviation University of China, Tianjin, China

Abstract. This study aimed to examine the effect of emotional stressor on human errors and arousal level in flight. 16 male pilot students were recruited as the participants whom were required to complete one flight scenarios with tasks in three different performance levels in low arousal (LA) and high arousal (HA) conditions (with or without emotional stressor) respectively. The photoplethysmograph physiological sensor was used to measure heart rate (HR) and heart rate variance (HRV) variables, the indicators of emotional arousal levels. The result indicated that emotional stressor can cause an increase in emotional arousal level and a decrease in human's skill-based and rule-based errors. Besides there was a quadratic trend for low frequency normalized unit (LF(nu)) and the number of calculation deviation (NCD), a dimension of problem-based error, which provided a plausible explain for the non-significance of the decrease in problem-based errors between HA and LA condition combined with inverted U-shaped curve. These findings have potential applications in preventing human errors before they occur.

Keywords: Emotional stressor · Human error · Arousal level · HRV

1 Introduction

As aircraft has become more reliable, human has played a more important role in contributing to flight accidents. The safety report provided by Civil Aviation Administration of China (CAAC) shows that the flight accidents caused by flight crews accounted for 67.90% of the total number in China in recent 10 years [1]. The unsafe acts consisting violations and errors of flight crews were regarded as the direct reason of flight accidents [2]. But unlike the violation, the error was defined as an unintentional deviation from operation procedure or standard. Moreover, nearly 40% aviation accidents caused by flight crews globally can be blamed for flight control errors, the skill-based error [3].

To prevent the harm caused by human errors, many studies used different theory models to investigate the mechanisms behind them. Norman [4] proposed the activation-trigger-schema system (ATS) to illustrate the formation of slip arguing that slip can be divided three categories: errors in formation of an intention, faulty activation of schemas and faulty triggering. Rasmussen [5] distinguished human behavior or performance according to basically different ways of representing the constraints in the behavior of a deterministic environment or system at three typical levels: skill-, rule- and knowledge-based performance. Based on Norman and Rasmussen' work, Reason [6] established

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 80–90, 2021.
https://doi.org/10.1007/978-3-030-77932-0_7

the generic error modelling system (GEMS) presenting an integrated picture of the error mechanism operating at the three levels of performance mentioned above and distinguishing these three basic error types at eight dimensions. In order to investigate human errors in aviation accidents, Wiegmann et al. [7] put forward Human Factors Analysis and Classification System (HFACS), in which the term of error was divided into three types: skill-based error, perceptual error and decision error. And rule- and knowledge-based errors were regarded as decision error because both of them were caused by wrong intentions. Landman [8] proposed the conceptual model of startle and surprise used to explain the impairments in flight performance in accident cases.

Besides theory model study, flight simulation test is also an important way to explain the mechanism under human errors in flight. Since 1980s there were lots of studies investigating human errors by flight simulation tests based on different theories, but a few referring to performance levels [9, 10].

In flight, pilots are suffering multiple sources of stressors, one type of which is emotional stressors [11]. According to Maymand [12], emotional stressors are related to living or intellectual activities. When perceiving emotional stressors, self-regulating processes start by automatically activating the autonomic nervous system (ANS) [13]. The ANS consists of the parasympathetic nervous system (PNS) and the sympathetic nervous system (SNS) [14]. Electrodermal activity (EDA) also known as skin conductance is an independent index of SNS activities while heart rate (HR) and heart rate variance (HRV) are influenced by interaction between the PNS and SNS activities [15]. So fluctuations in emotional arousal regulated by the ANS, are commonly indicated by HR, HRV and EDA. Besides pupil size modulated by balanced activities of PNS and SNS, is also considered an useful indicator of the level of emotional arousal [16].

There are some studies believing the high stress is a key contributor to human errors in aviation [17, 18]. However, other studies argue that stress can reduce the possibilities of errors and improve flight performance [12, 19]. The effect of stress on errors, a dimension of performance, can be well illustrated in term of arousal [20–22]. Yerkes-Dodson law speculates that the relationship between task performance and arousal level can be represented by an inverted U-shaped curve, which means that there is an optimal arousal level for a task performance and when one's arousal level is too high or too low, performance is predicted to be impaired [23, 24]. Moreover compared to difficult tasks, in simple tasks performers have a higher threshold for arousal, and therefore, an arousing stimulus can be tolerated well. So while a very high arousal level can be advantageous in simple tasks, for difficult tasks it can cause a decrease in performance [25]. For instance, Murray's work reveled a significant quadratic trend for low frequency normalized units (LF(nu)) and complex tasks including the post trail making task and the 4-choice reaction time task instead of simple tasks, and concluded that simple tasks can tolerate a larger variation of arousal levels and participants who were experiencing lower arousal would complete the simple reaction time task as well as someone who was experiencing high arousal [26].

In this study, we examine the effect of the emotional stressor on emotional arousal level and human errors in flight. Moreover to investigate the relationship between arousal level and human errors.

2 Method

2.1 Participants

16 male pilot students (20.11 ± 1.03 years) from Civil Aviation University of China(CAUC) were recruited as the participants. All participants had taken part in simulated flight before this test and were familiar with simulator environments. All of them were given written informed consent 1 h before carrying out the experiment in accordance with the local ethical board committee. This test was carried out in the Aviation Safety and Human Factors Lab of CAUC.

2.2 Equipment

The flight simulator which included two parts: flight control part and flying environment part, was used. The flight control part was composed of control column, a throttle controller and a pair of rudder pedals for flight control and flying environment part is made of three LCD screens. And the flight scenarios were designed based on the virtual world in X-Plane 11 (see Fig. 1).

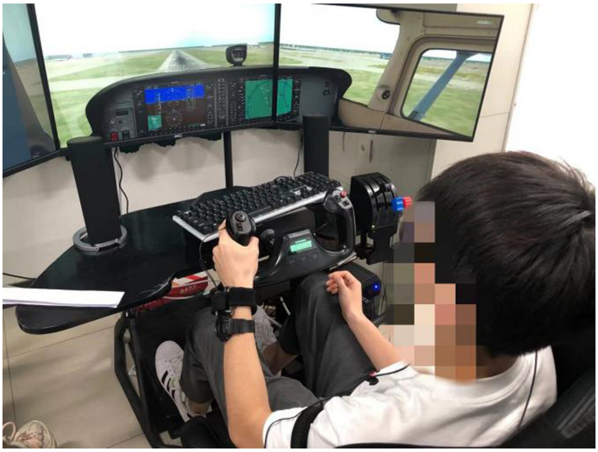


Fig. 1. Participant using the flight simulator

2.3 Experimental Setting

In this experiment, Cessna SP-G1000 was chosen as the aircraft. The experiment consisted in completing one flight scenarios with three different tasks in two conditions(with or without emotional stressor) respectively in order to assess the effects of emotional stressor on flight errors in three different performance level respectively.

Table 1. Description of flight error variables

Variable	Indicator	Unit	Description
Skill-based error	Number of Heading Deviation (NHD)	count	The number of deviation greater than or equal to 10° from the required heading
	Heading Deviation Time (HDT)	s	Total time spent in heading deviation
Rule-based error	Number of Altitude Deviation (NAD)	count	The number of deviation greater than or equal to 100 ft from the required altitude
Problem-based error	Number of Calculation Deviation (NCD)	count	The number of calculation different from the model answer
	Calculation Time (CT)	s	Total time spent in calculation

Flight Scenario. The traffic pattern lasting approximately 8 min from take-off to landing was chosen as the flight scenario. The required heading of these five legs of the traffic pattern were 342°, 252°, 162°, 72°, and 342°. During the flight scenario participants were required to complete three tasks designed to measure the number of errors in three different performance levels (skill-based level, rule-based level and problem-based level). At the first task, participants were asked to fly the airplane in required heading of each leg and any deviation greater than or equal to 10° from the required heading was counted as an error. What's more, the total time participants spent in such deviation was also recorded. For skilled pilots the cognitive process of flying the aircraft at correct heading is automated, so this task was designed to measure skill-based error. The second task required participants to report their altitude when they first reached 300 ft in the departure leg, 1000 ft in the cross-wind leg and 300 ft in the final approach, and any deviation greater than or equal to 100 ft from the required altitude was counted as an error, which can be described as rule-based because the stored rule that has been acquired through training or experience is applied: if (this situation occurs) then (do these actions). In the third task, participants needed to calculate five mathematical exercises given by our staff such as $45 + 37$ and after mentally calculating the answer, spoken out it loudly, and any deviation from the model answer was counted as an error. Besides the total time participants spent in answering these five mathematical exercises was also recorded. The difficulty of all calculations was approximately equal, which were presented to simulate problem-based tasks [9]. The detailed description of three types of errors can be seen in Table 1. All participants had two chances to fly in this flight scenario with the same tasks to ensure they were familiar with them before the formal test began.

Emotional Manipulation. We chose to manipulate emotional stressor using the similar method depicted in the study of Allsop and Gray [27, 28]. During the first simulated flight scenario considered as the low arousal (LA) condition, the participants were left alone in the flight simulator. When this scenario was completed, the participant was informed that a selection for the coming flight training would accord to the simulated flight performance and they could have one more chance to fly and the better performance would be recorded for the selection assessment, which is considered as the high arousal (HA) condition.

2.4 Physiological Measures of Emotional Arousal Level

As arousal level is regulated by activities of ANS, we chose SCL and pupil diameter as the complementation of HVR and HR to indicate arousal level. HRV and HR indicators was measured by Photoplethysmograph (PPG), a type of physiological sensor with a small size (43 * 25 * 12 mm) and a sampling frequency of 64 Hz, and the transmitter can wirelessly transfer the original data to computer. Meanwhile EDA was attained by EDA physiological sensor with the same size and sampling frequency as PPG. In addition, Tobii Glasses 2 eye-tracker including a glasses eye-tracker, a calibration card and a power equipment, was used to measure pupil diameter. The detailed description of indicators of emotional arousal level can be seen in Table 2.

Table 2. Description of indicators of emotional arousal level

Variable	Indicator	Unit	Description
HR	Mean HR	bpm	Heart rate corresponds to the mean number of beats per minute
HRV	Mean IBI	ms	Interbeat interval per minute
	PNN50	%	Proportion of n-n intervals which differ more than 50 ms with respect to the adjacent previous n-n interval
	LF(nu)	-	The normalized low frequency component of HRV
EDA	SCL	µs	Skin conductance level of EDA
Pupil size	Average pupil diameter	mm	The average diameter of left and right pupils

2.5 Statistical Analysis

All statistical analyses were carried out using SPSS 25.0. At first normality was checked for the indicator value distributions by Kolmogorov - Smirnov test. Then paired t test or Wilcoxon signed rank test was used to judge differences in indicators of arousal level and flight error between two conditions. Finally the regression analysis was used to examine the relationship between arousal level and flight errors.

3 Result

3.1 The Effect of Emotional Stressor on Arousal Level

Paired *t* test or Wilcoxon signed rank test was used to to test the differences in the value of physiological parameter representing arousal level between the two conditions according to their distribution. Compared with LA condition participants in HA conditions showed higher value in mean HR, and lower value in mean IBI and PNN50($p < 0.05$), and LF (nu) were also higher than LA condition, but its significance was just marginal (see Table 3). However the differences in SCL and pupil diameter between two conditions was not significant.

Table 3. The difference in HR and HRV between LA and HA conditions

Variable	Indicator	LA condition		HA condition		Paired *t* test		
		M	SD	M	SD	*t*	*d*	*p*
HR	Mean HR (bpm)	84.875	9.688	88.000	9.647	– 2.248	15	**0.040**
HRV	Mean IBI (ms)	716.171	80.862	689.339	72.233	2.599	15	**0.020**
	PNN50 (%)	12.229	11.204	9.236	10.064	2.821	15	**0.013**
	LF(nu)	0.586	0.132	0.638	0.135	– 1.812	15	0.090

3.2 The Effect of Emotional Stressor on Flight Errors

As to errors, Wilcoxon signed rank test or paired *t* test was used to test the difference in three different types of errors between two conditions. The results showed the skill-based error and rule-based error were reduced significantly ($p < 0.05$) (see Table 4).

Table 4. The difference in flight errors between LA and HA conditions

Variable	Indicator	LA condition		HA condition		Wilcoxon signed rank test	
		M	SD	M	SD	*z*	*p*
Skill-based error	NHD	4.188	3.146	1.813	1.721	– 2.365	**0.018**
	HDT (s)	33.230	35.652	10.543	11.891	– 2.430	**0.015**
Rule-based error	NAD	1.188	1.109	0.625	0.619	– 1.979	**0.048**

3.3 The Relationship Between Arousal Level and Flight Errors

Finally the regression analysis for the HRV variables was completed for the both conditions separately, only the HA condition revealed a significant quadratic trend for low frequency normalized units and NCD ($p < 0.05$) (see Table 5 and Fig. 2), and a quadratic trend for PNN50 and the CT ($p = 0.064$) (see Table 6 and Fig. 3). Please note the Y-axes of these two figures are inverted in order to demonstrate the inverted-U relationship.

Table 5. The quadratic trend for LF(nu) and NCD in HA condition

	Model abstract			Parameter value evaluation		
	R^2	F	P	*Constant*	b_1	b_2
LF (nu)	0.418	4.673	**0.030**	7.572	-25.642	22.342

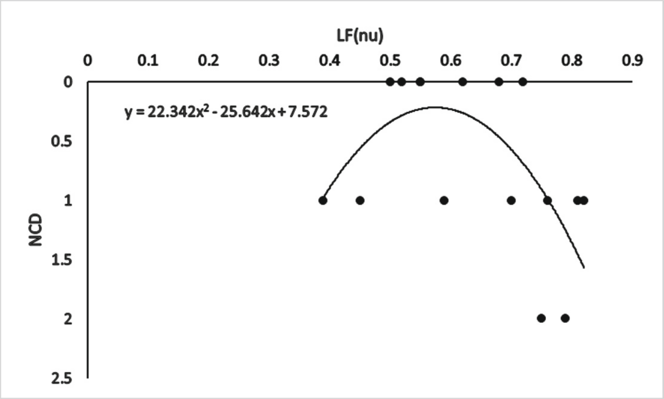


Fig. 2. The quadratic trend for LF(nu) and NCD in HA condition

Table 6. The quadratic trend for PNN50 and CT in HA condition

	Model abstract			Parameter value evaluation		
	R^2	F	P	*Constant*	b_1	b_2
PNN50	0.345	3.418	0.064	35.460	-1.564	0.039

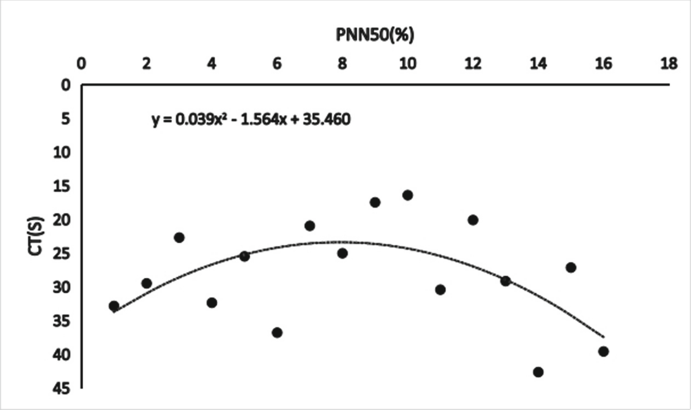


Fig. 3. The quadratic trend for PNN50 and CT in HA condition

4 Discussion

This study examined the effect of emotional stressor on human errors and the level of emotional arousal in flight. Moreover to investigate the relationship between arousal level and human errors. The participants were informed to perform the traffic pattern from take-off to landing, during which they need to complete three flight tasks belonging to different performance levels. This phrase was considered as LA condition. Participants were then transferred to HA condition where they were asked to perform the same traffic pattern and tasks. Lots of findings emerged from this study. First emotional stressor caused an increase in emotional arousal level. Second emotional stressor caused a decrease in human errors. Finally, there was a quadratic trend for LF(nu) and NCD. These findings are explained in details below.

Physiological measures indicated that emotional stress was successfully induced by the manipulation employed in the current test. Participants had significantly higher mean HR and lower mean IBI in HA condition, compared with LA condition. The same pattern as mean IBI was also found in PNN50. The participants' LF (nu) in HA condition were also higher than LA condition, but its significance was just marginal. Unfortunately the increase in SCL and pupil diameter in HA condition is not significant, compared with LA condition. These findings suggested that HA condition led to larger ANS response. As mentioned above, the fluctuation of emotional arousal regulated by the ANS can be indicated by HR and HRV [13] so the difference in these physiological indicators indicated that the level of emotional arousal in HA condition is higher. The increase in HR is in line with some previous studies with similar emotional stressor manipulated in our current test, suggesting anxiety or the emotional stress was induced and the level of arousal increased [19, 27, 28]. However these three previous studies did not measure HRV indicators or the difference in HRV indicators between two conditions was not significant, so the result in our studies can provide a new way of examining the effect of emotional stressor on arousal.

There was a significant reduce in human's skill-based errors and rule-based errors in HA condition, compared with LA condition. This finding supported the work of

Maymand [12] that emotional stressor can influence arousal level and improve human attention level, thus improving flight performance. Gray et al.[28] found participants with high scores on attention deviation showed a reduce in root mean square error of the vertical deviations of the aircraft from the ideal glide-slope in the anxiety phase, emphasizing the importance of attention in improving flight performance. According to cue-utilization theory [25, 29], both under-arousal and over-arousal would have a negative influence on attention and then cause an impairment in performance, but a moderate level of arousal was associated with better performance. So we can concluded that the arousal level in HA condition was more suitable for human's skill-based tasks and rule-based task, and therefore leaded to a reduce in errors in these two tasks compared with LA condition.

Moreover, an inverted-U relationship between NCD and LF(nu), the indicator of arousal level, was observed in HA condition. The same pattern is also found between CT and PNN50 although their relationship only reached significant level marginally ($p = 0.064$). In other two tasks there wasn't such a relationship. According to Maymand [12], physically manipulating the flight controls, the skilled-based behavior, is a simple task, meanwhile making logical decision(including both rule-based and problem-based tasks) is considered as a difficult task. But unlike problem-solving exercises that draw on high levels of cognitive resources, for rule-based tasks participants can resort to stored rules and just need retain ample mental capacity to comply with the relevant rules and procedures [9], which means problem-based task is more difficult than rule-based task. As mentioned in Yerkes-Dodson law, compared to difficult tasks, in simple tasks performers have a higher threshold for arousal. So while a high arousal level can be advantageous in simple tasks, for difficult tasks it can cause a decrease in performance [23, 24], which can provide a plausible explain the appearance of inverted-U curve in HA condition and the non-significance of the decrease in human's problem-based errors between HA and LA condition.

5 Conclusion

The presented study investigate the effect of emotional stressor on human errors of three different performance levels and the level of emotional arousal. Moreover to examine the relationship between arousal level and human errors in flight. Emotional stressor can cause an increase in emotional arousal level represented by HR and HRV indicators and a decrease in human's skill-based and rule-based errors. Besides there was a quadratic trend for LF(nu) and NCD, a dimension of problem-based error, which provided a plausible explain for the non-significance of the decrease in problem-based errors between HA and LA condition combined with inverted U-shaped curve. These findings have potential applications, as they suggest that it is possible to identify emotional arousal level during operational activity, via changes in HR and HRV indicators, and there is a correlation between emotional arousal level between human error. So the inclusion of physiological sensors such as PPG into a cockpit could be a way to potentially prevent human errors before they occur.

Acknowledgments. We appreciate the support this work received from the National Natural Science Foundation of China (grant no. 32071063).

References

1. Civil Aviation Administration of China: China aviation safety annual report 2019: CAAC, Beijing (2020)
2. International Air Transport Association. Safety Report 2019. IATA, Montreal (2020)
3. Shappell, S.A., Wiegmann, D.A.: The human factors analysis and classification system--HFACS. FAA, Washington, DC (2000)
4. Norman, D.A.: Categorization of action slips. Psychol. Rev. **88**(1), 1–15 (1981)
5. Rasmussen, J.: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Trans. Syst. Man Cybern. **3**, 257–266 (1983)
6. Reason, J.: Human Error. Cambridge University Press, London (1990)
7. Wiegmann, D.A., Shappell, S.A.: A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System. Routledge , London and New York (2017)
8. Landman, A., Groen, E.L., Van Paassen, M.M., Bronkhorst, A.W., Mulder, M.: The influence of surprise on upset recovery performance in airline pilots. Int. J. Aerospace Psychol. **27**(1–2), 2–14 (2017)
9. Morris, C.H., Leung, Y.K.: Pilot mental workload: how well do pilots really perform? Ergonomics **49**(15), 1581–1596 (2006)
10. Visciola, M., Armando, A., Bagnara, S.: Communication patterns and errors in flight simulation. Reliab. Eng. Syst. Saf. **36**(3), 253–259 (1992)
11. Bloguţ, A.: Stressing factors in aviation. Sci. Res. Educ. Air Force-AFASES **1**, 165–170 (2015)
12. Maymand, M.M., Shakhsian, F., Hosseiny, F.S.: The effect of stress on flight performance. World Appl. Sci. J. **19**(10), 1381–1387 (2012)
13. Dijkhuis, R.R., Ziermans, T., van Rijn, S., Staal, W., Swaab, H.: Emotional arousal during social stress in young adults with autism: insights from heart rate, heart rate variability and self-report. J. Autism Dev. Disord. **49**(6), 2524–2535 (2019). https://doi.org/10.1007/s10803-019-04000-5
14. Zantinge, G., van Rijn, S., Stockmann, L., Swaab, H.: Physiological arousal and emotion regulation strategies in young children with autism spectrum disorders. J. Autism Dev. Disord. **47**(9), 2648–2657 (2017)
15. Wang, C.A., Baird, T., Huang, J., Coutinho, J.D., Brien, D.C., Munoz, D.P.: Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. Front. Neurol. **9**(1029), 1–13 (2018)
16. Jerčić, P., Sennersten, C., Lindley, C.: Modeling cognitive load and physiological arousal through pupil diameter and heart rate. Multimedia Tools Appl. **79**(5–6), 3145–3159 (2018). https://doi.org/10.1007/s11042-018-6518-z
17. Vine, S.J., Uiga, L., Lavric, A., Moore, L.J., Tsaneva-Atanasova, K., Wilson, M.R.: Individual reactions to stress predict performance during a critical aviation incident. Anxiety Stress Coping **28**(4), 467–477 (2015)
18. Sexton, J.B., Thomas, E.J., Helmreich, R.L.: Error, stress, and teamwork in medicine and aviation: cross sectional surveys. BMJ **320**(7237), 745–749 (2000)
19. Hidalgo-Muñoz, A.R., Mouratille, D., Matton, N., Causse, M., Rouillard, Y., El-Yagoubi, R.: Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation. Int. J. Psychophysiol. **128**, 62–69 (2018)
20. Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., Wickens, C.D.: Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: a behavioral and eye tracking study. Int. J. Aerospace Psychol. **27**(1–2), 15–28 (2017)

21. Hebb, D.O.: Drives and the CNS (conceptual nervous system). Psychol. Rev. **62**(4), 243–255 (1955)
22. Takano, K.I., Nagasaka, A., Yoshino, K.: Experiment on validity of skin resistance level as an index of arousal level. Sangyo igaku. Japanese J. Ind. Health **35**(4), 257–268 (1993)
23. Yerkes, R.M., Dodson, J.D.: The relation of strength of stimulus to rapidity of habit-formation. Punishment: Issues Experiments 27–41 (1908)
24. Schaaff, K., Adam, M.T.: Measuring emotional arousal for online applications: evaluation of ultra-short term heart rate variability measures. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 362–368. IEEE (2013)
25. Ünal, A.B., de Waard, D., Epstude, K., Steg, L.: Driving with music: effects on arousal and performance. Transport. Res. F: Traffic Psychol. Behav. **21**, 52–65 (2013)
26. Murray, N.P., Russoniello, C.: Acute physical activity on cognitive function: a heart rate variability examination. Appl. Psychophysiol. Biofeedback **37**(4), 219–227 (2012)
27. Allsop, J., Gray, R.: Flying under pressure: effects of anxiety on attention and gaze behavior in aviation. J. Appl. Res. Mem. Cogn. **3**(2), 63–71 (2014)
28. Gray, R., Gaska, J., Winterbottom, M.: Relationship between sustained, orientated, divided, and selective attention and simulated aviation performance: training & pressure effects. J. Appl. Res. Mem. Cogn. **5**(1), 34–42 (2016)
29. Easterbrook, J.A.: The effect of emotion on cue utilization and the organization of behavior. Psychol. Rev. **66**(3), 183–201 (1959)

Multiple Physiological Indexes Analysis in Commercial Flight Tasks for Fatigue Evaluation

Xi Yang, Zhen Wang(✉), and Shan Fu

Department of Automation, Shanghai Jiao Tong University, Shanghai 201100, China
{yangxisc,b2wz}@sjtu.edu.cn

Abstract. To evaluate the flight fatigue induced by flight time and task operations, physiological indexes correlated with fatigue are selected and a comprehensive fatigue evaluation methodology is established by comprehensive analysis of multiple physiological indexes. The influence of workload on physiological data is avoided through subjective assessment of workload in each task. Every experiment is implemented with 14 training commercial flight tasks on a CRJ-200 cockpit simulator (Level D) to obtain physiological data of the pilots. Observation analysis and multiple linear regression analysis are applied to find out the relationships between fatigue level and physiological indexes. The achievements from this work could help to improve the efficiency of flight fatigue evaluation and contribute to real-time fatigue detection study in the future.

Keywords: Fatigue evaluation · Physiological index · Workload level · Observation analysis · Comparative analysis

1 Introduction

Benefited from the developments of automatic control systems, workload of pilots during flight decreases year by year. However, the unfortunate crash accidents of Boeing 737 Max tell the world that pilots' control authority is still the most important assurance of flight safety at present. Since pilots play a vital role in the flight, they must maintain sober and have rapid reactions. However, due to the complex flight control operations and long-time flight duration, the appearance of fatigue is inevitable.

To avoid human errors caused by fatigue, evaluation of flight fatigue is a primary task. Fatigue evaluation methods are generally divided into subjective evaluation and objective evaluation. However, low concentration and slow response which could cause human errors are usually ignored in subjective evaluations. Thus, objective evaluation is more comprehensive, accurate and reliable.

Researches on objective fatigue evaluation have made great progress in recent years, most of them use electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG) and eye tracker to obtain the physiological data and eye movement data for analysis.

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 91–100, 2021.
https://doi.org/10.1007/978-3-030-77932-0_8

Niels Egelund [1] found a significant relationship between 0.1 Hz HRV and long driving. Patel M. et al. [2] verified that spectral analysis of heart rate variability (HRV) data from ECG can be as an indicator of driver fatigue. Hiago M. et al. [3] investigated time-on-task effects on HRV and its relationship with self-reported and cognitive performance. Lal et al. [4] proposed that EEG is strongly related with driving fatigue, Wang et al. [5] set up an efficient means based on wireless EEG using power spectral density (PSD) with sample entropy to detect driving fatigue. Liu et al. [6] utilized eye movement indices to objectively assess the changes of attention allocation and workload. Carolina D.P. et al. [7] employed saccadic velocity to detect fatigue in military. Lisa C Thomas, et al. [8] constructed a statistical/machine learning model to accurately categorize fatigued data for pilots using pilots' combined sensor data with a success rate greater than 95%.

Most of the researches above use physiological indexes to evaluate fatigue, but rarely distinguish whether the changes are caused by fatigue or workload. Furthermore, when subjects' fatigue is induced by working on uncorrelated tasks or sleep deprivation, it is difficult to accurately simulate the fatigue under real conditions.

Hence, in our study, we propose a new method for index analysis in commercial flight tasks for fatigue evaluation, in attempt to compute and analyze fatigue via control the variable of workload level. We assume that fatigue is accumulated linearly with time. Then under the same workload level, we utilize the observation analysis and multiple linear regression method to carry out comprehensive analysis of multiple physiological indexes.

A simulated flight experiment containing 14 consecutive tasks was conducted to induce fatigue. During the experiment, multiple physiological parameters were measured. Tasks are classified under same workload levels according to the subjective assessment. The aim is to distinguish physiological changes caused by real fatigue rather than workload caused, and analyze the changes of same workload level in time order to discover the index change tendencies with fatigue and to establish a comprehensive fatigue evaluation methodology.

2 Experiment

2.1 Subjects

Subjects are 17 civil aviation pilots, healthy with normal eyesight. The pilots are divided into 11 crews, the average age is 39.1(±7.5) and average flying experience is 7173.2(±5113.6) h. All the pilots are well trained and familiar with the flight control operations.

2.2 Equipment

Experiment is conducted on a CRJ-200 Flight Simulator (Level D). Data acquisition equipment include 2 pairs of Tobii Glasses eye trackers developed by Tobii Technology AB, 2 Zephyr Bioharness multi-channel physiological measure systems and a Logitech Camera. Eye movement data are obtained from eye trackers. Heart rate and respiratory rate are obtained from the multi-channel physiological measure systems. The whole experiment is recorded by the camera.

2.3 Tasks and Procedures

Tasks. There are 14 commercial flight training tasks for each crew to conduct, such as Standard Instrumental Departure (SID), Standard Instrumental Arrival (STAR), Engine Fire, Fuel Imbalance, etc. For most crews, task orders are the same. And for the rest, task orders are designed to make differences. Each task lasts for about 10 min to 20 min, total experiment time is about 4 h, and a long rest is in the middle and divides the experiment into 2 halves.

Procedures. Before the experiment day, pilots are asked to have a good sleep without coffee or tea. Before experiment starts, pilots put on all the data acquisition equipment, eye trackers are adjusted. Pilots fill a NASA TLX questionnaire soon after each task is completed.

3 Data Analysis

The fundamental hypothesis in our study is that fatigue accumulates with flight time and tasks, so that physiological data can be statistically analyzed in time order. Data obtained in the experiment are preprocessed, and the following features are extracted for analysis, mainly include heart rate (HR), pupil diameter (PD), respiratory rate (RR) and their mean values of each task.

The data analysis includes observation analysis of task-time index change tendencies under workload levels, and multiple linear regression analysis.

3.1 Workload Level Determination

This research focuses on analyzing the effect of fatigue on physiological indexes, the interference of workload should be avoided in the analysis. Therefore, NASA Task Load Index (TLX) is applied to calculate the total workload of each task so that physiological data can be analyzed only associated with fatigue.

There are 6 factors in the TLX, which are mental demand, physical demand, temporal demand, performance, effort and frustration. Pilot chooses a more important factor for one task from a pair of factors, there are 15 pairs for each task. The pilot also gives each factor a score from 0 to 100. Total workload for one task is calculated as below, in which n_i means the times of i th factor is chosen and $Score_i$ means the score of i th factor.

$$Workload_{task} = \sum\nolimits_{i=1}^{6} Score_i \times \frac{n_i}{15}. \tag{1}$$

Five workload levels are defined as below.

$$workload_{min} + (Level - 1) \times \frac{workload_{max} - workload_{min}}{5} \leq workload \leq workload_{min} + Level \times \frac{workload_{max} - workload_{min}}{5} (Level = 1, 2, 3, 4, 5) \tag{2}$$

Then tasks are classified into the corresponding levels according to the total workload. Since subjects have bias in their own subjective assessment, there is no correlation between different subjects' task workloads.

3.2 Observation Analysis on Multiple Physiological Indexes

After the workload levels are defined, multiple physiological indexes are analyzed at same workload levels in time order to explore the change tendencies without the interference of workload. Due to total workload is calculated with tasks, so that physiological indexes are analyzed in form of mean value in task time. Several observation analyses are shown as below. Pilot B and Pilot C in Crew 4 are more experienced with more than 10000 flying hours. The flying hours of Pilot A and Pilot D are both less than 1000 and Pilot E is less than 3200, these three pilots are all less experienced.

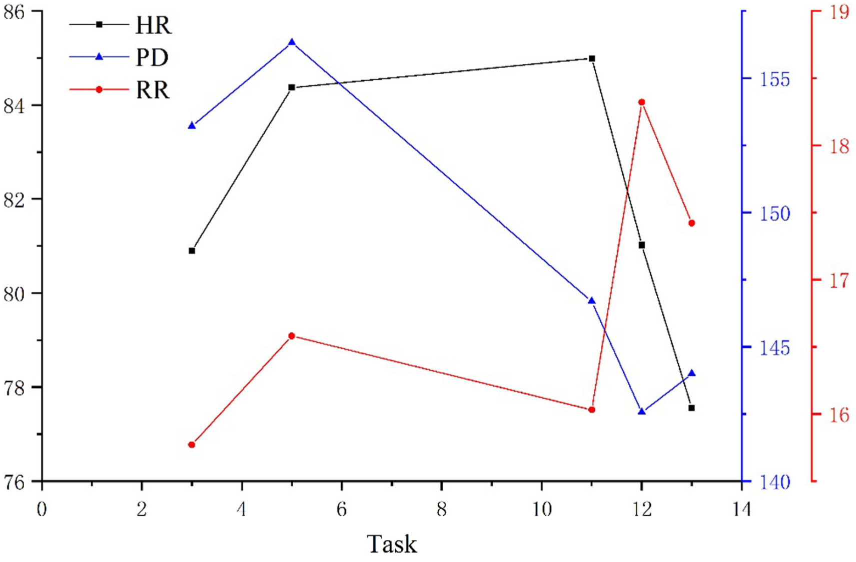


Fig. 1. Workload level 1 of Pilot A from Crew 8

Pilot A. In Fig. 1, there are five tasks in workload level 1 of Pilot A, first two tasks are from first half of the experiment and last three tasks are from the second half. All the three indexes increase in the first two tasks. Mean HR decreases in the last three tasks, and in the last task, mean HR is the lowest in five tasks. Mean PDs in the last three tasks show a rise-fall tendency and are obviously lower in values than those of first 2 tasks. Mean RRs in the last three tasks show a fall-rise tendency.

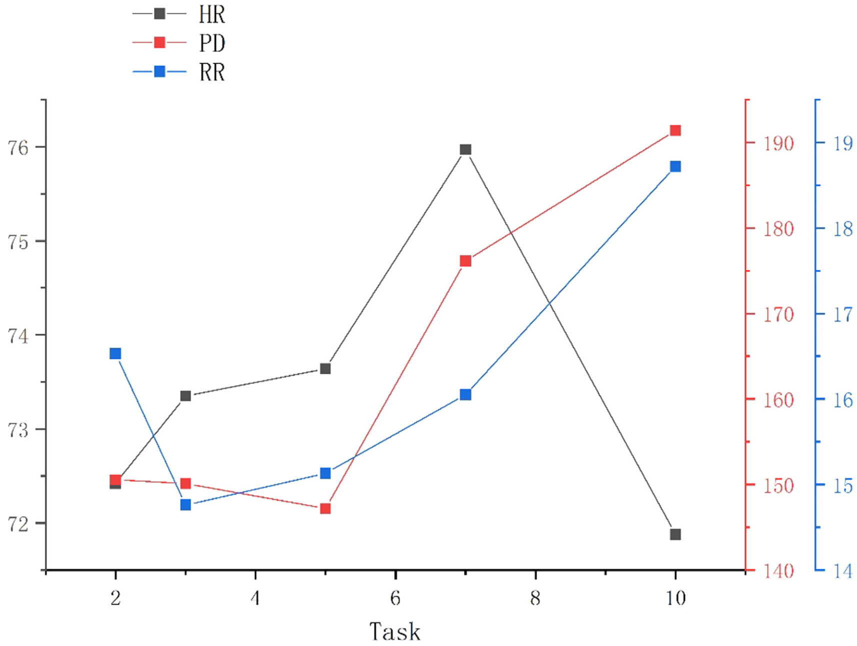


Fig. 2. Workload level 1 of Pilot B from Crew 4

Pilot B. In Fig. 2, there are five tasks in workload level 1, all the tasks are in the first half of the experiment. Mean HR changes in a rise-fall type, and in the last task, the mean HR is the lowest. There is a general trend of increase in mean PD, the values in first three tasks almost equal to each other and are much lower than those in the last two tasks. Mean RR shows a general rise up tendency except for the first task.

Pilot C. In Fig. 3, there are five tasks in workload level 3, the mean heart rate of Pilot C decreases in the first half experiment and increases in second half. Mean pupil diameter changes in the opposite way with mean heart rate. Mean respiratory rate shows a general decline tendency and becomes lowest in the last task.

Pilot D. In Fig. 4, there are five tasks from first half experiment in Level 2. All indexes obviously decline with time.

Pilot E. There are both three tasks in Level 2 and Level 3 (both are middle-stage workload levels), but three data for each index are not enough for observation analysis. Due to the largest workload deviation among the six tasks is less than 1.5 times of the distance of Level 2 and Level 3, physiological indexes of the six tasks are combined for analysis. Although there are two outliers in mean HR and mean PD, the lines in Fig. 5 still present general decline tendencies. It is obvious that mean RR declines with time.

Results of Observation Analysis. Line charts from five pilots are shown above. Through comparative analyses, mean HR of four pilots (except for Pilot C) changes

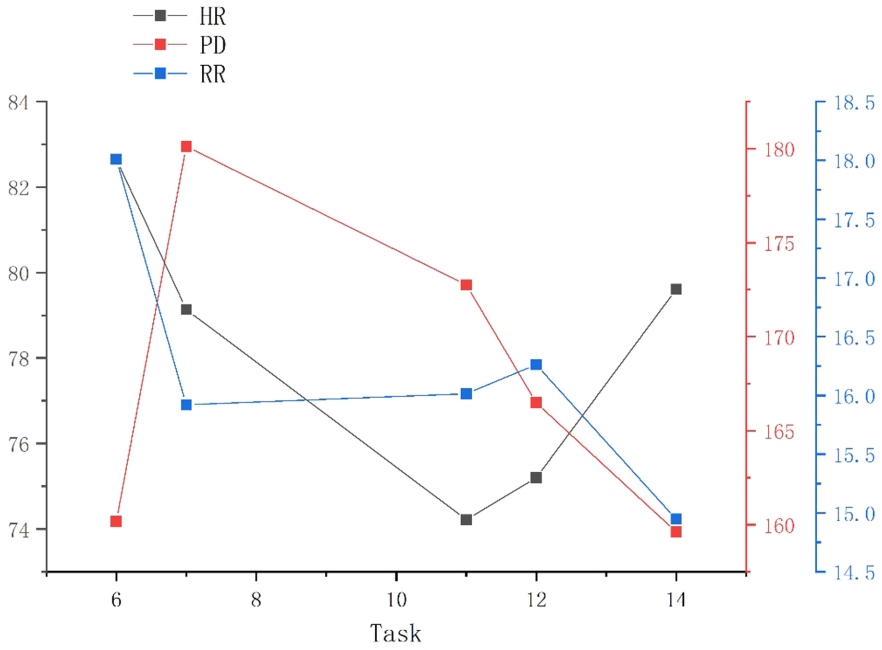


Fig. 3. Workload level 3 of Pilot C from Crew 4

in a similar general decline tendency. Mean HR of Pilot C changes in an opposite tendency compared with other four pilots. Mean RR of Pilot C, Pilot D and Pilot E, changes in a similar general decline tendency, but mean RR of Pilot A changes in a rise-fall type and that of Pilot B is in an increase trend. The situation of mean PD is almost the same as that of mean RR, except for the mean PD of Pilot A changes in a fall-rise type.

The physiological indexes of two more experienced pilots from same crew change in almost opposite tendencies. On the contrary, the physiological indexes of two less experienced pilots from same crew, Pilot D and Pilot E, change in an almost same type, all the three indexes show a decline trend with time.

3.3 Multiple Linear Regression for Fatigue Evaluation

Multiple linear regression is applied to investigate the effect of fatigue on physiological indexes and establish an equation for fatigue evaluation. According to the results of observation in Sect. 3.2, the physiological indexes of more experienced pilots change in opposite tendencies. Therefore, MLR is only applied to establish a fatigue equation for less experienced pilots, using the physiological data of middle-stage workload levels in five less experienced pilots with flying hours less than 3200 h.

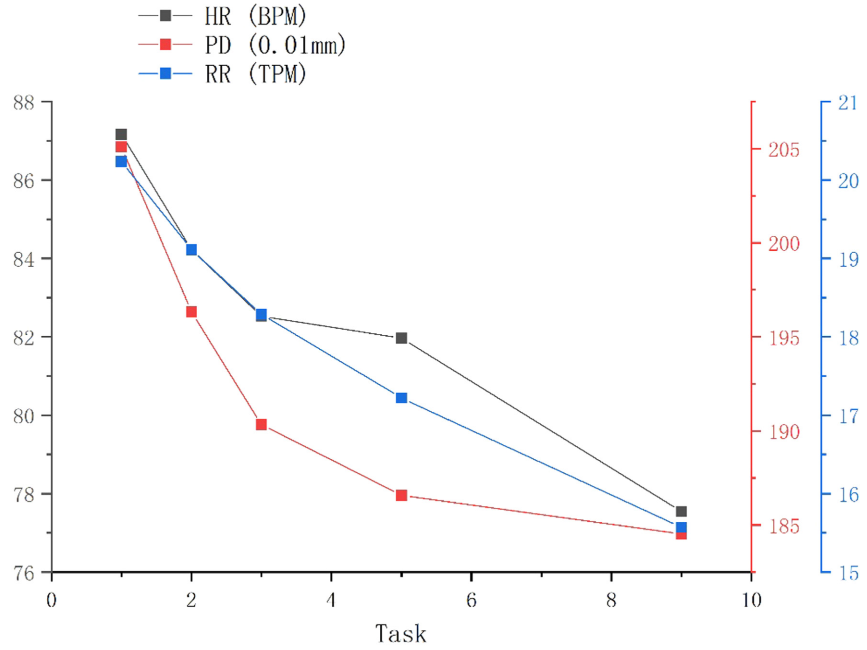


Fig. 4. Workload level 2 of Pilot D from Crew 11

In MLR, the dependent variable is fatigue and independent variables are mean HR, mean PD and mean RR. Since fatigue is assumed at the beginning of Chapter 3 to accumulate with time, it is scored linearly from 0 to 10 with task order. Moreover, because physiological indexes of pilots differ from each other, only variation values are meaningful for analysis, normalization is applied to avoid the differences in physiological base values of different pilots. Hence, normalized physiological data are utilized to build the MLR model. The MLR equation is shown as below:

$$FS = -0.983 \times HR + 0.136 \times PD - 6.693 \times RR + 8.409 \tag{3}$$

where the FS denotes fatigue score, HR, PD and RR are the normalized values of corresponding indexes. Fatigue score is calculated through this MLR equation, the value ranges from 0 to 10, which represent the fatigue levels. In the ANOVA results of MLR, $F = 39.918$, value $p < 0.001$ and $R^2 = 0.945$.

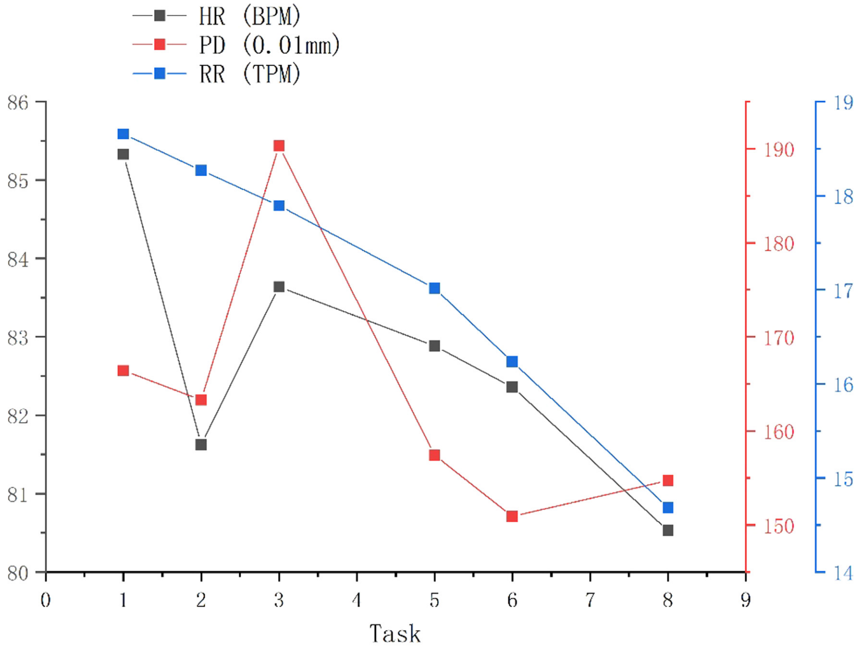


Fig. 5. Workload level 2 and Level 3 of Pilot E from Crew 11

4 Discussion

Analysis based on task workload level could explore the effect of fatigue accumulation on physiological indexes without the interference of workload. The use of mean value in analysis effectively removes the interference of instantaneous errors, but due to the data calculation time step is 10 to 20 min, short-time significant changes which can detect fatigue are also covered up. Therefore, this methodology is efficient when observing the fatigue-induced physiological indexes change tendencies in a long-time experiment, but it is not suitable for real-time fatigue detection. The most important meaning of multiple physiological indexes in fatigue evaluation is that multiple indexes can effectively avoid the inverses and omissions which could be easily caused by data errors in single physiological index situations.

The observation analyses carefully investigate objective effects of fatigue accumulation on multiple physiological indexes for each pilot. However, the results of the comparative analysis based on flying experience didn't come up with an expected result. Differences between pilots are described in detail in the observation analysis section.

Application of MLR is to establish a comprehensive equation to evaluate fatigue levels without the interference of workload. Results of MLR in this paper show that HR and RR of less experienced pilots are strongly affected by fatigue, and the PD is slightly affected by it. Young pilots usually have a better eyesight than older ones and their eyes are more difficult to get tired.

The experiment with 14 commercial flight tasks includes large amounts of complex operations, different pilots have different operation habits. Furthermore, blind operations cannot be applied during flight because of the big number of buttons and screen displays. Thus, the eye and body movements are random, the consequence of all the large amounts of random movements is that physiological indexes and their change tendencies of different pilots are usually with huge differences. Hence, the MLR equation of less experienced pilot cannot be directly applied to other pilots without adjustment.

5 Conclusions

In this paper, fatigue caused physiological index changes are analyzed without the interference of workload, observation analysis reveals the effects of fatigue on multiple physiological indexes. Comparative analysis is applied to investigate the relations of results from different pilots. Heart rate, pupil diameter and respiratory rate are chosen to establish multiple linear regression equation.

Through the observation on physiological index change tendencies in workload level, it is found that there is a general decline trend in mean HR and mean RR for most pilots during the experiment, but there is no significant change tendency in mean PD. More experienced pilots in same crew may have opposite physiological index change tendencies.

According to the MLR equation, it comes to a conclusion that there is a general decline trend in HR and RR for most less experienced pilots in the long-time experiment, and the pupil diameter tends to increase.

A methodology is established to help us evaluate fatigue in the future. First, workload levels are defined with subjective assessment. Then control the variables of workload level, change tendencies of physiological indexes can be individually observed with fatigue and avoid the interference of workload at same time. Multiple linear regression equation is established for fatigue evaluation of less experienced pilots.

References

1. Egelund, N.: Spectral analysis of heart rate variability as an indicator of driver fatigue. Ergonomics **25**(7), 663–672 (1982)
2. Patel, M., Lala, S.K.L., Kavanagha, D., et al.: Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert Syst. Appl. **38**(6), 7235–7242 (2011)
3. Melo, H.M., Nascimento, L.M., Takase, E.: Mental fatigue and heart rate variability (HRV): the time-on-task effect. Psychol. Neurosci. **10**(4), 428–436 (2017)
4. Lal, S.K., Craig, A.: Driver fatigue: electroencephalography and psychological assessment. Psychophysiology **39**(3), 313–321 (2002)
5. Wang, H., Dragomir, A., Abbasi, N.I., et al.: A novel real-time driving fatigue detection system based on wireless dry EEG. Cogn. Neurodyn. **12**(4), 36–376 (2018)
6. Zhongqi, L., Xiugan, Y., Wei, L., et al.: Analysis on eye movement indices based on simulated flight task. China Saf. Sci. J. **16**(2), 47–51 (2006)

7. Carolina, D.P., et al.: Fatigue in the military: towards a fatigue detection test based on the saccadic velocity. Physiol. Meas. **37**, 62–75 (2016)
8. Peng, C., Wang, Z., Lu, Y., Fu, S.: Research on workload-based prediction and evaluation model in power system. In: Harris, D. (ed.) HCII 2019. LNCS (LNAI), vol. 11571, pp. 92–105. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22507-0_8
9. Liu, W., Lu, Y., Huang, D., Fu, S.: An analysis of pilot's workload evaluation based on time pressure and effort. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10275, pp. 32–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58472-0_3

Electroencephalographic Signals and Pilot Situation Awareness During Simulated Flight: A Case for Enhanced Digital Technology in General Aviation

Alexia Ziccardi(✉), Kathleen Van Benthem, and Chris M. Herdman

Carleton University, Ottawa, ON K1S 5B6, Canada
alexiaziccardi@cmail.carleton.ca, {kathyvanbenthem, chrisherdman}@carleton.ca

Abstract. The present research reports on the effects of age and situation awareness (SA) on the neural processing of auditory tone stimuli presented during a simulated flight experiment to understand how pilots integrate auditory information into their mental models. Understanding the neuro-cognitive processes involved in transforming physical auditory stimuli into cognitive representations is important in the study of aviation psychology as a great deal of information about the environment that pilots receive and use to manage their flight is auditory (e.g., radio communication). Electroencephalogram (EEG) data was collected while 51 pilots conducted a one-hour flight in a Cessna 172 full-scale simulator and were presented with auditory tones. Markers were simultaneously added to the EEG data to reflect the onset of each tone and the pilot button response they were tasked with. Grand average event-related potentials (ERPs) related to the auditory tones compared neural responses for both the older (51+ years) and younger (<51 years) pilot groups. The relationship of age to accuracy of SA models was also investigated by indexing pilot mental models of the relevant environment. Findings showed that auditory information is not always well-integrated into SA models, and this was particularly true for older pilots. Furthermore, changes in how auditory information is processed in the brain may contribute to the negative age-effects seen in pilot SA. This research is important in informing efforts to enhance safety in general aviation. Effective strategies to improve pilot SA may include visual technologies to augment the auditory information available to pilots.

Keywords: Linguistics · General aviation · Human-computer interaction · Neuroscience · Cognition · Aging · General aviation

1 Introduction and Background

1.1 Background

Older age has been associated with higher risk of critical incidents in the general aviation population [1]. One plausible explanation for higher rates of accidents found among

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 101–115, 2021.
https://doi.org/10.1007/978-3-030-77932-0_9

older pilots is related to the normal declines in cognition seen in older age. For example, both standardized cognitive tests, as well as flight simulation research, has shown that older pilot cognition is subject to the same negative effects of age, as is found in the general population [2]. That is, unlike what may be speculated about this population, pilot experience and expertise may not offer protection against normal age-related declines in cognition [3, 4]. In aviation psychology research, basic cognitive functions show declines in older pilots, including information processing speed, attention, executive functioning, and memory, [2, 5, 6]. In a three-year longitudinal study by Taylor et. al (2007), older pilots performed worse on almost all performance tasks in simulator environments than did their younger counterparts (e.g., traffic avoidance, communication, and monitoring instruments) [7]. Furthermore, situation awareness (SA), may also be negatively impacted by older age [8].

SA is one of the key domain-specific cognitive abilities that pilots utilize during flight. As defined by Endsley (1998), SA is the ability to perceive elements in the environment and use this information to further project and plan for action in the future [9]. Air-to-air radio communication is a critical means by which pilots maintain up-to-date SA models and maintain safety during flight. In general aviation, SA requires the integration of auditory information via radio communication messages. Specifically, pilots use auditory information to build a mental operating picture of the position and characteristics of other aircraft in the relevant airspace. In light of the increased risk of critical incidents in older pilots, advancing our understanding of the effects of normal aging in aviators, and in particular, on SA, is critical to pilot safety [10, 11].

Situation Awareness. To avoid critical incidents, pilots must frequently update their mental models of the environment they are flying in and project for future action. SA drives decision-making and in turn is considered a crucial aspect of safe outcomes for pilots. Endsley (1997) outlined a three-level model of SA, where the first level involves the perception of elements in the environment [12]. Specifically, a pilot should be able to recognize items in their environment and distinguish their relevant features, including their status, attributes, and dynamics [12]. Pilots should also be aware of the properties of the aircraft or tools that they are using [12]. Visual and auditory information is prevalent at this level [12]. Level-1 SA is essential for building the second and third levels of SA, which involve the comprehension of the elements and dynamics of the environment and how to project these elements into future activity [12].

The effects of aging on SA have been examined in both the general population and the pilot population. Using a driving simulator experiment, Bolstad (2001) found that older adults performed worse on SA tasks while driving cars than did younger adults [13]. In a study examining change detection in air traffic and instrument malfunctions, Coffrey et al. (2007) found that older pilots struggle more than their younger peers to detect unexpected critical events in a simulated environment [14]. An association between older age and lower SA performance in a simulator have also been found in the literature [8]. Although the literature has shown these common findings, a more detailed investigation of age and factors involved in SA model building is needed in the literature.

Air-to-Air Radio Communication. One of the main ways in which pilots contribute to maintaining up-to-date and accurate mental models is through the processing of aurally presented information in the form of air-to-air radio communication (ATC) messages. Pilots are highly trained to utilize ATC communication as it is a contributing factor in Level-1 SA building. Radio messages contain critical information regarding other aircraft flying simultaneously (e.g., other pilot's aircraft type, call sign, location and intention). Pilots hear this information, process it conceptually and then manipulate it to make crucial decisions accordingly. The investigation into ATC communication and risk factors associated with this flight task is therefore an integral component of understanding the factors involved in safe flying.

Previous literature in aviation psychology has found older age to be associated with the reduction of ATC communication task abilities. In one previous study, older pilots showed greater difficulty reading-back ATC messages than younger individuals [15]. As part of their longitudinal study using a flight simulator, Taylor et. al (2005) investigated age-related effects on the executions of ATC communications [16]. The findings of this study showed significant age-related effects on communication performance such that older age was associated with lower performance scores on tasks involving the execution of tasks in ATC messages [16]. Since these broadcasts are auditory, these types of findings point to the importance of further investigating the effects of age on the processing of aurally presented information in general aviation.

Auditory Processing and EEG. Neurological studies are not as common in the domain of aviation psychology when discussing the processing of aurally presented information. However, neurological techniques such as electroencephalogram (EEG) have been useful, reliable, and objective tools used to understand the process by which individuals process auditory information. One type of neural analysis that can be conducted with EEG is Event-Related Potential (ERP). ERP is a measure of the neural response or electrical activity to a specific stimulus, measured in hertz (Hz). ERPs are time-locked, meaning that they are specific to the response at a certain time that the stimulus is present at and therefore, have great temporal resolution. When analyzing ERPs, components are examined, which are systematic or characteristic deflections in the data. Investigating different measures using ERP responses allows for a comparison of brain responses depending on the variables in place.

Many studies have used this technique to investigate the processing of auditory tones, commonly in oddball paradigms or attention tasks. These types of studies are reliable means to understand auditory processing along a temporal timeline, since stimulus detection (N1) and later processing (P2) components can be easily traced. An auditory ERP can be broken up into three phases [17]. In the earlier stage, within the first 10ms, the Auditory Brainstem Response occurs [17, 18]. These are seven slight peaks, generated in the auditory cortex and brainstem, that occur when there is a discharge of synchronous auditory neurons, generating their nerve and brainstem structures [17, 19]. In the following 50ms, the middle stage, auditory cortex neural activity responses occur [17, 18]. In the later stage, functional brain activity and information processing occur, mainly in the auditory and frontal regions, as well as in other areas [17, 18]. In the later stages, there is the N1 component, commonly related to stimulus detection and the P2 and P3 components, commonly related to neural processing in auditory and frontal regions [17, 18, 20, 21].

1.2 Present Research

The objective of the present research was to investigate the effects of age on the processing of auditory tone stimuli presented during a simulated flight experiment. SA behavioural data was analyzed to provide a greater understanding of what elements contribute to poorer SA abilities in the older pilot population. Additionally, ERPs frequently presented auditory tone stimuli were utilized as simple stimuli that can be investigated using an EEG system and provide insight in the pilot auditory processing mechanisms.

An exploration into the segments of communicated messages was conducted. The radio communication analyses investigated the effects of age on the types of information used in building operational pictures of the environment. A neural analysis was conducted to investigate age effects on processing auditory information at different cognitive levels (stimulus detection or processing). Pilot performance on SA tasks were also grouped (higher versus lower performers) and a neural analysis was additionally conducted to see if low performers showed patterns in their neural responses as compared to high performers. This research is important in informing the body of research in aviation psychology in efforts to enhance safety in general aviation. Effective training strategies or performance technologies can use these results to inform their system design.

2 Methods

The research outlined in this paper uses data collected as part of a larger study investigating general aviation safety and pilot factors. The present study was approved by the university ethics committee operating under the Canadian Tri-Council Code of Ethics for psychological research.

2.1 Participants

Licensed pilots were recruited from local pilot associations, flying schools and clubs. Inclusion criteria included having a pilot permit and medical certification to fly, as well as pilot-in-command hours in the past 24-months prior to the study. The final sample included 51 pilots (4 female). Pilot age varied (M = 46.29, SD = 17.44). For analyses, the participants were divided into a younger (n = 26, aged 17–50) and an older group (n = 25, aged 51–71). Participants provided written informed consent and received refreshments and paid parking as compensation for their participation.

2.2 Briefing

The researchers guided the pilots through an introductory PowerPoint presentation in order to familiarize the participants with the study purpose, the task requirements, and the equipment that will be used, for both the practice and experimental sessions. The presentation also contained information about the planned flight route, including the two legs they would be flying in, and flight parameters, such as altitudes, headings, and airspeeds for the outbound and inbound segments of the route. Participants ran through a practice session before beginning the experiment and were encouraged to ask questions if they were unsure of any tasks or procedures.

2.3 Materials

Pilots flew in a Cessna 172 Level 6 Flight Training Device (see Fig. 1). Within the cockpit, the flight instruments included a yoke, throttle, and flaps. Participants wore an EEG headset to record data pertaining to their neurological states. The EEG headset that was used was the EMOTIV EPOC + 14 channel wireless EEG system [22]. To record the EEG data, the EMOTIV software TestBench was used, applying a bandwidth of .2 to 45 Hz before further processing in EEGLAB, an open-source software running on MATLAB v. R2020a [23].

2.4 Flight Task Design

Participants were instructed to fly the first leg of the route, adhering to the planned headings, airspeeds, and altitudes. The route began at a small general aviation aerodrome at the Gatineau airport, heading westerly along two rivers. There was then a 10-min cross country portion before reaching the Carp local general aviation aerodrome (see Fig. 2 for a sample flight path). During the flight, air-to-air radio messages were broadcasted from local pilots. Participants were instructed to listen for these messages as they would be then asked questions about their content. The pilots wore earbuds to receive the broadcasts. The messages were presented in earbuds that the pilots wore during the flight. There were 14 radio messages in total, which contained information about other aircraft flying simultaneously (e.g., other pilot's aircraft type, call sign, location and intention).

Throughout the flight 1000 Hz auditory tones were presented every four to seven seconds and pilots were told to respond to them by clicking on a button on the yoke. The tones were 100 ms in length and were played from two speakers that were placed behind the pilot. The EEG headset recorded data pertaining to these auditory tone stimuli. Throughout the recording, markers (or triggers) were simultaneously added to the EEG data to reflect the onset of each tone and the pilot button response using Psychopy 3.0. Although pilots were tasked with clicking the button when they heard the tones, they were told to maintain the "aviate, navigate, and communicate" rule as their priority and to ignore the tones if they were busy attending to other tasks.

After completing a touch-and-go at the second aerodrome, experimenters paused the flight and participants were asked to indicate where other aircrafts were located on a map and report on the details of the radio messages heard. After the queries were completed, the participants continued onto the second leg, mirroring the first. Upon completion of the second leg, participants flew one touch-and-go at the first aerodrome and the aircraft turned onto the downwind. Once on the downwind, the simulation was over, participants exited the aircraft, and the same queries were conducted. The total duration of the flight was approximately 60 min.

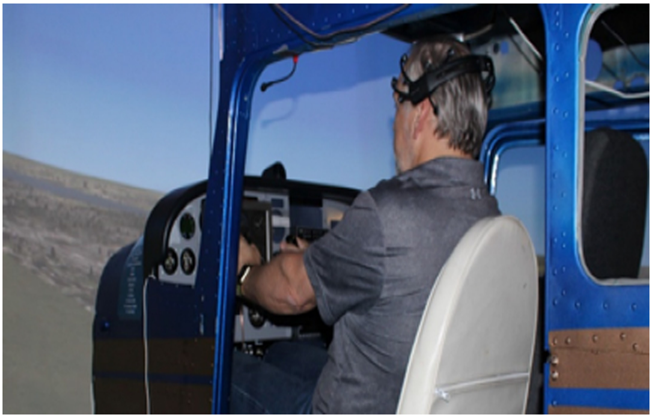


Fig. 1. Participant in the Cessna 172 simulator

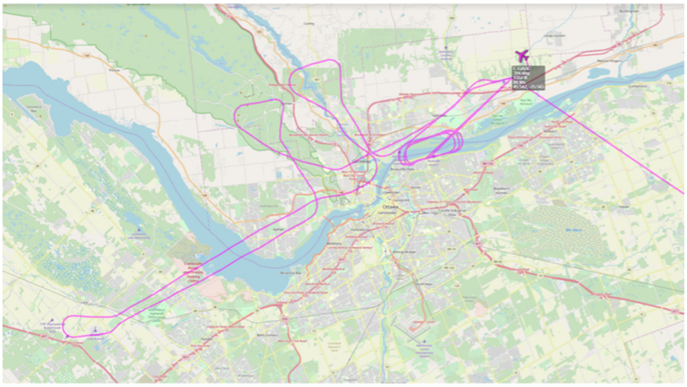


Fig. 2. Sample flight path

2.5 Behavioural Variables

Radio messages from other aircraft were separated into their common components: aircraft type, call sign, location and intention. In natural environments, pilots do not always verbalize all the components in the calls. Therefore, the message segment of greatest interest in this work was the Type segment, which was least frequently reported. Scores were based on the level of detail provided and ranged from 0 (no details of Type) to 2 (perfect recall of type of aircraft). The second index of SA of interest were responses related to pilot knowledge of the position of each aircraft at the time of the pause. Responses were marked on a map, and the Position scores were based on proximity to the actual position of the other aircraft. Scores ranged from 0 beyond a reasonable perimeter to 2 (very close to the actual aircraft position). Participants never saw the other aircraft, but instead had to build a picture of the airspace and its elements based on knowledge of aircraft characteristics and information from the aircraft radio messages.

For both indices of SA (Type and Position) the queried took place at the pause between the two segments of the flight.

2.6 EEG Variables

Electroencephalography Recording and Processing. EEG data was collected using the EMOTIV EPOC + 14 channel wireless EEG system (see Fig. 3). The EEG headset followed the international 10–20 system channel system and placements were referenced to electrodes P3 and P4. The recordings were collected at 2048 Hz, and downsampled to 256 Hz. Data was transmitted wirelessly from Bluetooth to an iMac desktop computer. Independent Component Analysis was used to clean the EEG data of noise components, such as eye, muscle or electrode movements and identify neural responses to the tones. Once flagged, components were automatically removed, and the recordings were re-referenced. A second cleaning was then performed. Following the automated pre-processing, subjective visual inspection was conducted to remove remaining high probability artifacts such as electrode "pops" or remaining blinks/lateral eye movements. The final subset of the sample was 33 participants.

Fig. 3. EMOTIV EPOC EEG System [22]

Event-Related Potential (ERP). Epochs were created based on the triggers inserted by Psychopy 3.0, indicating the onset of the tones at −2000 to + 2000 ms. The Study function in EEGLAB created a grand average of the ERPs across the final subsample of participants at key brain regions. Datasets were grouped according to age of participant (split at age of 50-years) to permit the examination of age effects on neural responses to the tones. A second analysis was conducted on SA performance groups. Using SPSS [24], Z-scores were created from the total radio message scores from both legs of the flight. The Z-scores were then split as below 0 (low scores) and 0 and above (high scores), creating two groups.

Brain Regions of Interest (ROIs). The areas of interest investigated in the analyses were auditory and frontal regions (see Fig. 4). Auditory regions (location at electrode T7 and T8) were of interest as these areas process sounds and their components. Frontal regions (location are the electrodes F3, F4, FC5 and FC6) were of interest as they relate to executive functioning following the sounds (e.g., decision making). In the ERP analysis, there was an investigation on whether there were differences in neural responses at the N1 component, related to stimulus detection, P2 and later components, related to information processing.

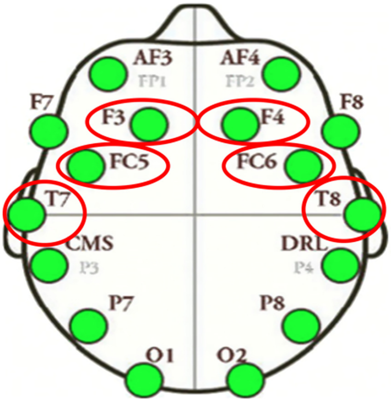


Fig. 4. EEG Electrode Location of the Brain ROIs

3 Results

The first results presented in this section are the SA performance analyses. The outcome variables for this first section are the mean SA performance scores for older and younger pilots of the radio call segment of Type. The second section will examine the neural responses to the repeated tone stimuli which were presented while the pilots were simultaneously engaged in the flight tasks. The outcome variables for these analyses will be the neural responses for both age groups as well as both performance groups, at the N1 component, related to stimulus detection, P2 and later components, related to information processing in different brain regions.

3.1 Situation Awareness Analysis

Univariate ANOVAs were conducted to investigate the effects of age on the mean SA performance scores corresponding to key SA queries conducted during the simulated flight. As shown in Fig. 5, for the first leg of the flight, there was a significant effect of age on the Type segment of the radio calls, such that older pilots were less likely to recall Type information than younger pilots, $F(1, 49) = 4.74$, $p = 0.034$, $\eta p^2 = 0.089$. This pattern was repeated for the second leg of the flight, where older pilots were again less likely to recall Type information from the radio messages, as compared to younger pilots, $F(1, 49) = 4.75$, $p = 0.034$, $\eta p^2 = 0.089$.

Univariate ANOVAs were also conducted to determine any effects of age on position information of other aircraft, another important aspect of the pilot's operational picture. There were no significant effects of age on position information in the second leg. However, in the first leg, older age was associated with poorer recall, $F(1, 49) = 4.28$, $p = 0.044$, $\eta p^2 = 0.082$.

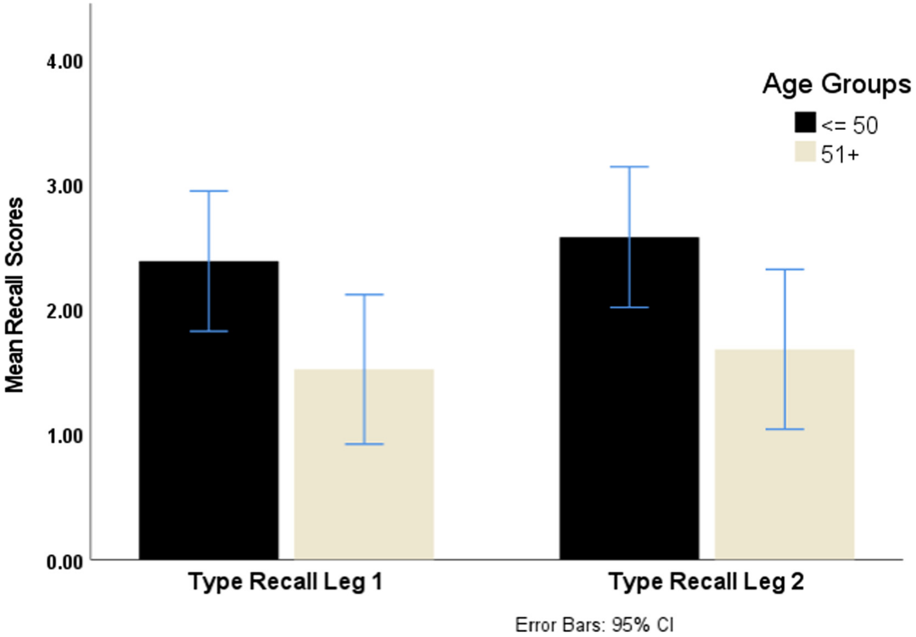


Fig. 5. Effect of age on Type segment of radio calls in both legs of flight.

3.2 Event-Related Potentials Analysis

To test for significant differences between the two groups, the EEGLAB software conducts *t*-tests at every millisecond post-stimulus for the selected epoch length.

SA and Event-Related Potentials. Illustrated in Fig. 6, is the average ERP linked to the onset of the tones for both the lower SA performing as well as the higher performing groups in a frontal region, closer to the frontal eye fields, (e.g., electrode F4). At the N1 component, there was no significant difference in the processing of the tone between the two groups. Also observed in this electrode's data, is a significant difference ($p < 0.05$) in the processing of the tone at the P200 component such that pilots who performed worse on SA tasks have greater amplitudes at this stage in the auditory processing pipeline.

Age and Event-Related Potentials. Illustrated in Fig. 7 is the average ERP linked to the onset of the tones for both the younger group as well as the older group of pilots in an electrode related to auditory processing (e.g., electrode T8). At the N1 component, there was no significant difference in the processing of the tone between the two groups. There is a significant difference ($p < 0.05$) observed in the processing of the tone after the P200 component such that older pilots have attenuated neural responses at this stage of auditory processing.

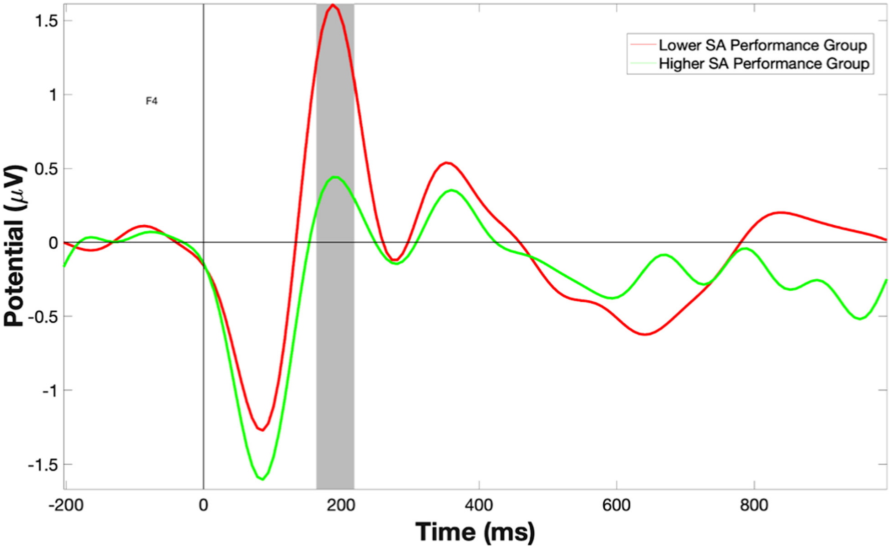


Fig. 6. SA Performance and Event-related Potentials in the Frontal Eye Fields. *Note***.** The grey bars indicate significance with a threshold of $p < 0.05$.

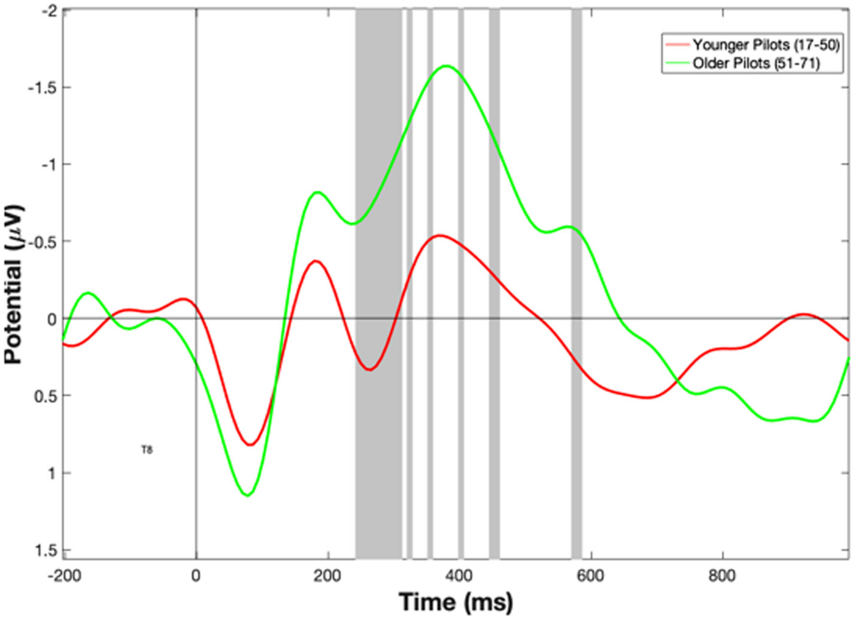


Fig. 7. Age and Event-related Potentials in the Auditory Cortex. *Note*. The grey bars indicate significance with a threshold of $p < 0.05$.

4 Discussion

The present research examined SA in general aviation through a behavioural and neural perspective. SA performance scores, collected during a simulated flight, were analyzed to investigate the age-related effects on pilots' abilities to build mental models of their environment from auditory information (e.g., ATC messages). Additionally, a neural analysis of auditory tone processing was used to examine whether an association can be made between amplitude responses to auditory stimulus between the lower and higher performing participants, as well as between older and younger age groups.

Behavioral Analysis. In the behavioural analysis, age-related effects on mean SA performance scores corresponding to key SA queries were examined. Of interest was the effect of age on mean performance on the Type segment, as this is a critical feature of ATC messages and was also the least presented information in the broadcasted messages. Significant effects of age were found on the mean performance scores for the Type message segment. Moreover, in the first leg flown, older pilots scored lower on one of the positioning tasks, where they had to indicate on a map the location of other aircraft.

These general age-effects on SA performance results support previous findings [8, 13]. The method of analysis in this study, examining the performance of message segment recall and positioning performance is uncommon in the literature. However, the deleterious effects of age on these indices of SA may be attributed to working memory (WM) ability declines, as seen in the aging population [25]. WM is needed in building SA models as new information must constantly be added to existing models in order to update the picture they have of the situation [12]. Declines in SA abilities have previously been found to be associated with declines in WM abilities [12, 26]. One assumption in the literature is that with an increase in age, there is a reduction in the capacity to retain information in the mind while simultaneously processing or integrating other information of other cognitive tasks [25]. In a flight scenario, pilots use ATC communication to integrate information into their SA models and make decisions accordingly. Perhaps, in the present scenarios, given that the segment for Type is normally at the beginning of the messages, older pilots may struggle to maintain these bits of information in their WM as they are still processing and retaining subsequent information.

Linguistic literature on speech comprehension found age-related difficulties to process complex syntactic structures as a result of poor WM abilities [27]. Although the radio messages may not be considered syntactic in the traditional sense, ATC messages follow a syntax in which pilots are trained to follow, understand and utilize during flight. The process of integrating ATC messages into semantic concepts mirrors regular sentence processing in that the meaning of the whole is dependent on the meaning of its parts [28]. If pilots are not keeping the first message segment in their WM, this can contribute to decreased SA abilities.

Another potential cause of the lower performance scores in the Type segment can be attributed to the linguistic features of the messages. Ziccardi et al. (2020) have investigated linguistic features of radio message segments and found that the aircraft Type segment had strikingly higher overall pitch than the other components [29]. There have been studies that found there to be age-related declines in auditory temporal processing and cognition for audiometrically normal hearing adults [30]. Therefore, it can be that at higher-cognitive processing of temporal information, such as pitch, older pilots may struggle more with these message segments. Future studies need to control for these factors and test message segments more in debt to better understand what is happening at this level.

Neural Analysis. In order for pilots to build SA models from radio communication messages, an extensive cognitive process occurs. In short, the message creates an acoustic wave, whose features are first processed by the auditory system and then further processed by the brain and transformed into concepts, which are then fed into other cognitive systems to be integrated into SA models [15]. At any stage of this process of converting ATC messages into mental concepts, SA can be affected. The neural analysis conducted focused on the point in which sounds are processed in the brain, right before the concepts are created and further processed into SA models.

Firstly, an investigation into how pilots process the auditory information presented during flight was conducted on lower and higher SA task performing groups. The electrodes relating to the Frontal Eye Fields showed significant effects of performance groups on neural activation. No significant differences were seen at the N1 component, however, at the P2 component, pilots who had poorer task performance presented larger P2 component amplitudes than pilots who had higher task performance. The frontal area of the brain is commonly associated with executive functions and decision making [31]. Therefore, individuals who performed better on SA tasks may have been more immersed in those flight-relevant tasks and allocated less neural resources to the auditory tone task. At the same time, the pilots who performed worse may have allocated more of their neural resources to the auditory tone task rather than the SA tasks. This is problematic as, although tone-related tasks are not usually present during flight, there are plenty of other flight-irrelevant visual or auditory cues that present themselves during such dynamic and complex flights and pilots still have to maintain up-to-date mental models of the environment they are flying in.

In the second neural analysis, an investigation into how older and younger pilots processed auditory stimuli presented during flight was conducted. For electrodes indexing activity over the auditory cortex significant effects of age on neural activation were found. No significant differences were seen for the N1 component, however, for the P2 component, older pilots presented larger component amplitudes after the P2 point than their younger peers. Previous literature of ERPs in auditory cortical regions, have associated the N1 component with the detection of sounds [20, 21, 32]. The lack of significant differences between these two age groups are not surprising as pilots should have audiometrically normal hearing and are tested for this. Therefore, the detection of sound should be similar in the auditory cortex at this component for both groups. Age effects were, however, observed in later components. The role of the P2 component and its subsequent peaks have been linked to neural activation in relation to information processing and task performance related to sound in the auditory cortical regions [18]. This result suggests that in the auditory cortex, there are age-differences in later, cognitive processing of the information related to sounds. These results support studies that show age differences in the P2 component in various auditory tasks [33] and studies that show older age to be associated with larger P2 components when presented with auditory information [34, 35]. Age differences in later processing of auditory stimulus can, therefore, negatively affect later SA model building constructed from auditory information.

5 Conclusion

This study can inform efforts to enhance pilot safety during flight. The results presented in this research suggest that age-related declines in SA awareness abilities may result from processing issues along the auditory pipeline. Results provide evidence that the cognitive processing components of messages may vary depending on what information is shared or the order in which certain components are presented. This aspect of research in ATC communication has not been extensively examined and requires further exploration to better understand the underlying mechanisms behind the phenomenon that aircraft type seems to be the parts of messages that older pilots struggle the most with.

A better understanding of these preliminary results can provide insight into creating different methods either of broadcasting information regarding other aircrafts or in training pilots to allocate those cognitive resources to this critical information. For instance, perhaps a more concise syntax can be created that would allow older participants to store all elements of messages in their WM. Or perhaps visual cues need to be present simultaneously to decrease the amount of auditory information pilots need to store while integrating into their mental models. These developments can potentially decrease the effect of older age on SA model building and in turn, decrease the greater critical incident rates in this pilot population.

Efforts in pilot communication technologies can be informed by the neural findings of this study. Lower SA performing pilots were associated with larger P2 component amplitudes in the frontal eye regions in comparison to the higher performing pilots. It is not clear whether the lower performers made the tone task a priority consciously or not, but these findings suggest that while simultaneously engaging in other flight activity, this group allocated great neural resources for executive functionning to flight-irrelevant tasks. This is an important finding as it suggests that age may not be the only contributing factor involved in poorer performance on SA tasks. Greater emphasis on flight training courses or devices (e.g., VR systems) need to be allocated to training on how to drop distractions and monitor important information. Further research into where pilots are allocating their neural resources should be conducted to further understand this finding.

Additionally, older age was associated with greater amplitudes in later components (following P2) in the auditory regions. This suggests older pilots struggle with higher-order processing of sound information. A more extensive investigation of sound features therefore, needs to be conducted to better understand what facets of sound processing are affected by age in the aviation population to create technologies that may potentially modulate certain sound features when presenting broadcasted information.

The findings in this research point to the importance of further researching the auditorily presented information along the auditory pipeline in pilots and creating alternative broadcasting technologies in order to counter the effects of aging on auditorily presented information and subsequent mental picture building.

References

1. Li, G., Baker, S., Qiang, Y., Grabowski, J., McCarthy, M.: Driving-while-intoxicated history as a risk marker for general aviation pilots. Accid. Anal. Prev. **37**, 179–184 (2005)
2. Hardy, D., Satz, P., D'Elia, L., Uchiyama, C.: Age-related group and individual differences in aircraft pilot cognition. Int. J. Aviat. Psychol. **17**, 77–90 (2007)
3. Hardy, D., Parasuraman, R.: Cognition and flight performance in older pilots. J. Exp. Psychol. Appl. **3**, 313–348 (1997)
4. Parasuraman, R., Sheridan, T., Wickens, C.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Making. **2**, 140–160 (2008)
5. Causse, M., Dehais, F., Arexis, M., Pastor, J.: Cognitive aging and flight performances in general aviation pilots. Aging Neuropsychol. Cogn. **18**, 544–561 (2011)
6. Van Benthem, K., Herdman, C.: Cognitive factors mediate the relation between age and flight path maintenance in general aviation. Aviation Psychol. Appl. Hum. Factors **6**, 81–90 (2016)
7. Taylor, J., Kennedy, Q., Noda, A., Yesavage, J.: Pilot age and expertise predict flight simulator performance: a 3-year longitudinal study. Neurology **68**, 648–654 (2007)
8. Van Benthem, K., Herdman, C.: The importance of domain-dependent cognitive factors in GA safety: Predicting critical incidents with prospective memory, situation awareness, and pilot attributes. Saf. Sci. **130**, 104892 (2020)
9. Endsley, M.: Design and evaluation for situation awareness enhancement. Proc. Hum. Factors Soc. Annual Meeting **32**, 97–101 (1988)
10. COPA Financial Reports. https://copanational.org/en/copa-financial-reports/
11. Bazargan, M., Guzhva, V.: Impact of gender, age and experience of pilots on general aviation accidents. Accid. Anal. Prev. **43**, 962–970 (2011)
12. Endsley, M.: Supporting situation awareness in aviation systems. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation (1997)
13. Bolstad, C.: Situation awareness: does it change with age? Proc. Hum. Factors Ergon. Soc. Annual Meeting. **45**, 272–276 (2001)
14. Coffey, E., Herdman, C., Brown, M., Wade, J.: Age-related changes in detecting unexpected air traffic and instrument malfunctions. In: 2007 International Symposium on Aviation Psychology (2007)
15. Morrow, D., Menard, W., Stine-Morrow, E., Teller, T., Bryant, D.: The influence of expertise and task factors on age differences in pilot communication. Psychol. Aging **16**, 31–46 (2001)
16. Taylor, J., O'Hara, R., Mumenthaler, M., Rosen, A., Yesavage, J.: Cognitive ability, expertise, and age differences in following air-traffic control instructions. J. Syst. Integr. Neurosci. **20**, 117–133 (2005)
17. Lakey, R.T., et al.: Event-related-potential (ERP) markers of traumatic brain injury (TBI) severity and cognitive function – Understanding how the brain works and thinks post TBI. J. Syst. Integr. Neurosci. 7 (2020)
18. Trainor, L.: Event related potential measures in auditory developmental research. Dev. Psychophysiol.: Theory Syst. Methods **69**, 102 (2008)
19. The Auditory Brainstem Response: Still an Important Tool for Neurodiagnostics! https://canadianaudiologist.ca/the-auditory-brainstem-response-still-an-important-tool-for-neurodiagnostics/
20. Hyde, M.: The N1 response and its applications. Audiol. Neurotol. **2**, 281–307 (1997)
21. Näätänen, R., Sams, M., Alho, K., Paavilainen, P., Reinikainen, K., Sokolov, E.: Frequency and location specificify of the human vertex N1 wave. Electroencephalogr. Clin. Neurophysiol. **69**, 523–531 (1988)

22. EMOTIV.: Emotiv Epoc+ User Manual. https://emotiv.gitbook.io/epoc-user-manual/ (2018). Accessed 10 Jan 2021
23. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods **134**, 9–21 (2004)
24. SPSS Software. https://www.ibm.com/analytics/spss-statistics-software
25. Salthouse, T.: Working memory as a processing resource in cognitive aging. Dev. Rev. **10**, 101–124 (1990)
26. Catherwood, D., et al.: Mapping brain activity during loss of situation awareness. Hum. Factors: J. Hum. Factors Ergon. Soc. **56**, 1428–1452 (2014)
27. Norman, S., Kemper, S., Kynette, D.: Adults' reading comprehension: effects of syntactic complexity and working memory. J. Gerontol. **47**, P258–P265 (1992)
28. Pelletier, F.: The principle of semantic compositionality. Topoi **13**, 11–24 (1994)
29. Ziccardi, A., Van Benthem, K., Herdman, C.M.: A language-oriented analysis of situation awareness in pilots in high-fidelity flight simulation. In: Stephanidis, C., Antona, M., Ntoa, S. (eds.) HCII 2020. CCIS, vol. 1294, pp. 639–646. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60703-6_82
30. Füllgrabe, C., Moore, B., Stone, M.: Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. Front. Aging Neurosci. **6** (2015)
31. Lara, A., Wallis, J.: The role of prefrontal cortex in working memory: a mini review. Front. Syst. Neurosci. **9** (2015)
32. Woldorff, M., Hillyard, S.: Modulation of early auditory processing during selective listening to rapidly presented tones. Electroencephalogr. Clin. Neurophysiol. **79**, 170–191 (1991)
33. Passow, S., et al.: Electrophysiological correlates of adult age differences in attentional control of auditory processing. Cereb. Cortex **24**, 249–260 (2012)
34. Daffner, K., et al.: Mechanisms underlying age- and performance-related differences in working memory. J. Cogn. Neurosci. **23**, 1298–1314 (2011)
35. Lubitz, A., Niedeggen, M., Feser, M.: Aging and working memory performance: electrophysiological correlates of high and low performing elderly. Neuropsychologia **106**, 42–51 (2017)

Cognitive Psychology in Air Traffic Control

A Preliminary Field Study of Air Traffic Controllers' Fatigue for Interface Design

Zhenling Chen[1], Jianping Zhang[1](✉), Wenchao Jing[2], Xiaofei Peng[1], Pengxin Ding[1], Yiyou Chen[1], Xiaoqiang Tian[1], Yanzhong Gu[1], and Guoliang Zou[3]

[1] The Second Research Institute of Civil Aviation Administration of China,
Chengdu 610041, China
zhangjp@caacsri.com

[2] Southwest Regional Air Traffic Management Bureau
of Civil Aviation Administration of China, Chengdu 610202, China

[3] Air Traffic Management Regulation Office, Civil Aviation Administration of China,
Beijing 10071, China

Abstract. Air traffic controllers' fatigue on duty affect their performance and even threaten aviation safety. But the design and promotion of air traffic management systems seldom consider dynamic changes of fatigue for controllers on duty. The aim of this study was to detect changes of fatigue for controllers on duty along with different task loads and working schedules. Fatigue influence sleep, mood and perception of controllers, and the three kinds of index employed to characterize their fatigue. A field investigation with 3 × 4 levels of two factors including task load and working schedule was carried out at one busy area control unit in China. 156 effective questionnaires (52 person times) were collected and analyzed with statistical methods. Primary results of repeated measure variance analysis showed that the accumulation of task load resulted in their sleep subscale scores increased very significantly, mood subscale scores and perception subscale scores decreased very significantly. Results of independent measure variance analysis showed that the influence of working schedule was also very significant to their sleep subscale, mood subscale and perception subscale scores. And the results of multiple factors covariance analysis showed that the interaction of task load and working schedule was significant. These primary results need to verify with large number of samples. These results clearly indicated that air traffic controllers' fatigue dynamically changed on duty. And the results will promote the design of new smart and adaptive air traffic management system which should take the dynamical change of controllers' fatigue and even capacity in to account.

Keywords: Air traffic controller · Fatigue · Task Load · Working schedule

1 Introduction

The rapid growth in commercial air travel, both in China and worldwide, is putting immense pressure on air traffic management (ATM). Consequently, ATM technological systems keep iteration and updating in order to meet the demands for enhanced capacity,

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 119–129, 2021.
https://doi.org/10.1007/978-3-030-77932-0_10

efficiency, and safety [1, 2]. And air traffic controllers (ATCOs) tend to work with more and more complex and powerful ATM systems, in which a proficient interaction between humans and computer systems is crucial to provide a safe and efficient service. Therefore, it has reached a consensus that the design of ATM systems needs to fit ATCOs ability demand [3]. For example, Brink, et al. established a new interface prototype that enables controllers to manipulate multiple flows of traffic by facilitating interaction with a path-planning algorithm [4]. IJtsma, et al. proposed to design an adaptive automation system to balance air traffic controller workload between underload and overload [5]. Borst, et al. put forward a prototype ecological interface for tactical conflict detection and resolution in the horizontal plane to assist ATCOs in fault diagnosis of automated advice [6]. Most of these studies regard ATCOs' ability and performance during operational duty as static stability. However, is the ATCOs' capacity to perform their duties stable during operational duty although they have got training and hold certifications?

The responsibility of ATCOs work is to avoid collisions of aircrafts with barriers on the ground, and with other aircrafts, and to guide aircrafts orderly and efficiently flow [7]. Therefore, ATCO has to monitor the aircraft assumed in the air-sector, and command the pilots for safe, efficient and smooth navigation, intervening in case of risk of infringing the prescribed safety separations between aircrafts. They tend to suffer fatigue during operational duty under such a multitask mode and stressful work environment [8, 9]. Recent years, along with the rapid increasement of flights in China, ATCOs' task loads increase quickly and their safe pressure increases fast in their daily work. For example, In July 2014, ATCO at Tianhe Airport, Wuhan slept on duty [10]. Obviously, fatigue is able to reduce ATCOs' capacity to perform their duties [8, 9]. What's more, fatigue results in human body biochemical metabolic changes [11, 12], makes his mood worse to cause aggressive behavior, and brings about perceptive ability decreased [13, 14]. Bad mood and reduced perception affect ATCOs' capacity to perform their duties very possibly. All this means that the state of ATCOs' ability during operational duty is dynamically changing. In this paper, we examined the changes of ATCOs' fatigue characterized as sleep, mood and perception three aspects during operational duty with a cross sectional field investigation in order to provide the trend information on dynamical changes of ATCOs for the design of smart and adaptive ATM systems.

2 Experiment

2.1 Study Design

ATCOs' task load and working schedule were designed as two independent variables in this study. A two factors mixed experimental design with 3 × 4 levels showed in Table 1. The 3 levels for task load set as 0 for rest, 0.5 for working in the middle of a shift, and 1.0 for complete working of a shift. The 4 levels for working schedule set as shift I (morning, 8:00–12:00), shift II (afternoon, 12:00–18:00), shift III (evening, 18:00–24:00), shift IV (deep night, 24:00–6:00).

2.2 Participants and Questionnaire Survey

ATCOs were recruited as volunteers in this study, who worked at one busy area control unit in China where the number of controlling flights was nearly 4000 a day during our

Table 1. Two factors mixed experimental design with 3 × 4.

Factor	Task load	Working schedule
Levels	0	Shift I
	0.5	Shift II
	1.0	Shift III
		Shift IV

survey period (from 26 to 28 Aug. 2019). They were regularly assigned to work for continuous two days and rest next continuous two days. All of them got quite enough rest before work. All the ATCO volunteers participating in the survey held certificates of competency. The general information of ATCO volunteers showed in Table 2.

Table 2. General information of ATCO volunteers in the study.

Age	Gender	People number	Shift I	Shift II	Shift III	Shift IV
23–42	Male	52	15	15	15	7
Mean 29.6						

On the premise to limit disturbance to minimum possible to ATCOs' work, we selected one scale developed in our previous work as the fatigue questionnaire which was simple and easy to understand. And the reliability and validity of the scale was verified with ATCOs during operational duty in the previous work [15]. The scale contains 7-item as sleepy, anxious, feel confident, irritable, able to concentrate, energetic, and slow degree of nerve, respectively. There are 10 degrees from 10 to 100 for each item, 10 represented the lowest degree, and 100 represented the highest degree. After reading the instructions on the scale, the volunteers were asked to select the value of each item to indicate how they currently feel.

The ATCO volunteers were required to fill out the questionnaire just before, in the middle, and just after their duties. Sixty ATCOs were recruited and fifty-eight ACTOs completed the scale three times as before, in the middle and after their duties.

The study protocol was approved by the Ethics Committee of the Second Research Institute of Civil Aviation Administration of China. All ATCOs participating in the experiment were provided with and signed an informed consent form. All relevant ethical safeguards have been met with regard to subject protection.

2.3 Statistical Analysis

Collected questionnaires were analyzed with statistic methods mainly including repeated measure variance analysis, independent variance analysis, multiple factors covariance analysis, and so on. In order to conveniently analyze and easily understand the results, 7-item of the scale was divided into three subscales as sleep subscale (SS) including sleepy, mood subscale (MS) including anxious, feel confident and irritable, and perception subscale (PS) including able to concentrate, energetic and slow degree of nerve.

3 Results and Discussion

156 questionnaires (as 52 persons) past the box plot test and applied to statistical analysis. The reliability of the questionnaires was very high (Cronbach's α, 0.85).

3.1 Influence of Task Load

Firstly, we investigated the influence of task load to ATCOs' fatigue with three aspects including sleep, mood and perception, respectively. Before ATCOs' duty, their task loads set as 0. In the middle of ATCOs' duty, their task loads set as 0.5 considering that there are regulations on balance of task loads by opening or closing sectors for very high or very low task loads in the 'Air traffic control rules for civil aviation' in China. And after ATCOs' duty, their task loads set as 1.0. The scores of ATCOs' SS, MS, and PS before, in the middle, and after duty showed in Fig. 1.

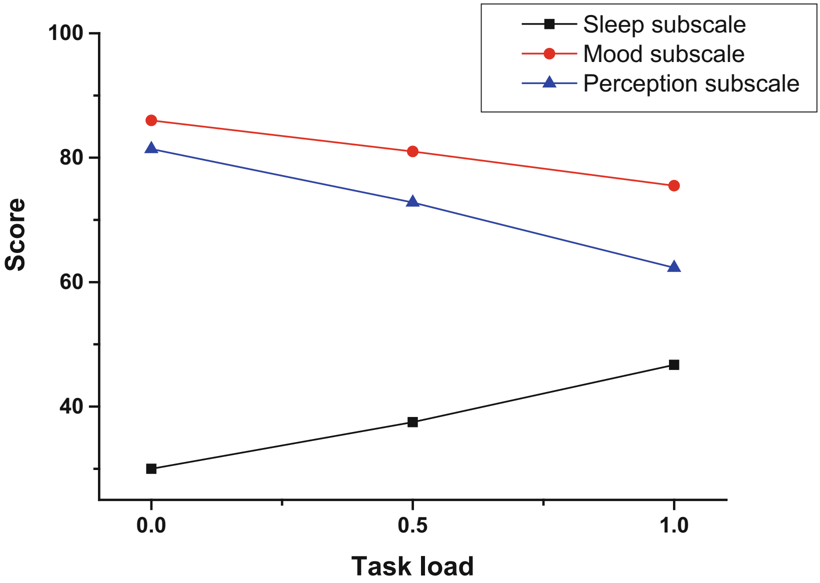


Fig. 1. Changes of ATCOs' fatigue with their task loads. (Color figure online)

Repeated measure variance analysis employed to the collected effective questionnaires to investigate the influence of task load to ATCOs' sleep, mood and perception, respectively. Results of ATCOs' SS showed that the influence of task load was very significant ($P < 0.001$, Fig. 1, black line). Multiple comparison analysis of the repeated measure data discovered that the SS scores in the middle of duty (mean 37.5) and after duty (mean 46.7) were very significant increased compared with that before duty (mean 30.0), respectively ($P = 0.006$, $P < 0.001$), and the SS scores after duty (mean 46.7) increased very significantly compared with that in the middle of duty (mean 37.5) ($P < 0.001$). These results indicated that the ATCOs' task loads resulted in their SS scores rapid increasing, and they kept to increase rapidly along with work time extension as task load accumulation. It was worth to notice that in the middle of duty, ATCOs' sleep changed very significantly compared with that before duty when they had performed duty only for 2 or 3 h. Therefore, it's a legitimate implication that ATCOs' sleep was a dynamical change process during they performed operational duty.

Then, for ATCOs' MS, results showed that the influence of task load was very significant ($P < 0.001$, Fig. 1, red line). Multiple comparison analysis of the repeated measure data discovered that the MS scores in the middle of duty (mean 81.0) and after duty (mean 75.5) were significant and very significant decreased compared with that before duty (mean 86.0), respectively ($P = 0.016$, $P < 0.001$), and the MS scores after duty (mean 75.5) decreased very significantly compared with that in the middle of duty (mean 81.0) ($P < 0.001$). These results indicated that the ATCOs' task loads resulted in their MS scores rapid decreasing, and they kept to decrease rapidly along with work time extension as task load accumulation. The results were similar with that of another survey for military drivers on their mood in China [14]. It's important to notice that in the middle of duty, ATCOs' mood changed significantly compared with that before duty when they had performed duty only for 2 or 3 h. It's reasonable to speculate that ATCOs' mood was a dynamical change process during they performed operational duty.

At last, for ATCOs' PS, the results, similar to that of MS, showed that the influence of task load was very significant ($P < 0.001$, Fig. 1, blue line). Multiple comparison analysis of the repeated measure data discovered that the PS scores in the middle of duty (mean 72.8) and after duty (mean 62.3) were significant and very significant decreased compared with that before duty (mean 81.4), respectively ($P = 0.011$, $P < 0.001$), and the PS scores after duty (mean 62.3) decreased very significantly compared with that in the middle of duty (mean 72.8) ($P < 0.001$). These results indicated that the ATCOs' task loads resulted in their PS scores rapid decreasing, and they kept to decrease rapidly along with work time extension as task load accumulation. The results were in accord with that of the survey for Chinese military drivers on their perception [14]. Similar to the mood, in the middle of duty, ATCOs' perception changed significantly compared with that before duty when they had performed duty only for 2 or 3 h. It's reasonable to speculate that ATCOs' perception was a dynamical change process during they performed operational duty.

The above results showed that ATCOs' fatigue characterized as sleep, mood and perception were influent significantly by task load. And the increasement of their fatigue presents as SS scores increasing, MS and PS scores decreasing. It's a legitimate implication that ATCOs' fatigue was a dynamical increase process during they performed operational duty. The results were consistent with that of another research for the fatigue of ATCOs at one terminal unit in China using the Stanford Sleeping Scale [16].

3.2 Influence of Working Schedule

Secondly, we investigated the influence of working schedule to ATCOs' sleep, mood and perception, respectively. In order to provide 24-h continuous air traffic management service, there were many working schedules at the area control units. We selected four typical working schedules as morning, afternoon, evening, and deep night as shift I, II, III, and IV for this study. The scores of ATCOs' SS, MS, and PS on the four shifts before duty showed in Fig. 2.

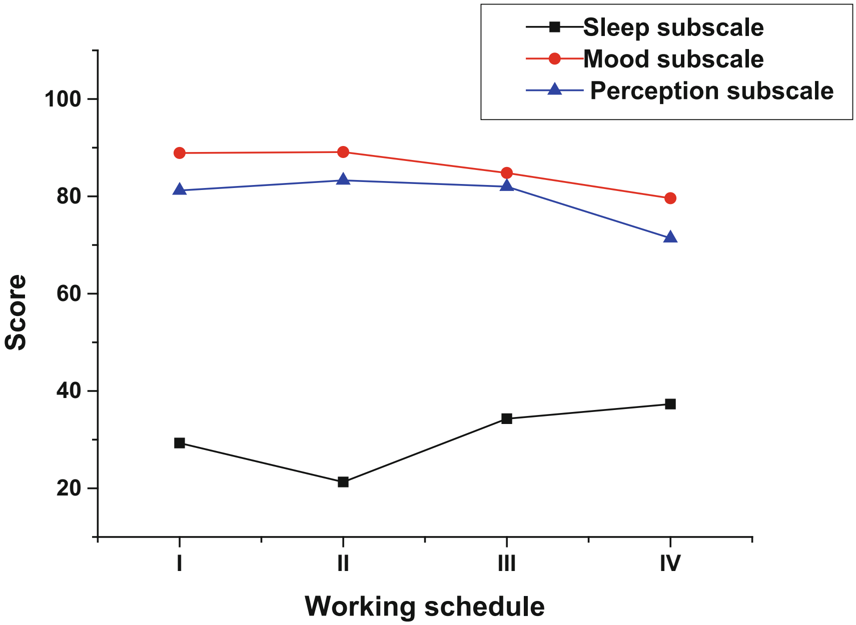


Fig. 2. Changes of ATCOs' fatigue along with their working schedules before duty. (Color figure online)

Independent measure variance analysis employed to the questionnaires before duty to investigate the influence of working schedule to ATCOs' sleep, mood and perception, respectively. The changes of ATCOs' three parameters before duty caused mainly from their basic factors such as circadian rhythm, age, and so on, without disturbance from task load considering they got enough rest before duty. And these basic factors reflected in the working schedule. For ATCOs' sleep (Fig. 2, black line), the variance analysis of independent measurement in the four shifts discovered that SS scores in shift IV increased significantly compared with that in shift I ($P = 0.030$), II ($P = 0.027$), and increased very significantly compared with that in shift III ($P = 0.006$), respectively. And there was no significant difference appeared for the SS scores among shift I, II and III. For ATCOs' mood (Fig. 2, red line), the statistical results showed that MS scores in shift IV decreased significantly compared with that in shift I ($P = 0.027$), II ($P = 0.035$), and III ($P = 0.041$), respectively. And there was no significant difference appeared for the MS scores among shift I, II and III, either. For ATCOs' perception (Fig. 2, blue line), the statistical results, similar to mood, PS scores decreased very significantly compared with that in shift I (P < 0.001), II (P = 0.002), and III (P = 0.001), respectively. And there was no significant difference appeared for the PS scores among shift I, II and III, either.

These results were consistent with the change of the circadian rhythm. In the day time, people remain clear-headed to work in accordance with shift I to III. And at night, people become fatigue and sleep to gather strength in accordance with shift IV [17, 18]. The above results also agreed with that of ATCOs at the terminal control unit in China [16].

3.3 Influence of Interaction Between Task Load and Working Schedule

Then, we investigated the interaction influence between task load and working schedule to ATCOs' sleep, mood and perception, respectively. Multiple factor covariance analysis employed to examine the interaction influence.

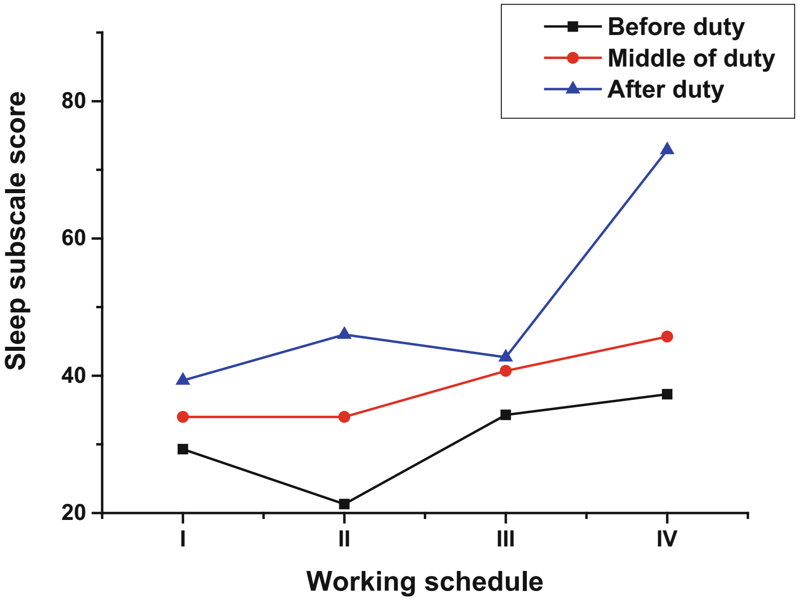


Fig. 3. Changes of ATCOs' sleep along with their task loads and working schedules

For ATCOs' sleep during operational duty, the changes of SS scores with different working schedule showed in Fig. 3. Results of covariance analysis showed that the interaction influence of task load and working schedule was very significant for ATCOs' sleep (P = 0.007). And the interaction influences enhanced ATCOs sleep. Therefore, this process is a dynamically changing course and is worth to pay attention in the design of new air traffic management systems.

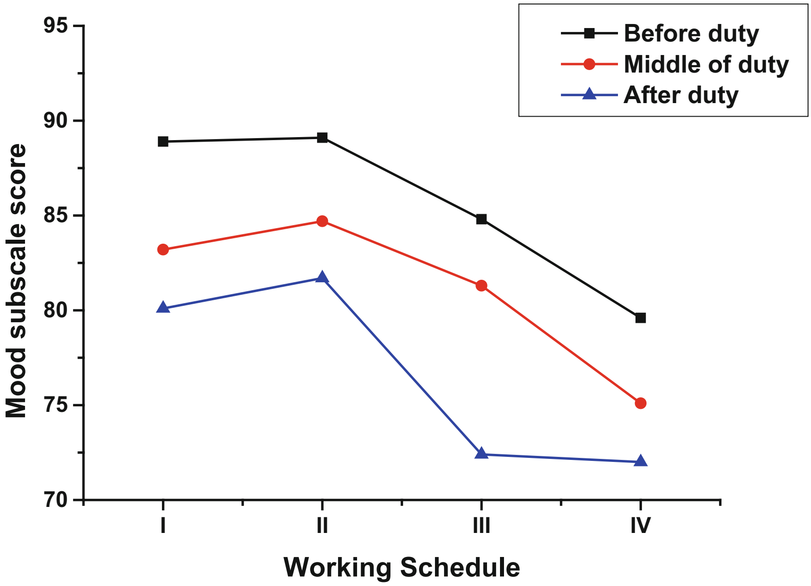


Fig. 4. Changes of ATCOs' mood along with their task loads and working schedules.

For ATCOs' mood during operational duty, the changes of MS scores with different working schedule showed in Fig. 4. Results of covariance analysis showed that the interaction influence of task load and working schedule was significant for ATCOs' mood (P = 0.046). To our knowledge, for the first time, we preliminarily demonstrated the interaction influence of task load and working schedule in ATCOs' mood with the questionnaire survey. Moreover, the interaction influence needs a greater number of samples to verify. The interaction influences reduced ATCOs mood and made their temperature more awful. And the process of ATCOs' bad mood is a dynamically changing course.

For ATCOs' perception during operational duty, the changes of PS scores with different working schedule showed in Fig. 5. Results of covariance analysis showed that the interaction influence of task load and working schedule was significant for ATCOs' perception (P = 0.031). And the interaction influences reduced ATCOs perception. To our knowledge, for the first time, we preliminarily demonstrated the interaction influence of task load and working schedule in ATCOs' perception with the questionnaire survey. Obviously, the reduced perception leads to lessen the capacity of ATCOs to perform their duties. Similarly, the process of ATCOs' perception decreasing is a dynamically changing course and worth to consider in the design of new ATM systems.

These results were also consistent to the result of the investigation for ATCOs at one terminal control unit in China in another work [16]. Those results indicated that ATCOs fatigue was a dynamically increasing process along with the accumulation of task loads and work at night shift. And fatigue is able to reduce the ATCOs' ability to conduct their duties [8, 9]. Therefore, this process is worth to pay attention in the design of new air traffic management systems.

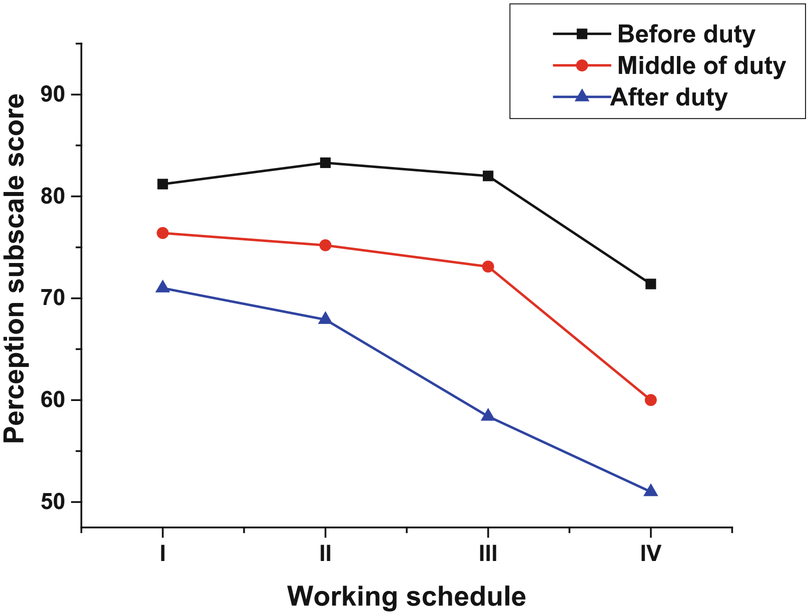


Fig. 5. Changes of ATCOs' perception along with their task loads and working schedules.

In brief, the primary results of this cross section field investigation clearly showed that during operational duty, ATCOs' fatigue dynamically changed. What's more, even they performed duty only for 2 or 3 h the three parameters changed significantly. These results full demonstrate that the dynamical process of the three parameters for ATCOs is not neglectable and is able to influence their capacity of duty. Therefore, the dynamical process is worth to pay special attention in the design of new ATM systems, otherwise, the new designed system will most likely not able to achieve desired results. Brink et al. applied their designed interface prototype to test with human participation, the sector robustness and efficiency did not improve the solution as compared the previously designed interface [4]. As for IJtsma's adaptive automation system, lower triggering thresholds increased the frustration level of participants and decreased acceptance of the support [5]. It may be a practical good try to take the dynamic process of ATCOs' fatigue into account for improvements of these designed systems. For example, it may decrease the frustration level of participants and increase acceptance of the support to set higher triggering thresholds at the onset of ATCOs' duty and then to decrease thresholds gradually for the adaptive automation system of IJtsma et al., considering their fatigue gradually increasing during operational duty.

4 Conclusion

In conclusion, the continuous increasing commercial flights require to improve the smart adaptive new ATM systems. The design of the systems seldom pay attention to the dynamical change of ATCOs fatigue to perform their duty during operational duty. In this paper, we primarily demonstrated that ATCOs' fatigue changed dynamically during they performed operational duty with a cross sectional field investigation. In particular, their sleep was tended to increase, their mood and their perception were tended to decrease on duty. Task load and working schedule as two main influent factors for the sleep, mood and perception of ATCOs were investigated. The results showed that the influence of task load was very significant to ATCOs' sleep resulted in it increased, and was also very significant to ATCOs' mood and perception resulted in them decreased. The results indicated that the influence of working schedule was very significant to ATCOs' sleep caused it increased at night, and was very significant to ATCOs' mood and perception resulted in them decreased at night. What's more, the results, for the first time, primarily discovered that the interaction influence between task load and working schedule exited indeed to ATCOs' fatigue. The results of the study were just primary results and needed a great number of samples confirm.

All these results demonstrated that on duty ATCOs' fatigue was dynamically changed with task load and working schedule. And the changes of their fatigue necessarily result in the changes of their capacity for work. Therefore, the design of new ATM systems should take the dynamical change of ATCOs' capacity into account. It means that the parameters for the design of systems include ATCOs' dynamical state as a variable in place of a constant. The expression or the change rule of the variable needs to further research.

Acknowledgement. We would like to thank all air traffic controllers who participated in this study. This study was supported by the National Natural Science Foundation of China (No. 52072406 and No. 22074157), Safety Foundation of Civil Aviation Administration of China (No. [2020]116), and Chengdu Science and Technology Project (2020-XT00–00001-GX).

References

1. EUROCONTROL. Five major challenges of long-term air traffic growth. https://www.eurocontrol.int/eec/public/standard_page/ETN_2009_1_Growth.html. Accessed 29 Dec 2020
2. Airbus. Global market forecast: Mapping demand 2016–2035. Airbus, Technical report. D14029463, August 2016. http://www.team.aero/files/airbusforecast/Global_Market_Forecast_2016-2035.pdf
3. Arico, P., Borghini, G., Flumeri, G., et al.: Human factors and neurophysiological metrics in air traffic control: a critical review. In: IEEE Reviews in Biomedical Engineering. hal-01511343 (2017). https://doi.org/10.1109/RBME.2017.2694142, https://hal-enac.archives-ouvertes.fr/hal-01511343
4. Brink, D., Klomp, R., Borst, C., et al.: Flow- based air traffic control: human-machine interface for steering a path-planning algorithm. In: 2019 IEEE International Conference on Systems, Man and Cybernetics, pp. 3186–3191. Bari, Italy (2019)
5. IJtsma, M., Borst, C., Mercado-Velasco, G., et al.: Adaptive automation based on air traffic controller decision-making. In: 19th International Symposium on Aviation Psychology, pp. 461–466. Dayton, Ohio, USA (2017)
6. Borst, C., Bijsterbosch, V.A., van Paassen, M.M., Mulder, M.: Ecological interface design: supporting fault diagnosis of automated advice in a supervisory air traffic control task. Cogn. Technol. Work **19**(4), 545–560 (2017). https://doi.org/10.1007/s10111-017-0438-y

7. International Civil Aviation Organization: Procedures for air navigation services air traffic management. 16th Edition (2016)
8. International Civil Aviation Organization: Manual for the oversight of fatigue management approaches. Doc 9966, 2nd Edition (2016)
9. Nealley, M.A., Gawron, V.J.: The effect of fatigue on air traffic controllers. Inter. J. Aviat. Psychol. **25**(1), 14–47 (2105)
10. Southcentral Regional Administration of Civil Aviation Administration of China: Report the investigation report on the air traffic controller sleeping on duty causing CEA MU2528 go around, Plain code telegram form the Southcentral Regional Administration [2014]1289 (2014)
11. Chen, Z., Xu, X., Zhang, J., et al.: Application of LC-MS-based global metabolomic profiling methods to human mental fatigue. Anal. Chem. **88**(23), 11293–11296 (2016)
12. Chen, Z., Zhang, J., Ding, P.: Fatigue detection of air traffic controllers using metabolomic methods. In: 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology, pp. 471–474. IEEE Press, Kunming (2019)
13. Lajunen, T., Parker, D.: Are aggressive people aggressive drivers? A study of the relationship between self-reported general aggressiveness, driver anger and aggressive driving. Accid. Anal. Prev. **33**(2), 243–255 (2001)
14. Yu, J., Yang, J., Zhang, S.: Research on state of fatigue and emotion changes in military drivers before and after driving. J. Occup. Health Damage. **30**(1), 48–51 (2015)
15. Chen, Z., Zhang, J., Ding, P., et al.: A scale to assess fatigue, concomitant mood and perception of air traffic controllers: a field study. In: 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology, pp. 874–877. IEEE Press, Weihai (2020)
16. Chen, Z., Zhang, J., Zou, G., et al.: Interaction effect of workload and circadian rhythm in air traffic controllers' fatigue. In: Proceedings of the 20th International Conference on Man-Machine-Environment System Engineering, pp. 235–241. Springer, Zhengzhou (2020)
17. Den, R., Toda, M., Nagasawa, S., et al.: Circadian rhythm of human salivary chromogranin A. Biomed. Res. **28**(1), 57–60 (2007)
18. Gu, Y., Chen, Z., Zhang, J., Zou, G., Ding, P., Deng, W.: A PERCLOS method for fine characterization of behaviour circadian rhythm. In: Long, S., Dhillon, B.S. (eds.) MMESE 2020. LNEE, vol. 645, pp. 243–249. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-6978-4_30

Supervising Multiple Remote Tower Operations: How to Develop and Test a New Work Position in the ATC Domain?

Maik Friedrich(✉), Felix Timmermann, and Jörn Jakobi

German Aerospace Center, Lilienthalplatz 7, 38108 Braunschweig, Germany
Maik.Friedrich@dlr.de

Abstract. Since single remote tower operations are already reality, further research concentrates on the provision of Air Traffic Services (ATS) to multiple airports in parallel by a single Air Traffic Control Officer (ATCO). The multiple remote tower center (MRTC) concept addresses how air traffic service is provided and also the uncertainties that influence the ATCO's ability to provide ATS. The next essential research step is a flexible and optimal allocation in between of several multiple remote tower modules (MRTM). The MRTC foresees the position of a supervisor that supports the active MRTM. For the design of the supervisor workplace a user-centered design approach was selected. This workplace and its procedures still have to be defined. This paper focuses on the initial steps to identify the workplace environment and the necessary procedures. Three workshops with 5 system matter experts were conducted and the results are presented. These results are used to design a validation set-up that could help to identify the overall benefit of the MRTC supervisor position. In summary the paper shows the initial steps in the human centered design approach applied to a developing workplace to handle a new task in the ATC domain.

Keywords: Multiple remote tower operations · System matter experts · Human centered design · Validation preparation

1 Introduction

A central interest of today's economy is using the available resources as efficient as possible. This is also true for the task of an air traffic control officer (ATCO). Remote tower research began with the development of single remote tower operations and its technical and practical challenges. In the last 15 years research developed a remote tower workplace as an alternative for the traditional tower controller workplace. From a remote tower workplace, ATCOs can provide safe and secure air traffic service (ATS) independent from the airport tower. Remote Tower Operation (RTO) is a solution for small airports that have a low amount of traffic that leads to under-utilized ATCOs and an inefficient use of resources.

Because single remote tower operations are already reality, further research concentrates on the provision of ATS to multiple airports in parallel by a single ATCO.

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 130–141, 2021.
https://doi.org/10.1007/978-3-030-77932-0_11

Therefore, the Multiple Remote Tower Center (MRTC) concept was defined. It not only addresses how a navigational service (ANS) is provided in parallel from a Multiple Remote Tower Module (MRTM) but also the uncertainties that influence the ATCO's ability to fulfill his/her duties. In an effort to keep safety at an all-time high and increase efficiency, MRTM have to be evaluated in a MRTC. A key aspect of the MRTC is the improvement in connection with work share and balancing between different MRTM. In a traditional tower, the coordination of tasks between controller working positions is handled by a supervisor. In analogy to that, the MRTC foresees the working position of a remote tower center supervisor that could support up to 5 active MRTM with a maximum of 15 active airports.

Defining a new position in the air traffic domain is long-lasting process. Starting with the conceptual necessity, its integration into the process, the requirements for needed information and the design of the workplace itself, including validation cycles after each major step. Even the concept of the MRTC is not yet conclusively defined and depends strongly on the current success of RTO. Keeping this in mind, this paper focuses on the initial steps to define the MRTC supervisor workplace and develops a valid set-up to test the ideas with system matter experts.

2 Design Procedure

EUROCONTROL and FAA [1] describe design factors that influence operators in aviation and thereby set guidelines for when developing a new workplace. Their twelve factors are summarized in Fig. 1, with the workflow of design, evaluation and measurement around them. The workflow around the factors is essential to ensure the safety of new developments. The factors in Fig. 1 represent all factors that should be considered for the development of workplaces within the ATC domain. The development of the MRTC supervisor has a unique impact on each of these twelve factors. For the human centered design approach, within this paper, the influential factors "Procedures/Working practice" (PWp) and "Workstation & Equipment" (WorkE) were selected to support the initial concept for the MRTC supervisor.

Friedrich [3] describes the workplace in general as an abstract representation of the environment allowing the operator to gather all needed information and interact with the environment as found necessary for the task. Endsley [4] states that user-centered design integrates all necessary information in ways that fit the goals, tasks, and needs of the users. For a structured approach, the user-centered design was separated in the following three steps. First, the design of a MRTC supervisor concept in relation to the selected influential factor Procedures/Working practice. System matter experts (ATCOs) evaluate the connection to analyze how such a workplace could be integrated into a workflow between different MRTM. Second, the identification of requirements in cooperation with system matter experts (ATCOs) that worked as traditional supervisors. The traditional supervisor workplace is thereby used as a baseline and enhanced with additional tasks that might become relevant in a MRTC. This is done in connection to the influential factor Workstation & Equipment. Third, in order to validate system matter experts' interaction with the MRTC supervisor workplace, realistic use cases had to be identified.

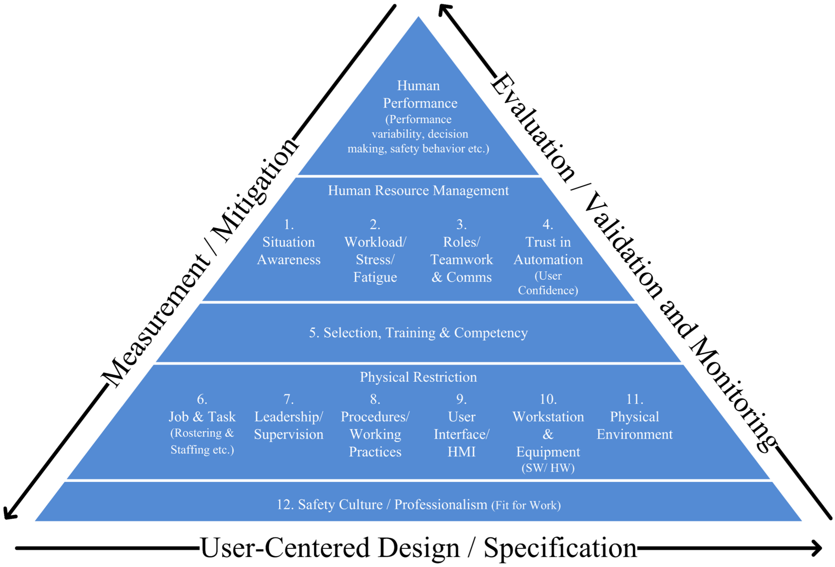

Human Performance (Performance variability, decision making, safety behavior etc.)
Human Resource Management
1. Situation Awareness
2. Workload/ Stress/ Fatigue
3. Roles/ Teamwork & Comms
4. Trust in Automation (User Confidence)
5. Selection, Training & Competency
Physical Restriction
6. Job & Task (Rostering & Staffing etc.)
7. Leadership/ Supervision
8. Procedures/ Working Practices
9. User Interface/ HMI
10. Workstation & Equipment (SW/ HW)
11. Physical Environment
12. Safety Culture / Professionalism (Fit for Work)
Measurement / Mitigation
Evaluation / Validation and Monitoring
User-Centered Design / Specification

Fig. 1. Influence of human performance in connection with new systems developed within ATC [adapded from 2]

3 Initial Concept Multiple Remote Center

The main responsibility of the MRTC supervisor is to ensure continuous, safe traffic flow and high quality of ATS provided at each MRTM. The role of an MRTC supervisor is to provide an efficient set up at all times and guarantee a flexible system by means of; maintaining overall supervision of all aerodromes within the MRTC; managing the allocation of staff and MRTMs; performing planning, administration, allocation of tasks and supervision of technical systems. The concept for the MRTC supervisor workplace was initially defined within the constraint's technical feasibility, required safety level, and increase in efficiency in relation to current procedures. The MRTC supervisors have to focus on balancing workload between the different MRTMs accommodated within a MRTC. The environment of a MRTC supervisor is a flexibel number of MRTM with either one, two, or three active aerodromes controlled by one ATCO. The environment also comprises additional communication requirements to various stakeholders on each airport e.g. approach control or fire station. Figure 2 presents an overview of the MRTC supervisor workplace and its possible connections to the MRTM. A direct line of communication, either by phone or due to close proximity, is thereby essential for time critical reaction. The MRTC supervisor also receives tactical information from each airport e.g. weather situation and amount of traffic. The MRTM workplace includes an optical sensor presentation replacing the direct out-of-the-window (OTW) view, radar, flight stripes, and of course radio communication to each airport that is controlled. The MRTM is not in the focus of this paper, see [5, 6] for more details.

In addition to the coordination of MRTM, the supervisor is responsible for gathering pre-tactical data from the airports in order to manage the opening and closing in a safe and

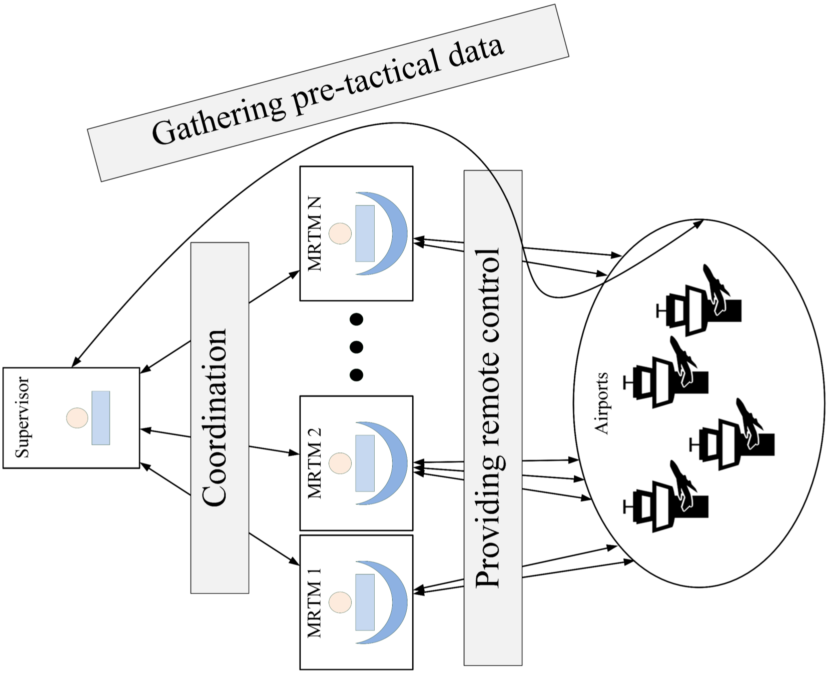


Fig. 2. General set-up for a possible Multiple Remote Tower Center

efficient way. The MRTC PWp foresees two ways to trigger the MRTC supervisor. Either he/she decides to initiate an action (top-down) or is requested by one of the MRTM to react (bottom-up). Following the top-down approach, the MRTC supervisor decides that an action is necessary, based on the gathered pre-tactical data. The bottom-up approach suggests an event on an airport is triggering an ATCO and he/she requests support from the MRTC supervisor. Both approaches have to be analyzed and considered during the workshops.

In combination with the PWp the MRTC concept foresees the design of a MRTC supervisor tool that is integrated in WorkE. In general, the MRTC supervisor workplace should include basic ATC functionalities, e.g. radio communication, telephones and weather information. This information is necessary to provide an overview. For balancing the task load between the MRTM, the MRTC supervisor needs information about the available staff and the traffic situations ahead.

4 Workshops

Three design workshops with system matter experts were conducted. The workshops were full day virtual telephone conferences on the 30th of May 2020, the 6th of October 2020 and the 20th of January 2021. In addition to voice communication, Microsoft Powerpoint was used to share and illustrate important talking points. In each workshop, 5 system matter experts participated. Two participants were from the ANS provider Oro navigacija (Lithuania) and three from PANSA (Poland). All participants were questioned together and in English. Each participant had knowledge about the traditional supervisor workplace and 4 of them worked in this position actively. The participants did not change between the workshops. All workshops had an equal duration of approximately 6 h.

Each workshop was designed as structured interview with a specific topic predefined. For each topic a set of questions was predefined and distributed to all the participants in advance. For the first workshop the focus was on the requirements defining the WorkE. The second workshop focused on the PWp. The final workshop was again for PWp, but focused on additional tasks derived from the traditional supervisor position and the MRTC supervisor tool.

The procedure was kept similar for each workshop. Two weeks before the workshop the participants had to read the MRTC concept and prepared questions that focused on the topic of the workshop. As an option the participants could provide additional information in connection to the topic as preparation to the workshop. At the beginning of each workshop a briefing was conducted and participants were instructed to answer as if they were the acting supervisor in a MRTC. The workshop moderator collected and summarized the given answers to the questions and comments in connection to the concept and workplace design. To increase the generalizability, the answers were prioritized if participants from both ANS providers agreed on the answer or if similar comments were provided from both.

As mentioned above, the first workshop focused on the responsibilities and the requirements defining the MRTC supervisor workplace, and in particular the design of a supervisor tool. These three aspects are interconnected, e.g., the responsibility to communicate, leads to the requirement to have a radio connection and the necessary frequencies on the display, which allows for a structured and stepwise approach. The following questions were prepared for each part to be discussed and answered at the end of the workshop:

- What would the responsibilities for a MRTC supervisor be? How would that differ from a supervisor in a traditional tower?
- What would the requirements for a MRTC supervisor position be?
- What is the necessary information for MRTC supervisor tool?

The requirements were collected in a structured manner by completing the categories title, description, rationale, category and importance. Title, description, and rationale were completed cooperatively. The moderator provided the category for each requirement. The importance was voted by the participants when all other categories were completed.

The second workshop focused on PWp by defining use cases that described specific tasks and procedures for a MRTC supervisor in connection with the ATCOs at the MRTM. For each use case the categories title, description, trigger, procedure, duration, supervisor tool implication and assumption. Title, description, and duration, supervisor tool implications and assumptions were completed in cooperation of all participants. The moderator decided if the bottom-up or top-down approach was applied to the use case, depending on the trigger. In addition to the categories and under the assumption

"Imagine to work this use case as a MRTC supervisor, ...", the following questions were answered in detail for each use case:

- Which kind of information do you need?
- What types of input/interaction would you need?
- What could be the main problems to avoid in such a situation?

The final workshop was equal to the second workshop in terms of procedure and structure. In order to complete the user-centered design approach, the focus was on additional MRTC supervisor tasks that are not in connection to the MRTM. The participants were instructed to identify use cases that are not directly related to the workload balancing between the MRTM. The categories and additional questions for each use case were also identical to the second workshop.

5 Results

The results show a structured overview of the workplace design process. The method of video conference workshops supported by Powerpoint was effective to collect answers for the questions and additional comments. The workshops were focused on their individual topics and allowed for a flexible scheduling. The moderator reported that all participants were equally motivated to participate. Even though there are differences between the two ANS providers concerning the role of the traditional supervisor or a MRTC supervisor, the overall requirements and use cases presented in this chapter are acceptable for both. The results in this paper focus on the must requirements in detail.

The results of the first workshop are divided into the three aspects responsibilities, requirements for the MRTC supervisor and requirements for the supervisor tool. The responsibilities of the MRTC supervisor are

- to contact and communicate with adjacent ATC sectors, other responsible organizations and internal units of the company about the relevant issues related with ATS provision.
- to contact and inform the Head of the Centre in case of incidents or violations related with ANS provision.
- to take all the necessary actions to support active ATCOs and maintain the efficient traffic flow.
- to monitor expected traffic load, weather forecast with respect to personal experience.
- to organize and control the shift of each ATCO ensuring the ANS is provided in a safe and efficient manner.
- to organize daily ATCO work plans in respect with ATCOs endorsements and/or potential training requirements.
- to organize shift's briefing and de-briefing.
- to prepare shift reports and other documentation related with ANS provision.
- to check the validation of ATCOs' license before the shift.

The workplace requirements from the must category are:

- MRTC supervisors must be equipped with a voice communication system which has every frequency and recording, received or transmitted from MRTC.
- MRTC supervisor must have flight plans, e-Flight strips, arrival/departure plans, and activations for all the airports on all the MRTM he/she supports.
- Standard equipment (e.g. UTC clock) must be clearly visible.
- MRTC supervisor must have the capability to monitor surveillance information and switch between multiple airports. It's useful for monitoring traffic load and traffic mix. It can also be helpful for short term staff planning.
- For the support tool 10 must requirements (and 4 should requirements) are identified. The must requirements are:
- The support tool must enable the MRTC supervisor with access to all frequencies and recorded radio transmissions.
- The supervisor tool must have chat functionality to every MRTM.
- The supervisor tool must show planned traffic in any desired period of time within a day.
- All information represented in aircraft labels must be seen at the ongoing operational time (current view) and the forecasted view.
- The airport labels must contain the information – runway in use, jurisdiction, NOTAMS and other information.
- The airport labels must include information about current and forecasted air traffic information from every airport (departing/arriving/transit traffic).
- The airport label must include information about current and forecasted NAV aids information from every MRTM, airport or overall equipment.
- The supervisor tool must directly provide weather conditions: visibility, cloud base, wind direction, wind shear.
- The supervisor tool must provide present and forecasted weather on each airport.
- The supervisor tool must provide the information about each ATCO at the MRTMs.

The results of the first workshops show that communication, planning and reporting are an important part of the MRTC supervisors' workload. Table 1 presents the summary of the comments and helps defining the basic description of the WorkE from the perspective of the system matter experts.

Table 1. Summary of the first workshop results

	Responsibilities	Requirements	
		MRTC supervisor	Super-visor tool
Communication	2	1	2
Planning	4	2	6
Reporting	2	1	1
Relation to total items	8 out of 9	4 out of 4	9 out of 10

The results of the second workshop are 11 use cases that describe the PWp. These use cases are the main tasks for the MRTC supervisor in relation to the MRTM. They are validated by both participating ANS providers. Table 2 presents the use case opening of an airport in detail.

Table 2. The use case opening up an airport in detailed description as collected during the workshop.

Title	Opening of an airport
Description	The scheduled opening of an airport needs to be handled
Trigger	Supervisor (bottom-up)
Procedure	1. Check if assigned ATCO is suitable 2. coordinating with the ATCO 3. collect response of successful opening
Duration	Approx. 15 min
supervisor tool implication	Activate an airport Checklist to inform as supervisor Support feature: Reminder for the Supervisor (e.g. 15 min before)
Assumption	The ATCO has already been briefed on airport situation
Which information do you need?	• Available ATCOs that fit to the airport • Available MRTM that could open up the airport The planned traffic for the next hours at this airport • What is the current situation at the airport opening up, weather, runway condition, NOTAM, Technical maintenance
what types of input/ interaction would you need?	• Communication between ATCO and SUP • ATCO does the actual opening of the airport • Green Button / automatic setting the airport active
what could be the main problems to avoid in such a situation?	• Technical issue at the airport is not available • Technical problem with the MRTM (e.g. no visual)

Table 3 presents an overview of all use cases that were defined in the second workshop. For each use case the degree of detail is equal to the example in Table 2. All use cases require to be either initiated by the supervisor (top-down) or the ATCO (bottom-up).

The 11 use cases are triggered twice by top-down, three times by bottom-up, and 6 times by both approaches. 8 use cases have an approximate duration of 10 to 15 min. 3 use cases have a time horizon depending on the situation. 9 use cases can be handled

Table 3. Title and description of the use cases in the second workshop.

Title	Description
Scheduled airport opening	The scheduled opening hours of an airport start and the airport needs to be opened
Unplanned airport opening	An aircraft requests landing for an airport that is closed
Scheduled airport closing	The scheduled closing hours of an airport begins and the airport needs to be closed
Unplanned airport closing	Due to severe weather events in the near future (e.g. low visibility) a specific airport has to be closed
Unplanned runway closing	Due to a technical failure an aircraft blocks the runway on a specific airport
Scheduled Workload increase	Due to excepted increase in workload one airport is handed over from the assigned MRTM to a different ATCO on second MRTM
Reduce ATCOs	To increase the efficient handling of the open airports and reduce the number of applied ATCOs, an airport needs to be merged into another already active MRTM
Daily planning	Due to an unexpected event an ATCO is not available for his/her shift that starts in a couple of hours
Fitting ATCO planning	Due to the shift planning an ATCO is handing over his MRTM to the next ATCO. The task is to find a new optimal allocation of the airports to the MRTM with a replacement ATCO who has different endorsements
Handling ATCO Request	Due to increased traffic volume on a specific airport, the ATCO on a MRTM either requests an additional ATCO for a specific airport or gets proposed an additional ATCO by the supervisor The supervisor then has to identify a suitable ATCO on hold to support the MRTM
Unscheduled ATCO replacement	Due to unexpected circumstances, an ATCO has to be relieved and replaced for some time by another ATCO (Ex. health issues) from his/her MRTM

in parallel with other use cases. In an attempt to define the major task for the MRTC supervisor the following three areas were identified: “Opening or closing of airports”; “Splitting or merging MRTMs”; and “Selection and checking of suitable ATCOs”.

The result of the third workshop are 9 use cases that describe the PWp. These use cases are tasks for the MRTC supervisor that are not directly related to the MRTM. Table 4 presents an overview of all use cases that were defined in the third workshop. For each use case the degree of detail is equal to the example in Table 2.

The use cases from the third workshop can be clustered into maintenance, coordination, and individual requests. 7 use cases have an approximate duration of 5 to 10 min

Table 4. Title and description of the use cases in the third workshop

Title	Description
Interaction with the technical unit	Maintenance work and operations information, depending on ATCO briefing, supervisor tool, or daily log entry
Interaction with Airport Maintenance	Airport Authority calls on construction work/maintenance work
Cleaning team	Cleaning in MRTC – a cleaning team would like to do their tasks
Facility Issues	Non-operational equipment failure, requires separate facility log/tickets for facility management
Flow Control Coordination	Calls or software inputs with/for flow control from a unit on one of the airports
MET Office Coordination	Coordination calls with MET office: e.g., because of wind forecasts or thunderstorms
Capacity coordination	Capacity shortcomings a few hours in the future, coordinate with the tower unit chief
Pilot Requests	Pilot calls for example to request a photo flight/ aerial survey flights or training flights
Drone Operation	Notification on camera drone operation/balloon "operation" within CTR - > coordinated directly with SUP

and consist of a single phone call. 2 use cases have a time horizon depending on the situation. All 9 use cases can be handled in parallel with other use cases from the second or the third workshop.

In addition to the structured interview, the participants also supported the theoretical application of the MRTC and made suggestions for the standard scenario and environment. This also helps to further specify the MRTC concept itself. The following comments have a connection to a possible validation of the MRTC concept:

> "the number of 5 MRTM with 3 airports at each MRTM seems to be a total maximum if the airports have simple schematics"
>
> "There is no need for the supervisor to hear communication of all airports, only non-nominal situations"
>
> "the maximum number of airport endorsements can be assumed to be 4"

Following the workshops, the next step in the user centered design process is to develop a validation based on the inputs.

6 Implication for Validations

This section discusses how the workshop results influence a possible initial validation for the MRTC concept. Such a validation has to focus on implementing WorkE and the PWp in a realistic environment. Therefore, the results from the workshops are essential to create an environment and implement procedures with a high degree of realism.

It is the authors opinion, that as validation methodology a real time simulation would be the best fit to create a realistic environment that incorporates the results from the workshops. The real time simulation should be based on computer simulation rather than video images from the airports. The flexibility to change and adapted WorkE or PWp feedback directly is needed for the validation. Of course, the participants are aware that they do not have to deal with real traffic, but this might enable them even more to generate additional ideas about the workplace. It also ensures a possibility to perform high capacity tests and abnormal scenario trials without a risk for live traffic.

As mentioned above, the operational scope of the validation includes the workshop results in connection to WorkE and PWp. The main focus of the validation should be the interaction between the MRTC supervisor and several MRTM. The MRTC supervisor coordinates up to a maximum of 15 airports controlled via a maximum of 5 MRTM and has to balance the ATCOs according to the requirements and endorsements.

To validate the defined use cases from the second workshop in a real time simulation, three working positions (one MRTC supervisor and two ATCOs at MRTMs) are required at minimum. This is especially due to the use cases "Scheduled Workload increase" or "Reduce ATCOs", with interaction between the MRTM. Each of the two MRTMs provide ATS to three small sized aerodromes, all operating in one-runway configuration. The rest of the MRTM and airports that the supervisor coordinates can be simulated.

The first validation within the design process for a new workplace is difficult because no reference scenario exists to measure improvement. Therefore, to evaluate Human Performance, Safety, Usability, and Capacity, the validation needs different scenarios, each with a different number of major tasks for the MRTC supervisor (see results of the second workshop). The validation scenarios are also to check and reevaluate the maximum number of responsibility and requirement from the first workshop. Besides that, each scenario should incorporate an equal amount of use cases from the third workshop to increase the external validity of the validation and to provide realistic environment conditions for the MRTC supervisor.

Objective feedback can be provided as workload measurements by the application of a standardized questionnaire between the scenarios or different use cases e.g., the NASA-TLX [7]. Subjective feedback is provided by a qualitative measure e.g. a structured debriefing. The subjective feedback should also be used to discuss and re-evaluate the results from the workshops and also the MRTC concept. An overall result of the validation is also operational feedback regarding the arrangement of the visual information of the MRTC supervisor workplace and the implementation of the support tool.

7 Summary

This paper shows the initial steps in the human centered design approach applied to a new workplace in the ATC domain. The structured approach decreases the possible need of redesign after implementation and helps to clarify the concept. The influential factors of human performance in connection with new ATC systems were used to set-up workshops investigating the WorkE and PWp.

The validation helps to understand if the responsibilities, requirements and use cases collected in the workshops are applicable in a realistic environment. Everything from interface design, amount of traffic, or methods of measurement can be influenced by these results. But, the results of the workshops do not only influence the validation but also help to define the MRTC concept in more detail. This helps to increase the quality of the concept, but also increases the acceptance in the community of ATCOs.

Future work should concentrate on implementing the validation with the described implications from the workshops. This should be followed by a structured analysis of additional influential factors of human performance for new workplaces in the ATC domain.

Acknowledgements. The authors thank SESAR Joint Undertaking for funding this project as well as their project partners PANSA, Oro Navigacija and Frequentis for their collaboration. This project has received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874470.

References

1. EUROCONTROL, and FAA, A Human Performance Standard of Excellence (2015)
2. EUROCONTROL/FAA, Human Performance in Air Traffic Management Safety: A White Paper, Action Plan 15 Safety (2010)
3. Friedrich, M.: Designing a workplace in the aviation domain: the transition to a remote air traffic control workplace by analysing the human-computer interaction. Technischen Universität Chemnitz, online, Fakultät für Human- und Sozialwissenschaften (2019)
4. Endsley, M.R.: Designing for Situation Awareness: An Approach to User-Centered Design: CRC Press (2016)
5. SESAR Joint Undertaking, *Remote Tower for Multiple Airports*: Project funded by SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 730195 (2019)
6. Friedrich, M., Hamann, A., Jakobi, J.: An eye catcher in the ATC domain: Influence of Multiple Remote Tower Operations on distribution of eye movements
7. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. Human Mental Workload **1**, 139–183 (1988)

Psychophysical Coherence Training Regulating Air Traffic Controller's Heart Rate Variability and Resilience to Fatigue

Wen-Chin Li[1](✉), Jingyi Zhang[1], Peter Kearney[2], and Graham Braithwaite[1]

[1] Cranfield University, Bedford, UK
Wenchin.li@cranfield.ac.uk
[2] Irish Aviation Authority, 11-12 D'olier Street, Dublin, Ireland

Abstract. Heart Rate Variability (HRV) can reflect individuals' cognitive workload objectively. HRV measurement is a non-invasive method to evaluate relevant physiological changes in a human body. Physiological changes and cognitive processes are associated with the cardiac dynamic autonomic control, and thus to influence individuals' ability to cope with fatigue and achieve resilience. As a dynamic process to be learned, resilience can be formed and improved through Quick Coherence Training in a short time. To study the regulatory effects of coherence training on Air Traffic Controllers' (ATCOs) HRV and resilience ability, the HRV parameters of 34 qualified ATCOs before and after Quick Coherence Training (QCT) are collected and analysed by paired T-Test. Also, participants' HRV at rest are recorded as baseline. The results show that the coherence, mean RR interval, SDNN, and RMSSD after QCT are significantly higher than before. The mean heart rate after QCT is significantly lower than before. The findings demonstrate that coherence has effective and efficient regulatory effects on coherence, mean RR interval, SDNN, mean heart rate, and RMSSD. Psychological coherence training can be an efficient method for ensuring ATCOs recover from fatigue and achieve resilience quickly during the short break in long-time continuing monitoring and controlling. The findings in the current research can have potential to be further developed for Fatigue Risk Management System which is required by ICAO (International Civil Aviation Organization) to be applied for Air Navigation Services providers.

Keywords: Heart rate variability · Psychophysical resilience · Quick Coherence Training · Fatigue risk management

1 Introduction

Fatigue risk management is considered to be most effective when it is integrated into, or supported by, an SMS, thereby forming an FRMS [1]. Fatigue risk management system (FRMS) have been introduced on the basis that "better results (both in terms of safety and productivity) might be obtained from approaches that are more comprehensive, more flexible, and better tuned to current scientific understanding of key factors in fatigue

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 142–150, 2021.
https://doi.org/10.1007/978-3-030-77932-0_12

prevention" [2]. In fact, the prescriptive approach has been under increased scrutiny as to its benefits in preventing fatigue. The main assumption is that regulations are not able to take into account the complexity of fatigue [1]. For instance, because of circadian rhythms, a break will not have the same recovery value depending on the time of the day, the timing of the break being more important than the duration of the break itself. This is the reason why alternatives to a prescriptive approach, such as FRMS, are becoming more popular [3].

Resilience refers to the ability for individuals to resist or recover from suffering negative experiences such as stress and fatigue [4]. Resilience has been proven to be associated with some inherent traits of individuals. There are significant correlations between psychological resilience and the big five personality traits: resilience is positively correlated with conscientiousness, agreeableness, openness to experience and extraversion, and negatively correlated with neuroticism [5]. Also, resilience is associated with social intelligence and coping styles related to task and emotion [6, 7]. Furthermore, there are some psychiatric illnesses or cognitive disorders which could influence the ability of resilience [7, 8]. On the other hand, the formation and improvement of resilience are considered as the dynamic process which could be learned at any period of life [9]. Resilience could be acquired over a period of time by using a process rather than coming all at once [10]. Intervention techniques based on cognitive behavioral therapy or mindfulness have been proven to have a positive impact on resilience [11].

Heart rate variability (HRV) can reflect a person's cognitive workload objectively [12, 13]. Healthy biological systems exhibit mathematically chaotic variations. HRV is composed of variations in the interval between successive heartbeats, known as the heartbeat interval (IBIs). A healthy heart is not a metronome. The oscillations of healthy hearts are not only complex but also varied. This enables the cardiovascular system to quickly adapt to any challenges for the homeostasis. Due to the influence of respiration, blood pressure and skin temperature on the heart control mechanism, the heart rate of normal people will undergo physiological changes. HRV index is the neurocardiac function, which is generated by the cardiac and brain interaction and the dynamic nonlinear Autonomic Nervous System (ANS) process. From a clinical perspective, the relationship between specific psychologic states and patterns of autonomic physiological responses is a particularly important issue for further research. HRV reflects the physiological changes in the human body by a non-invasive measure method [14, 15]. Mental states and processes can have an impact on cardiac dynamic autonomic control [16]. The HRV parameters including basal characteristics, time domain, frequency domain and non-linear are proved to be associated with fatigue and perceived workload [17, 18]. HRV basal characteristics include Coherence, Mean R-R interval, Mean HR and STDHR; Time-domain parameters include SDNN, RMSSD, NN50, and pNN50; Frequency domain measurements include LF, HF, Total Power, and Ratio LF/HF. The research aim is to develop FRMS for air navigation services providers by psychophysical coherence training.

2 Method

2.1 Participants

Thirty-four qualified Air Traffic Controller Operators (ATCOs, 27 male and 7 female) participated in this research. The ages of participants ranged between 23 and 58 years of age (M = 41.21, SD = 7.49), and their work experience as air traffic controllers varied from 1 to 38 years (M = 17.28, SD = 9.38). The collected data was gathered from human subjects; therefore, the research proposal was submitted to the Cranfield University Research Ethics System for ethical approval. As stated in the consent form, participants have the right to terminate the experiment at any time and to withdraw their provided data at any moment even after the data collection.

2.2 Apparatus

HRV Measurement Device. The HeartMath emWave device was applied to collect HRV data while participants performed various tasks. The device is equipped with an ear sensor which can gather the Inter Beat Interval (IBI) parameters, and it is possible to connect it using Bluetooth and export HRV data to other mobile devices through Inner Balance application (Fig. 1). Kubios is an advanced tool for investigating the variability of heart-beat intervals. Due to the wide variety of different analysis options and easy-to-use interface, the software is suitable for researching various premises. The software is suitable for clinical and public health researchers working on human HRV. This research applied the latest version of Kubios HRV developed at University of Eastern Finland [17]. The Kubios software is used to analyze participants' HRV parameters including time domain, frequency domain and non-linear (Fig. 2).

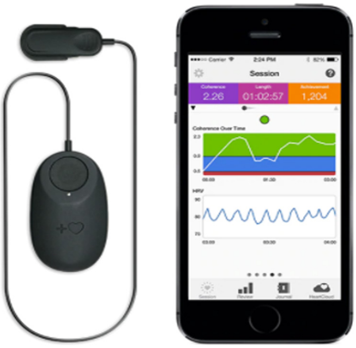
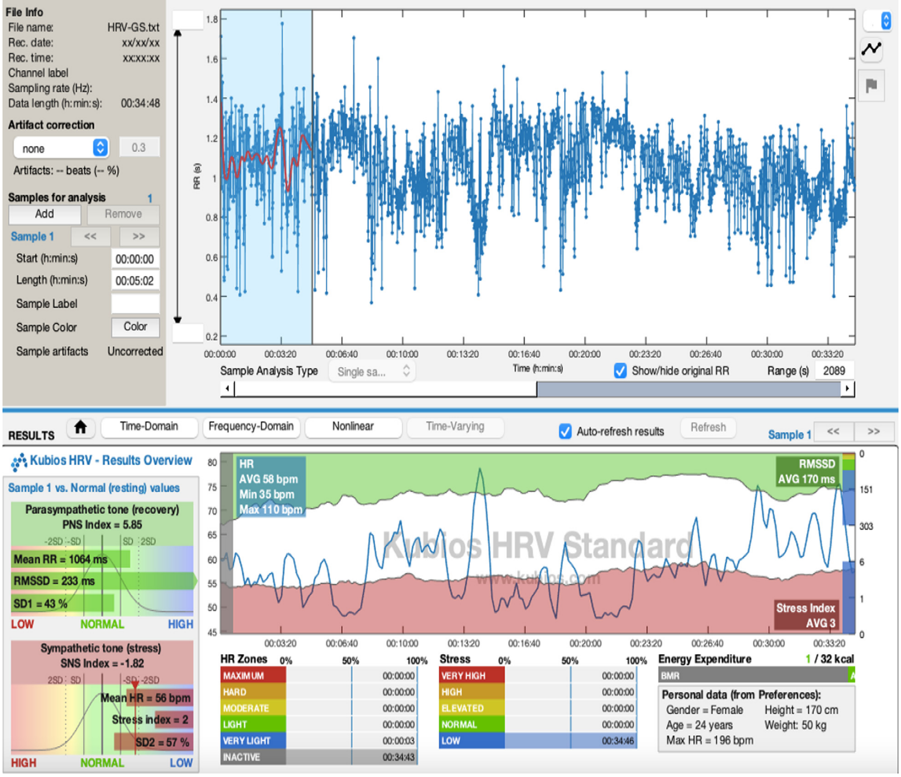


Fig. 1. Inner Balance Bluetooth sensor for Android & iPhone

HeartMath Quick Coherence Technique. The Quick Coherence Training (QCT) proposed by HeartMath can shift from stress and frustration to balance and resilience in about one minute with simple, but powerful steps. Coherence refers to a state where thoughts



File Info
File name: HRV-GS.txt
Rec. date: xx/xx/xx
Rec. time: xx:xx:xx
Channel label
Sampling rate (Hz):
Data length (h:min:s): 00:34:48

Artifact correction
none 0.3
Artifacts: -- beats (-- %)

Samples for analysis 1
Start (h:min:s) 00:00:00
Length (h:min:s) 00:05:02
Sample Color
Sample artifacts Uncorrected

Sample Analysis Type Single sa...
☑ Show/hide original RR
Range (s) 2089

☑ Auto-refresh results

Kubios HRV - Results Overview

Sample 1 vs. Normal (resting) values

Parasympathetic tone (recovery)
PNS Index = 5.85
Mean RR = 1064 ms
RMSSD = 233 ms
SD1 = 43 %

Sympathetic tone (stress)
SNS Index = -1.82
Mean HR = 56 bpm
Stress index = 2
SD2 = 57 %

HR
AVG 58 bpm
Min 35 bpm
Max 110 bpm

RMSSD
AVG 170 ms

Stress Index
AVG 3

HR Zones	
MAXIMUM	00:00:00
HARD	00:00:00
MODERATE	00:00:00
LIGHT	00:00:00
VERY LIGHT	00:00:03
INACTIVE	00:34:43

Stress	
VERY HIGH	00:00:00
HIGH	00:00:00
ELEVATED	00:00:00
NORMAL	00:00:00
LOW	00:34:46

Energy Expenditure 1 / 32 kcal
BMR

Personal data (from Preferences):
Gender = Female Height = 170 cm
Age = 24 years Weight: 50 kg
Max HR = 196 bpm

Fig. 2. Example of HRV parameters analysis by Kubios

and emotions are balanced, this is reflected in more balanced heart rhythms and it facilitates brain function and access to higher intelligence. The HeartMath Quick Coherence Training is an effective technique to achieve energy, mental clarity and resilience. There are two simple and quick steps for Coherence Training by the technique of heart-focused breathing.

2.3 Research Design

Each participant carries out the same procedures follows; (1) briefing the research aims (five minutes); (2) providing the demographical data including age, gender, qualifications, and working experience (five minutes); (3) wear Heart Math HRV measurement device for data collection on air traffic control working position (60 min); (4) conducting Quick Coherence Training (30 min); (5) Collecting post-training session of HRV (5 min).

3 Results

The HRV parameters of 34 participants before and after QCT are collected. Also, participants' HRV data at rest was recorded as baseline. For the convenience of statistical analysis and based on the research goals, five parameters: Coherence, Mean RR, SDNN, Mean HR, and RMSSD are involved in the statistical analysis in the current study. Mean RR is the heart inter-beat (R-R) interval; SDNN is the standard deviation of normal to normal (N-N) R-R intervals, reflecting the sympathetic and parasympathetic activity influencing HRV; Mean HR is the mean heart rate in beats per minute; and RMSSD is the square root of the mean of the sum of the squares of differences between adjacent NN intervals. Paired T-Test was applied for data analysis. Cohen's d was calculated as the effect size metric. The sample characteristics and T-Test results of HRV parameters on three phrases are shown as Table 1.

Table 1. The means and standard deviation of HRV parameters and paired T-Test results between before and after QCT

HRV parameters (Baseline)	QCT	M	SD	N	T-Test			
					t	df	p	Cohen's d
Coherence (27.01)	Before	25.74	5.22	34	29.14	33	.000	5.00
	After	70.56	9.19					
Mean RR (ms) (796.20)	Before	814.53	105.54	34	4.05	33	.000	0.69
	After	922.65	105.10					
SDNN (ms) (85.10)	Before	71.15	39.83	34	2.87	33	.007	0.49
	After	101.43	46.45					
Mean HR (bpm) (76.48)	Before	74.91	9.98	34	− 4.22	33	.000	-0.72
	After	65.85	7.45					
RMSSD (ms) (88.94)	Before	76.08	51.79	34	2.60	33	.014	0.45
	After	109.62	53.91					

The T-Test results show that the Coherence after QCT (M = 70.56, SD = 9.19) is significantly higher than before (M = 25.74, SD = 5.22), t = 29.14, p < .001, d = 5.00. The Mean RR after QCT (M = 922.65, SD = 105.10) is significantly higher than before (M = 814.53, SD = 105.54), t = 4.05, p < .001, d = 0.69. The SDNNs after QCT (M = 101.43, SD = 7.96) is significantly higher than before (M = 39.82, SD = 6.83), t = 2.87, p < .01, d = 0.49. The Mean HR after QCT (M = 65.85, SD = 7.50) is significantly lower than before (M = 74.90, SD = 9.98), t = −4.22, p < .001, d = −0.72. The RMSSD after QCT (M = 109.62, SD = 53.91) is significantly higher than before (M = 76.08, SD = 51.79), t = 2.60, p < .05, d = 0.45.

4 Discussion

The T-Test results demonstrated that coherence training has significant regulatory effects on coherence, mean RR, SDNN, mean HR, and RMSSD. Coherence is the state when the heart, mind and emotions are in energetic alignment and cooperation. Gaining coherence is a strategy to build resilience and accumulate personal energy, recovering and keeping more energy for positive outcomes. A QCT program or training which increases sense of coherence could contribute to decrease stress and fatigue, as well as improve resilience [19]. Furthermore, compared to various complicated physiological parameters, coherence could be considered as a simple and quick indicator for evaluating individuals' general mental workload. The correlational and causative relationships between coherence and HRV as well as various other physiological systems have been revealed [20]. Therefore, it is effective and efficient as a means of fatigue management to monitor ATCOs' coherence level. If used during break times it should permit ATCO's to achieve resilience throughout their working day.

In terms of the fluctuation of HRV parameters, the statistical analysis reveals that coherence training could help ATCOs make a good recovery from Air Traffic Control tasks to a parallel or even better level with baseline HRV in a short time. The coherence increased significantly indicating participants have built a good resilience via coherence training. Moreover, the increased mean RR interval, SDNN, and RMSSD indicates the recovery from work pressure and mental stress [21–23]. Heart rate could increase with higher workload and complexity of tasks [24]. The decreasing heart rate of participants indicates effective relaxation and recovery from workload and fatigue. Through coherence training, participants are in a relatively relaxed state and able to embrace further ATC tasks and mental challenges.

However, participants' coherence and HRV parameters did not simply return to baseline state at rest, especially for the mean RR interval and heart rate (Fig. 3). On the one hand, this finding confirms the effectiveness of coherence training for resilience and recovery. On the other hand, it is uncertain that the baseline HRV parameters demonstrate the optimum state for more ATC tasks. It is noted that mean RR interval increase and heart rate decrease during work, which means less stress and mental workload. This phenomenon could be attributed to the influence of passive fatigue, which might be caused by boredom, task underload or monotony [25]. Therefore, it is essential to find proper points to initiate coherence training and engage in tasks without passive fatigue.

Table 1 demonstrated that ATCOs' mean RR had significantly increased from 814.53 ms to 922.65 ms with SDNN increasing from 71.15 ms to 101.43 ms by practicing quick coherence training compared not using QRT on the working position. While the HeartMath tools are intentionally designed to be easily learned and used in day-today life, professionals suggest that these techniques often facilitate profound shifts in perception, emotion and awareness. Moreover, extensive laboratory research performed at HMI has shown that the physiological changes accompanying such shifts are dramatic. Several studies using various combinations of these QCT techniques have found significant correlations between psychophysical coherence and improvements in cognitive function. For example, a study of school students with Attentional Deficit Hyperactivity Disorder (ADHD) showed a wide range of significant improvements in short and long-term memory, ability to focus and significant improvements in behaviours both at home and in school [26]. A study of 41 fighter pilots engaging in flight simulator tasks found a significant correlation between higher levels of performance and heart coherence as well as lower levels of frustration [27]. A study of patients diagnosed with Post-Traumatic

Stress Disorder (PTSD) found that relatively brief periods of HRV coherence training combined with practicing the Quick Coherence Technique resulted in significant improvements in the ability to self-regulation along with a wide range of cognitive functions [28]. McCraty et al. concluded that psychologically the coherence state promotes a calm, emotionally balanced, yet alert and responsive state with a sense of enhanced subjective well-being, which is conducive to cognitive function and task performance, including problem-solving, decision-making, and activities requiring perceptual acuity, attentional focus, coordination, and discrimination [29]. Edwards evaluated the influence coherence training both quantitatively and qualitatively. The findings indicated significant improvements in health, mindfulness, and spirituality perceptions, as well as meaningful learning experiences and positive evaluations of the coherence training feedback for health, sport, exercise, performance, meditation, and daily life contexts [30]. Field et al. also provided consistent evidence that the coherence training program is both feasible and effective in improving heart rate variability, physiological relaxation and mindfulness [31]. The technique of quick coherence did increase ATCOs' resilience to fatigue and improve ATCOs' health and well beings.

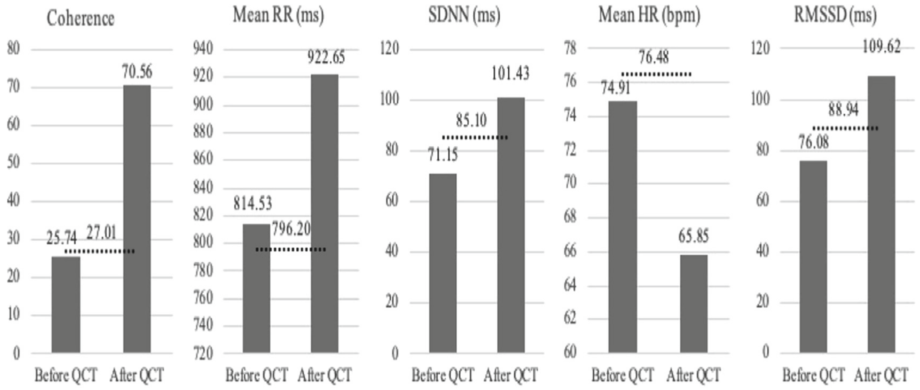


Fig. 3. Participant' coherence, mean RR, SDNN, mean HR and RMSSD before QCT and after QCT, and at rest as baseline (dotted lines)

5 Conclusion

The statistical results demonstrated that coherence training could regulate ATCOs' HRV parameters effectively and efficiently. Psychological coherence training could contribute to recovery from heavy workload and achieve resilience to cope with fatigue. However, the proper initial state and optimum outcomes of HRV parameters for coherence training are still uncertain. The negative influence on ATCOs HRV parameters and resilience ability of passive fatigue should be confirmed and would benefit from further research. QCT could be an efficient method for ensuring ATCOs' recovery and resiliency in a short break form long-time shift work. In summary, the current research could have the potential to be further developed for Fatigue Risk Management System which is required by ICAO (International Civil Aviation Organization) to be applied for Air Navigation Services providers.

References

1. Dawson, D., McCulloch, K.: Managing fatigue as an integral part of a safety management system. In: Proceedings of Fatigue Management in Transport Operations Conference, Seattle, USA (2005)
2. Civil Aviation Safety Authority: Fatigue management: suggested alternatives to prescribed flight and duty times. Civil Aviation Safety Authority, Canberra, Australia (2004)
3. Cabon, P., Deharvengt, S., Grau, J.Y., Maille, N., Berechet, I., Mollard, R.: Research and guidelines for implementing fatigue risk management systems for the French regional airlines. Accid. Anal. Prevent. **45**(Supplement), 41–44 (2012). https://doi.org/10.1016/j.aap.2011.09.024
4. Rutter, M.: Resilience, competence, and coping. Child Abuse Negl. **31**(3), 205–209 (2007). https://doi.org/10.1016/j.chiabu.2007.02.001
5. Fayombo, G.: The relationship between personality traits and psychological resilience among the Caribbean adolescents. Int. J. Psychol. Stud. **2**(2), 105–116 (2010). 10.1.1.665.3361
6. Friborg, O., Barlaug, D.A.G., Martinussen, M., Rosenvinge, J.A.N.H.: Resilience in relation to personality and intelligence. Int. J. Methods Psychiatric Res. **14**(1), 29–42 (2005). https://doi.org/10.1002/mpr.15
7. Campbell-Sills, L., Cohan, S.L., Stein, M.B.: Relationship of resilience to personality, coping, and psychiatric symptoms in young adults. Behav. Res. Ther. **44**(4), 585–599 (2006). https://doi.org/10.1016/j.brat.2005.05.001
8. Hechtman, L.: Resilience and vulnerability in long term outcome of attention deficit hyperactive disorder. Can. J. Psychiat. **36**(6), 415–421 (1991). https://doi.org/10.1177/070674379103600606
9. Flach, F.F.: Psychobiologic resilience, psychotherapy, and the creative process. Compr. Psychiatry **21**(6), 510–518 (1980). https://doi.org/10.1016/0010-440X(80)90054-1
10. Fine, S.B.: Resilience and human adaptability: who rises above adversity? Am. J. Occup. Ther. **45**(6), 493–503 (1991). https://doi.org/10.5014/ajot.45.6.493
11. Joyce, S., Shand, F., Tighe, J., Laurent, S.J., Bryant, R.A., Harvey, S.B.: Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open **8**(6), 1–9 (2018). https://doi.org/10.1136/bmjopen-2017-017858
12. Aasman, J., Mulder, G., Mulder, L.J.: Operator effort and the measurement of heart-rate variability. Hum. Factors **29**(2), 161–170 (1987). https://doi.org/10.1177/001872088702900204
13. Kramer, A.F.: Physiological metrics of mental workload: a review of recent progress. In: Multiple-Task Performance, pp. 279–328. Taylor and Francis, London (1991). https://apps.dtic.mil/dtic/tr/fulltext/u2/a223701.pdf
14. Durantin, G., Gagnon, J.F., Tremblay, S., Dehais, F.: Using near infrared spectroscopy and heart rate variability to detect mental overload. Behav. Brain Res. **259**, 16–23 (2014). https://doi.org/10.1016/j.bbr.2013.10.042
15. Rowe, D.W., Sibert, J., Irwin, D.: Heart rate variability: indicator of user state as an aid to human-computer interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 480–487, Los Angeles, CA, USA (1998). https://doi.org/10.1145/274644.274709
16. Stein, P.K., Bosner, M.S., Kleiger, R.E., Conger, B.M.: Heart rate variability: a measure of cardiac autonomic tone. Am. Heart J. **127**(5), 1376–1381 (1994). https://doi.org/10.1016/0002-8703(94)90059-0

17. Tarvainen, M.P., Niskanen, J.P., Lipponen, J.A., Ranta-Aho, P.O., Karjalainen, P.A.: Kubios HRV–heart rate variability analysis software. Comput. Methods Programs Biomed. **113**(1), 210–220 (2014). https://doi.org/10.1016/j.cmpb.2013.07.024
18. Watson, D.W.: Physiological correlates of heart rate variability (HRV) and the subjective assessment of workload and fatigue in-flight crew: a practical study. In: Proceedings of the Second International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres, pp. 159–163. Institution of Engineering and Technology, Manchester, UK (2001). https://doi.org/10.1049/cp:20010453
19. Foureur, M., Besley, K., Burton, G., Yu, N., Crisp, J.: Enhancing the resilience of nurses and midwives: Pilot of a mindfulness-based program for increased health, sense of coherence and decreased depression, anxiety and stress. Contemp. Nurse **45**(1), 114–125 (2013). https://doi.org/10.5172/conu.2013.45.1.114
20. Edwards, S.D.: HeartMath: a positive psychology paradigm for promoting psychophysiological and global coherence. J. Psychol. Afr. **25**(4), 367–374 (2015). https://doi.org/10.1080/14330237.2015.1078104
21. Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M., Pecchia, L.: Acute mental stress assessment via short term HRV analysis in healthy adults: a systematic review with meta-analysis. Biomed. Signal Process. Control **18**, 370–377 (2015). https://doi.org/10.1016/j.bspc.2015.02.012
22. Lackner, H.K., Papousek, I., Batzel, J.J., Roessler, A., Scharfetter, H., Hinghofer-Szalkay, H.: Phase synchronization of hemodynamic variables and respiration during mental challenge. Int. J. Psychophysiol. **79**(3), 401–409 (2011). https://doi.org/10.1016/j.ijpsycho.2011.01.001
23. Visnovcova, Z., et al.: Complexity and time asymmetry of heart rate variability are altered in acute mental stress. Physiol. Meas. **35**(7), 1319–1334 (2014). https://doi.org/10.1088/0967-3334/35/7/1319/meta
24. Jorna, P.G.A.M.: Heart rate and workload variations in actual and simulated flight. Ergonomics **36**(9), 1043–1054 (1993). https://doi.org/10.1080/00140139308967976
25. Körber, M., Cingel, A., Zimmermann, M., Bengler, K.: Vigilance decrement and passive fatigue caused by monotony in automated driving. Proc. Manuf. **3**, 2403–2409 (2015). https://doi.org/10.1016/j.promfg.2015.07.499
26. Lloyd, A., Brett, D., Wesnes, K.: Coherence training in children with attention-deficit hyperactivity disorder: cognitive functions and behavioral changes. Altern. Ther. Health Med. **16**(4), 36–42 (2010). http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=1&sid=9431f558-0fd6-431c-85eb-b063f76b26c5%40sessionmgr4008
27. Li, W.-C., Chiu, F.-C., Kuo, Y.-S., Wu, K.-J.: The investigation of visual attention and workload by experts and novices in the cockpit. In: Harris, D. (ed.) EPCE 2013. LNCS (LNAI), vol. 8020, pp. 167–176. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39354-9_19
28. McCraty, R., Zayas, M.A.: Cardiac coherence, self-regulation, autonomic stability, and psychosocial well-being. Front. Psychol. **5**, 1–13 (2014). https://doi.org/10.3389/fpsyg.2014.01090
29. McCraty, R., Atkinson, M., Tomasino, D., Bradley, R.T.: The coherent heart heart-brain interactions, psychophysiological coherence, and the emergence of system-wide order. Integr. Rev. **5**(2), 10–115 (2009). http://www.integral-review.org/issues/vol_5_no_2_mccraty_et_al_the_coherent_heart.pdf
30. Edwards, S.D.: Evaluation of heart rhythm coherence feedback training on physiological and psychological variables. South Afr. J. Psychol. **44**(1), 73–82 (2014). https://doi.org/10.1177/0081246313516255
31. Field, L., Edwards, S., Edwards, D., Dean, S.E.: Influence of HeartMath training programme on physiological and psychological variables. Global J. Health Sci. **10**(2), 126–133 (2018). http://eprints.staffs.ac.uk/5956/

Implementation Digital Tower for Apron Control on a Large-Scale of International Airport

Pan Zou[1](✉), Wen-Chin Li[2](✉), and Jingyi Zhang[2](✉)

[1] Norin Intelligent Aviation Software Co. Ltd., Guangzhou, China
zoupan@norinaviation.com

[2] Safety and Accident Investigation Centre, Cranfield University, Cranfield, UK
{wenchin.li,Jingyi.zhang}@cranfield.ac.uk

Abstract. Digital tower has been applied for middle and small sizes of airports for many years. How to adapt digital tower to be used in a large-scale international airport still not be fully investigated. The concept of digital tower operations is that a controller can control any apron from a distant virtual control center. The view of the apron under control is displayed in real time on screens and all traffic movements can be controlled effectively. Therefore, the system designs on digital tower related to human-computer interaction (HCI) and visual presentations which controllers shall be able to identify for safety operations are the main concern of regulator. The research objective is to conduct safety assessment of implementing digital tower for apron control on an international airport. There are six subject-matter experts including four apron controllers and two human factors experts participated in current research. The Hierarchical Task Analysis (HTA) method was used to break down activities, scenarios, and operational steps into each scenario related to apron control in an international airport. The HTA method enables a comprehensive step-by-step description of the task activities associated with each specific scenario related to departure and arrival aircraft and ground vehicle. These operational steps were analysed by human error template (HET) while implementing digital tower for apron control. The result of current research has no safety concern on the safety assessment of digital tower operations from apron controllers' point of view. Designing and managing innovative digital tower on apron control require an understanding of the principles of cognitive systems, allocation of functions and team adaptation between human operators and innovative technology.

Keywords: Apron control · Digital tower · Hierarchical Task Analysis · Human performance · Safety assessment

1 Introduction

At airports, ramp controllers are responsible for the safety and efficiency of the ground movements. Therefore, the effective monitoring of traffic within the controlled zone

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 151–159, 2021.
https://doi.org/10.1007/978-3-030-77932-0_13

by controllers is an important safety mechanism. The paradigm of digital tower operation will allow air traffic services (ATS) be delivered remotely without direct observation from a local tower [1, 2]. Controllers will be supported by an innovative augmented visualization system equipped with pan-tilt-zoom (PTZ) cameras allowing enhanced visual tracking of runway, taxiway, controlled areas on aircraft and vehicle and foreign objects. Based on the concept of digital tower operations, apron control tower on international airport may also get benefits from this innovative technology for capacity, safety, cost efficiency and reduced controllers' workload [3, 4]. The concept of digital tower operations is that a controller can control eastern and western aprons from a distant virtual digital center. The view of the apron under control is displayed in real time on screens and all traffic movements can be controlled effectively. Therefore, the system designs on digital tower related to human-computer interaction (HCI) and visual presentations which controllers shall be able to identify for safety operations are the main concern of regulator [5].

Apron controllers shall be able to perform their tasks using surveillance cameras, data-communication links and the display systems in a remote tower center [2, 6]. The technical evolution of controller working position (CWP) on a remote tower module comprises with Out the Window (OTW) visualization supported by Radar Data Processing (RDP), Electronic Flight Strips (EFS), Information Data Processing (IDP), and a voice communications system (VCS) to provide air traffic information to with a mix of visual flight rules and instrument flight rule [7]. European Union has established remote tower research project to improve safety, minimize costs and environmental impact, and at the same time increase efficiency and capacity in order to meet the requirements of expanding air traffic [1, 8]. Many air navigation service providers (ANSPs) have developed automated systems using video-panorama cameras for synthetic outside view, to increase capacity at airports and to improve cost efficiency by minimizing personnel to meet cost efficiency targets. The remote tower research has been established over the last 20 years [9].

There are some limitations on current apron control in the large-scale international airport including blind spot, distracted multiple information displays and design of automation systems. As the flight volume increases, the negative impact of the complex operating environment on controllers will gradually increase. However, the application of augmented video-panorama technologies has been demonstrated as increasing controller's monitoring ability and situation awareness (SA) during digital tower operations [10]. The technology of digitalized tower can also be applied as contingency tower for large international airports. It offers a more cost-efficient solution than building an additional stand-alone contingency tower [11]. The display of distant aircraft via pan-tilt-zoom (PTZ) cameras enables the images to be augmented superimposed with digital information to reduce controller's head-down time. However, the observable resolution on the human-machine interface (HMI) at the video-panorama may be reduced due to the imperfect system integration on digital tower [12]. Therefore, we have to conduct safety assessment on implementation digital tower for apron control on an large-scale international airport [13].

2 Method

2.1 Participants

Participants consist with qualified Apron Controllers and human factors experts participated in this research. The ages of participants ranged between 25 and 56 years old, and their work experience from 3 to 30 years. The research proposal was submitted to Cranfield University Research Ethics System for ethical approval (CURES/12291/2020). As stated in the consent form, participants are anonymous and have the right to terminate the experiment at any time and to withdraw their provided data at any moment even after the data collection.

2.2 Research Framework

The objective of this research is to conduct safety assessment before implementing digital tower for apron control on a large-scale international airport. The design and development of the digital tower operations can be supported by a formal cognitive work and task analysis. The results of cognitive work and task analysis serve as input data for the simulation of the controller's decision-making processes at the controller working position (CWP) by the Hierarchical Task Analysis (HTA) method. It was used to break down scenarios, goals of activities, and operational steps into each scenario related to apron control [14, 15]. The HTA method enables a comprehensive step-by-step description of the task activities associated with each specific scenario related to departure and arrival aircraft and ground vehicle by apron digital tower center. These operational steps were analyzed by human error template (HET) while controllers using digital tower for apron control. HET is one of the Human Error Identification (HEI) techniques are used to identify potential human or operator error in complex, dynamic systems [6, 16].

2.3 Scenario

The scenarios are including all apron activities relevant to departure, arrival, apron activities, abnormal operations and special environment operations. Based on current standard operational procedures (SOPs) to develop the procedures and guidance for future digital tower operations.

2.4 Data Collection Process

Six subject-matter experts familiar with safety assessment methods and apron control on a large-scale international airport participated focus group sessions. The Hierarchical Task Analysis (HTA) method was used to break down all scenarios and tasks into operational steps. This methodology enables a comprehensive step-by-step description of the task activities associated with the goal of each scenario [17]. The step by step breakdown of apron control on digital tower operations detailed description controller's operational behavior involving human-computer interaction such as interactions with different systems and parties during task performance. All of these 31 operational steps will be analyzed by HET [18]. The main advantages of the HET method are that it is simple to learn and use, requiring very little training and it is also designed to be convenient to apply in a field study. The error taxonomy used is comprehensive as it is based on existing error taxonomies from a number of Human Error Identification (HEI) methods. For each credible error the domain experts should give a description of the form that the error would take. Furthermore, the participants have to determine the outcome or consequence associated with the error and got agreement to the outcome. Finally, the participants have to determine the likelihood of the error (low, medium or high) and the criticality of the error (low, medium or high).

3 Results and Discussion

3.1 Task Decomposition on Apron Control

The results of HTA demonstrated that summed up 98 operational steps consisted by 11 operational steps for Arrival; 31operational steps for Departure; 20 operational steps for Other apron activities; 24 operational steps for Abnormal operation and 12 operational steps for Special environment operations. For example, the detailed decomposition on the scenario of departure for controller's 31 operational behaviors on apron control shown as Fig. 1.

3.2 Application of HET for Error Prediction

The HET matrix is based on the operational steps of HTA which constructed with the vertical-axis assigned as 'likelihood', while the error 'criticality' index was placed on the horizontal-axis. Likelihood and criticality were combined through a multiplication process (likelihood x criticality) to give a 'Pass' or 'Concern' of predicting error related to apron control. A condition determined through the HET to have achieved a likelihood and criticality combination between M and L was assigned as 'Pass', a score between H classified as 'Concern' [18]. An example of operational step 1.2.3 "Issue push-back and start-up command" shown as Table 1. All of those 98 operational steps on HET are passed which meant controllers had no safety concerns on apron control by using digital tower operations.

3.3 Safety Assessment of Apron Control on Digital Tower

The controller of apron digital tower holds the responsibility for the safety and efficiency of aircraft movement and ground transportation activities. Therefore, maintaining the visual attention resources of the controller's working position to monitor the traffic in a controlled area is an important safety mechanism [5, 19]. The safety assessment methodology adapted from Eurocontrol followed a structured step wise process as followings (Fig. 2). The safety plan defines a safety program that is planned, integrated and developed in conjunction with other design, development, production and quality control activities. It details safety activity timelines and deliverable in accordance with the higher project plan. It requires regulatory endorsement and approval. System safety

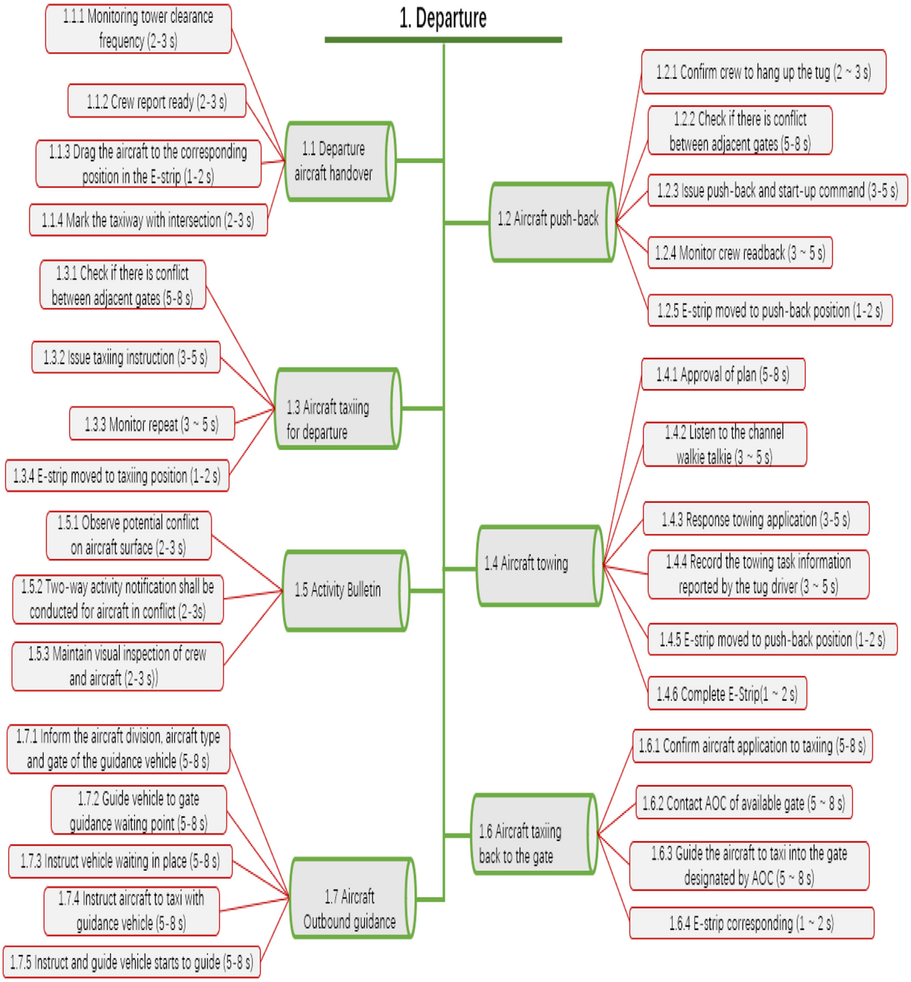
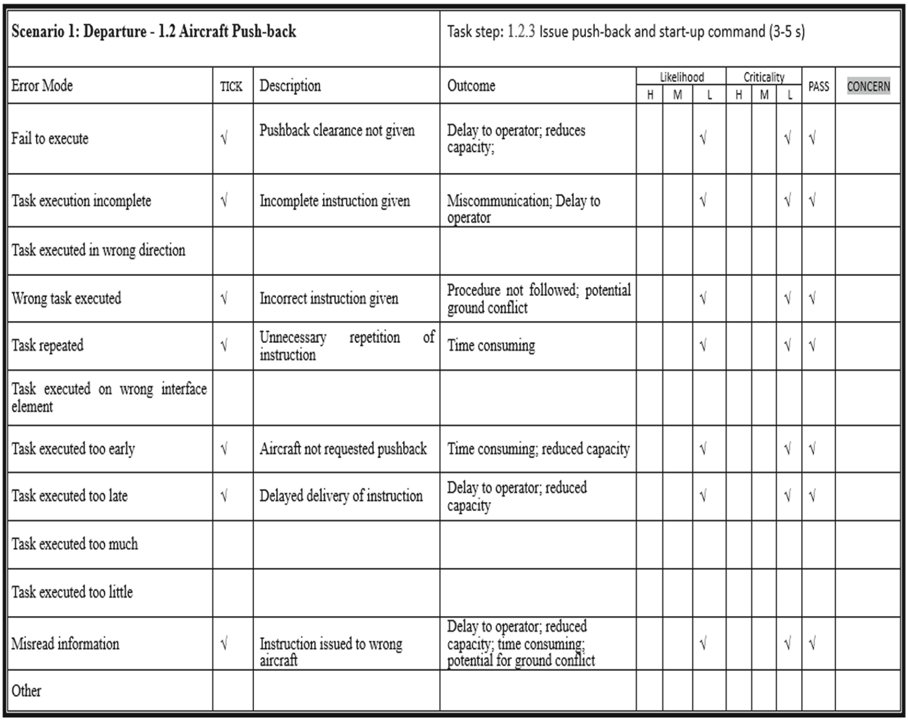

1. Departure
1.1 Departure aircraft handover
1.1.1 Monitoring tower clearance frequency (2-3 s)
1.1.2 Crew report ready (2-3 s)
1.1.3 Drag the aircraft to the corresponding position in the E-strip (1-2 s)
1.1.4 Mark the taxiway with intersection (2-3 s)
1.2 Aircraft push-back
1.2.1 Confirm crew to hang up the tug (2 ~ 3 s)
1.2.2 Check if there is conflict between adjacent gates (5-8 s)
1.2.3 Issue push-back and start-up command (3-5 s)
1.2.4 Monitor crew readback (3 ~ 5 s)
1.2.5 E-strip moved to push-back position (1-2 s)
1.3 Aircraft taxiing for departure
1.3.1 Check if there is conflict between adjacent gates (5-8 s)
1.3.2 Issue taxiing instruction (3-5 s)
1.3.3 Monitor repeat (3 ~ 5 s)
1.3.4 E-strip moved to taxiing position (1-2 s)
1.4 Aircraft towing
1.4.1 Approval of plan (5-8 s)
1.4.2 Listen to the channel walkie talkie (3 ~ 5 s)
1.4.3 Response towing application (3-5 s)
1.4.4 Record the towing task information reported by the tug driver (3 ~ 5 s)
1.4.5 E-strip moved to push-back position (1-2 s)
1.4.6 Complete E-Strip(1 ~ 2 s)
1.5 Activity Bulletin
1.5.1 Observe potential conflict on aircraft surface (2-3 s)
1.5.2 Two-way activity notification shall be conducted for aircraft in conflict (2-3s)
1.5.3 Maintain visual inspection of crew and aircraft (2-3 s))
1.6 Aircraft taxiing back to the gate
1.6.1 Confirm aircraft application to taxiing (5-8 s)
1.6.2 Contact AOC of available gate (5 ~ 8 s)
1.6.3 Guide the aircraft to taxi into the gate designated by AOC (5 ~ 8 s)
1.6.4 E-strip corresponding (1 ~ 2 s)
1.7 Aircraft Outbound guidance
1.7.1 Inform the aircraft division, aircraft type and gate of the guidance vehicle (5-8 s)
1.7.2 Guide vehicle to gate guidance waiting point (5-8 s)
1.7.3 Instruct vehicle waiting in place (5-8 s)
1.7.4 Instruct aircraft to taxi with guidance vehicle (5-8 s)
1.7.5 Instruct and guide vehicle starts to guide (5-8 s)

Fig. 1. The example of Hierarchical Task Analysis (HTA) decomposition the task of departure on digital tower operation in a large-scale International Airport

assessment on the digital tower for apron control has to record the evidence, arguments and assumptions to verify and validate that the system design configuration met safety requirements. It also describes specific operating and maintenance requirements necessary to assure safety and provides arguments to support the claim that the systems will not affect the safety of apron control [8]. The layout of digital tower on apron control centre shown as Fig. 3.

In order to meet the safety metrics requirements, the scenarios and operating procedures of digital tower system must be standardized and verified in detail. This research provides the basis, supplements and assumptions for the safe and sustainable operation of the apron digital tower system and lists all constraints and the scope of activities outlined in the safety assessment plan by subject-matter experts team. Furthermore, the

Table 1. The example of error prediction on the operational step of Push-back on Departure by the HET for Error Mode, Likelihood and Criticality analysis



Scenario 1: Departure - 1.2 Aircraft Push-back			Task step: 1.2.3 Issue push-back and start-up command (3-5 s)								
Error Mode	TICK	Description	Outcome	Likelihood			Criticality			PASS	CONCERN
				H	M	L	H	M	L		
Fail to execute	√	Pushback clearance not given	Delay to operator; reduces capacity;			√			√	√	
Task execution incomplete	√	Incomplete instruction given	Miscommunication; Delay to operator			√			√	√	
Task executed in wrong direction											
Wrong task executed	√	Incorrect instruction given	Procedure not followed; potential ground conflict			√			√	√	
Task repeated	√	Unnecessary repetition of instruction	Time consuming			√			√	√	
Task executed on wrong interface element											
Task executed too early	√	Aircraft not requested pushback	Time consuming; reduced capacity			√			√	√	
Task executed too late	√	Delayed delivery of instruction	Delay to operator; reduced capacity			√			√	√	
Task executed too much											
Task executed too little											
Misread information	√	Instruction issued to wrong aircraft	Delay to operator; reduced capacity; time consuming; potential for ground conflict			√			√	√	
Other											

time frame of each operational step identified in the HTA is under normal operations, it is likely that should a critical event occur or an abnormal pilot request to controllers occurred, there is potential for workload to increase and time pressure to become more acute. The processes of functional hazard assessment and preliminary system safety assessment were taking into accounts in this research. The results of HTA and HET demonstrate that advanced technology integrated with augmented visualization design improved controllers monitoring performance for controlling aircraft from two airports simultaneously. A fundamental principle for the introduction of any new technology is that it must first achieve at least the same level of safety of ATS provision as that which is provided using the traditional physical tower. The analysis of human performance by HET can provide the evidence, arguments and assumptions to support this principle. An important technical requirement is that the visual presentation of aircraft and vehicles by the remote tower system shall not exceed the 1,000 ms of end-to-end delay to fit the requirements of safety assessment. There is a requirement for further research on controller's visual behaviors and operational behaviors related to safety and human performance using digital tower for apron control [13].

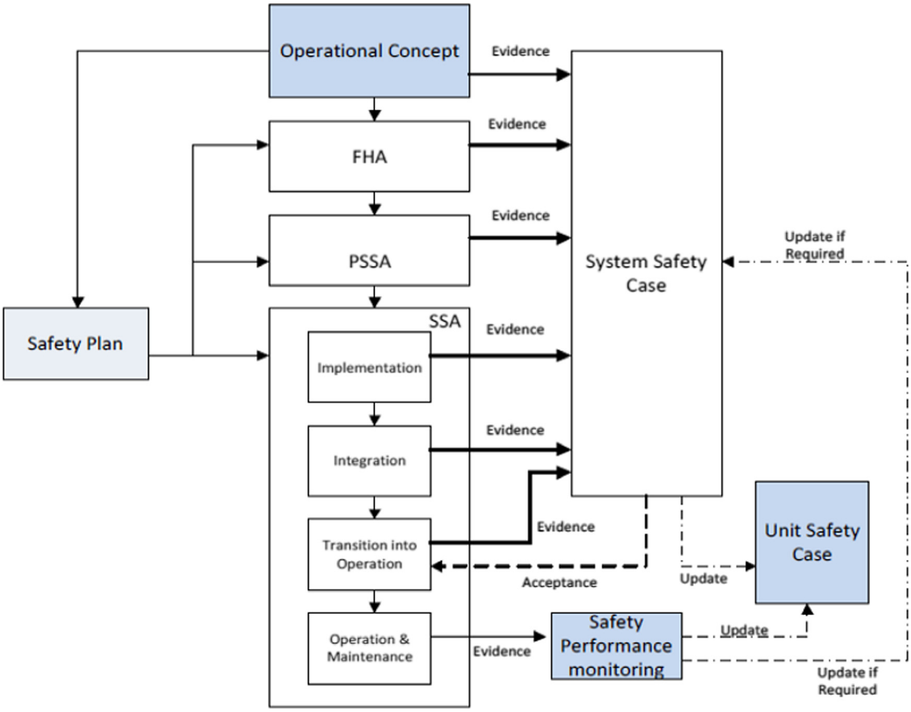


Fig. 2. The processes of safety assessment conducted for digital tower operation in a large-scale International Airport

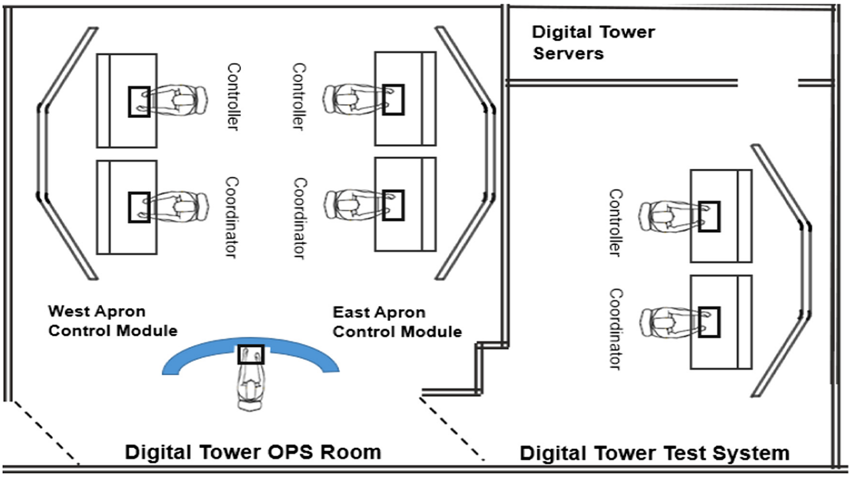


Fig. 3. The layout of digital tower control center including controller and coordinator on each module (left) and testing room (right)

4 Conclusion

Digital tower has brought advantages and benefits to the safety operation on air traffic management. The technical route is relatively mature, and the development direction is basically clear. The overall design scheme mainly includes augmented visualization design on enhanced visual monitoring system, intelligent controller working position and joint digital AOC center construction. Although there are different opinions and concerns either on safety or cost-efficiency for apron operation. The result of current research has no safety concern on the safety assessment of digital tower operations from apron controllers' point of view. The processes of validation and certification before implementation are critical factors to ensure future safety on operations. There are still many questions to be further explored on the application of digital tower in large-scale international airports including how to deal with the complex runway-taxiway structure and apron environment; complex airport surrounding airspace environment; complex weather environment; complex data and system environment; complex operation coordination environment. This creates a need for further research on how to deal with the HCI issues including system integrations and change management for digital tower operations and controllers' training.

Acknowledgements. Authors would like to express special thanks to all apron controllers for their contributions to this research. Their supports and the enthusiasm of their respective teams were invaluable in facilitating the authors' research efforts.

References

1. European Aviation Safety Agency: Term of reference for a rulemaking task: Technical requirements for remote tower operations, Cologne, Germany (2014)
2. van Schaik, F.J., Roessingh, J.J.M., Lindqvist, G., Fält, K.: Detection and recognition for remote tower operations. In: Fürstenau, N. (ed.) Virtual and Remote Control Tower. RTA, pp. 53–65. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28719-5_3
3. Kearney, P., Li, W.C., Zhang, J., Braithwaite, G., Wang, L.: Human performance assessment of a single air traffic controller conducting multiple remote tower operations. Hum. Factors Ergon. Manuf. Serv. Ind. **30**(2), 114–123 (2020). https://doi.org/10.1002/hfm.20827
4. Van Beek, S.D.: Remote towers: a better future for America's small airports (2017). https://reason.org/policy-study/remote-air-traffic-control-towers-a-better-future-for-americas-small-airports/. Accessed 11 July 2017
5. Li, W.C., Kearney, P., Braithwaite, G., Lin, J.J.: How much is too much on monitoring tasks? Visual scan patterns of single air traffic controller performing multiple remote tower operations. Int. J. Ind. Ergon. **67**, 135–144 (2018). https://doi.org/10.1016/j.ergon.2018.05.005
6. Kearney, P., Li, W.C., Yu, C.S., Braithwaite, G.: The impact of alerting designs on air traffic controller's eye movement patterns and situation awareness. Ergonomics **62**(2), 305–318 (2019). https://doi.org/10.1080/00140139.2018.1493151
7. Ellis, S.R., Liston, D.B.: Visual features used by airport tower controllers: some implications for the design of remote or virtual towers. In: Fürstenau, N. (ed.) Virtual and Remote Control Tower. RTA, pp. 21–51. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28719-5_2

8. Kearney, P., Li, W.C.: Multiple remote tower for Single European Sky: the evolution from initial operational concept to regulatory approved implementation. Transp. Res. Part A: Policy Pract. **116**, 15–30 (2018). https://doi.org/10.1016/j.tra.2018.06.005
9. Leitner, R., Oehme, A.: Planning remote multi-airport control—Design and evaluation of a controller-friendly assistance system. In: Fürstenau, N. (ed.) Virtual and Remote Control Tower. RTA, pp. 139–160. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28719-5_7
10. Ziegler, J.: Cut the cost: how can remote tower and virtual tower solutions improve the efficiency and reduce the operational costs of ATC services for regional airports? Int. Rev. Air Traffic Technol. Manag. **2017**, 1–3 (2017)
11. SESAR Joint Undertaking: Human performance assessment report of Remote and Virtual Tower. 06-09-03, Brussels, Belgium (2016)
12. Fürstenau, N., Schmidt, M.: Remote tower experimental system with augmented vision videopanorama. In: Fürstenau, N. (ed.) Virtual and Remote Control Tower. RTA, pp. 163–192. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28719-5_8
13. European Aviation Safety Agency: Guidance material on the implementation of the remote tower concept for single mode of operation. Cologne, Germany (2015)
14. Stanton, N.A.: Hierarchical task analysis: developments, applications and extensions. Appl. Ergon. **37**(1), 55–79 (2006). https://doi.org/10.1016/j.apergo.2005.06.003
15. Kearney, P., Li, W.-C., Braithwaite, G., Greaves, M.: The investigation human-computer interaction on multiple remote tower operations. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10275, pp. 301–309. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58472-0_23
16. Kirwan, B.: Human error identification in human reliability assessment. Part 1: overview of approaches. Appl. Ergon. **23**(5), 299–318 (1992). https://doi.org/10.1016/0003-6870(92)90292-4
17. Annett, J.: Hierarchical task analysis. In: Stanton, N.A., Hedge, A., Brookhuis, K., Salas, E., Hendrick, N.A. (eds.) Handbook of Human Factors and ergonomics methods, pp. 329–337. CRC Press, Boca Raton (2004)
18. 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, Farnham (2013)
19. Irish Aviation Authority: Remote Towers Demonstration Report: LSD 02.04. Dublin, Ireland (2016)

Studies on Cognitive Processes

How Metacognitive Monitoring Feedback Influences Workload in a Location-Based Augmented Reality Environment

Wenbin Guo(✉) and Jung Hyup Kim

Department of Industrial and Manufacturing Systems Engineering, University of Missouri, Columbia, USA
wgk95@mail.missouri.edu, kijung@missouri.edu

Abstract. This research aims to investigate the impact on workload caused by metacognitive monitoring feedback (MCMF) in a location-based augmented reality (AR) learning environment. MCMF helps learners to monitor and control their cognitive processes and influences their learning behaviors. However, it should be studied further how MCMF affects student workload while using the AR system. In this study, we conducted an experiment to compare perceived mental workload between two groups (with MCMF vs. without MCMF). The results show that MCMF does not increase students' workload. It means that MCMF can be used effectively without workload increment while learning. The current study advanced our understanding of metacognitive strategies on subject interaction in a location-based AR environment. Furthermore, the study outcomes could develop better metacognitive strategies without increasing learners' workload in the AR environment.

Keywords: Augmented reality · Engineering education · Learning performance

1 Introduction

Metacognition has been studied in many research [1–3]. These studies suggest that metacognitive strategies might help students monitor and control their learning processes in an augmented reality (AR) environment. Metacognitive prompting could be defined as a specific question, action, or thought to learners that they are cognitively activated or reflected on their metacognition [4]. This study used retrospective confidence judgments as a metacognitive prompting to assess the likelihood of students' responses corresponding to a test [5]. Retrospective confidence judgments represent students' confidence levels about the learning before knowing their outcomes. On the other hand, metacognitive monitoring feedback [6, 7] is a tool to calibrate students' understanding of their learning and judgments. There are pros and cons to metacognitive monitoring techniques. One of the possible drawbacks is a significant increment in mental workload. Galy, Cariou [8] found that mental workload could be related to students' learning strategies and a conscious search for patterns in learning material. Hence, in this study,

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 163–171, 2021.
https://doi.org/10.1007/978-3-030-77932-0_14

we investigated the workload impact of metacognitive monitoring feedback (MCMF) in an AR learning environment. NASA Task Load Index [9, 10] has been used to measure subjective workload in many studies. In this study, Students' mental workload with MCMF was measured by using the NASA-TLX. Then, the hypothesis was presented.

Hypothesis: Metacognitive monitoring feedback (MCMF) influences students' perceived workload in location-based AR learning environments.

The NASA-TLX contains ratings and weights. Ratings have six dimensions of experience evaluation from subjects. The numerical range of each is from 1 to 100 (low to high). When performing a task in a pair of options, weights represent the most significant contributor to human workload. Six dimensions constitute 15 pairs of weights, and the number of weights is from 1 (not relevant) to 5 (most relevant). The overall workload depends on the weighted mean of ratings and weights [10].

NASA-TLX measures the overall human workload while immediately after completing a task [9]. It has been applied in various areas, including military simulation, personal driving, complex air traffic control [11–16]. Much research exhibited NASA-TLX is a valuable tool to use for workload measurement. Six dimensions (mental, physical, temporal, performance, effort, and frustration) represent independent clusters of variables to assess workload. Mental, physical, and temporal demands relate to the subject dimension, while effort, frustration, and performance interact with a given task.

Using NASA-TLX, this study will advance our understanding of how MCMF influences workload and student engagement in a location-based AR learning environment. In addition, this research outcome could be used to develop better MCMF without increasing learners' workload in a location-based AR environment.

2 Related Works

A previous study found that AR impacts students' workload on mental, performance, effort, and frustration dimensions in engineering education [17]. Understanding learners' workload in an AR environment helps researchers design appropriate AR content in various areas to improve learning performance. Metacognition refers to the human ability to use metacognitive knowledge to monitor and control a cognitive process [18, 19], which includes three components: knowledge related to cognition, metacognitive monitoring, and control [20, 21]. Most metacognition research focuses on monitoring and controlling [22]. Monitoring involves the awareness of learner's cognitive processes—a flow of cognitive information from cognition to metacognition. Researchers used metacognition to improve student learning in different disciplines [23].

Metacognition has been applied to influence cognitive activities in an AR environment. According to the influence of metacognition on learning processes, metacognitive prompting is provided to collect students' metacognitive judgment, and metacognitive monitoring feedback (MCMF) was developed as a tool to monitor and calibrate their metacognitive judgments. Bannert found that metacognitive prompting did not increase workload in the computer-based learning environments [24], while it is not verified in an AR environment. Fiorella utilized Cognitive Load Theory to investigate the relationship between metacognition and workload [22]. The researchers found that phrase-based promptings might influence learners' workload, and they were distracted by additional information from the metacognitive promptings. If MCMF increases learners' workload in the AR learning system, it would not be beneficial to implement when students are under a high workload condition.

3 Methods

3.1 Apparatus

HoloLens (see Fig. 1) is a prominent augmented reality device with Windows 10 operating system, which is used to create biomechanics AR content in the experiment. HoloLens is an independent mobile computer, and students are entirely free to move around to observe learning content in the experiment area. This device can mix real and virtual world objects and present them in front of the users. Students gaze at the targets they are observing. Several human gestures can be used to connect with virtual objects to understand the learning content better.

Fig. 1. Microsoft HoloLens

3.2 Participants

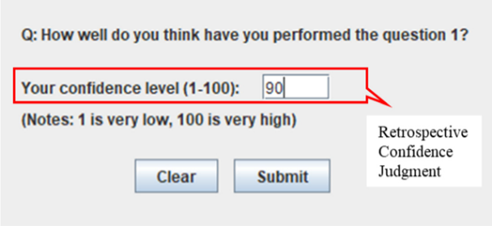
A total of forty-two students from the University of Missouri with an average age of 21.91 years (StDev = 3.51, age range from 19 to 39) participated in the experiment, consisting of twenty-eight male and fourteen female students. The research was conducted in two conditions, as shown in Table 1. Group 1 students learned biomechanics in a location-based AR environment without MCMF. Group 2 students had MCMF in the AR environment.

Table 1. Group description

Group	Retrospective Confidence Judgments (RCJs)	Metacognitive Monitoring Feedback (MCMF)
Group 1 (n = 16)	Yes	No
Group 2 (n = 26)	Yes	Yes

3.3 Metacognitive Monitoring Techniques

The retrospective confidence judgment is one of the metacognitive prompting techniques to self-evaluate the learner's confidence levels before knowing actual test scores. Figure 2 shows an example of retrospective confidence judgment in the question, "How well do you think have you performed the question 1? (1% - low confidence level, 100% - high confidence level)". Figure 3 presents an example of metacognitive monitoring feedback at the end of the question. It displays the comparison of confidence levels and actual test scores. Metacognitive techniques would influence the students' confidence judgments, changing their learning behaviors, and progress.



Q: How well do you think have you performed the question 1?

Your confidence level (1-100): 90

(Notes: 1 is very low, 100 is very high)

Retrospective Confidence Judgment

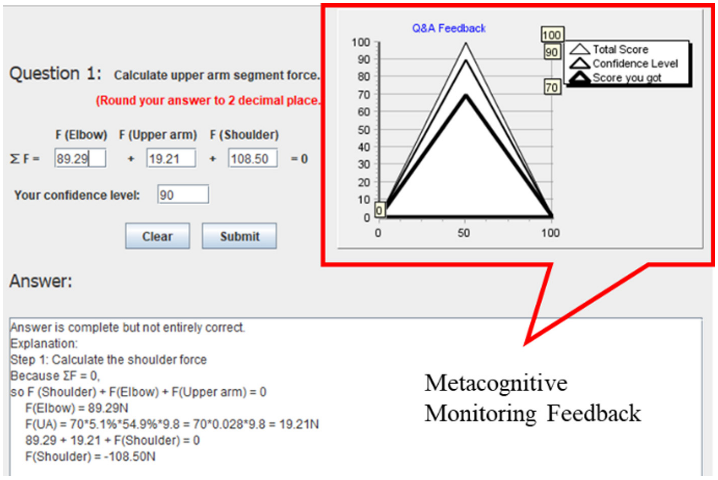
Fig. 2. Retrospective confidence judgment

3.4 Learning Content

This study used biomechanics as the AR learning content for engineering students (Fig. 4). Biomechanics knowledge showed the complexity of learning with the internal and external forces and moments acting on body segments. Accordingly, the AR environment might help students recognize spatial visualization and be beneficial for spatial awareness.

3.5 Experiment Design

There were two groups in this study. As a control group, the participants in Group 1 received RCJs and biomechanics test questions, but the feedback screen only included the biomechanics questions and correct answers without MCMF. The participants in Group 2 were asked RCJ probes and biomechanics test questions, and they viewed the feedback screen with the correct answers of the biomechanics questions with the MCMF.



Question 1: Calculate upper arm segment force.

(Round your answer to 2 decimal place.

	F (Elbow)		F (Upper arm)		F (Shoulder)	
Σ F =	89.29	+	19.21	+	108.50	= 0

Your confidence level: 90

Answer:

Answer is complete but not entirely correct.
Explanation:
Step 1: Calculate the shoulder force
Because ΣF = 0,
so F (Shoulder) + F(Elbow) + F(Upper arm) = 0
F(Elbow) = 89.29N
F(UA) = 70*5.1%*54.9%*9.8 = 70*0.028*9.8 = 19.21N
89.29 + 19.21 + F(Shoulder) = 0
F(Shoulder) = -108.50N

Fig. 3. Metacognitive monitoring feedback

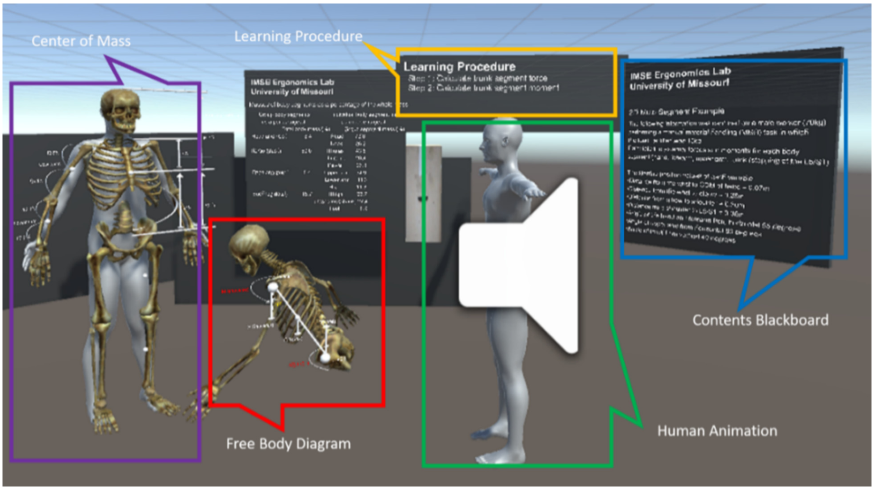


Fig. 4. Biomechanical AR module

Figure 5 shows the experiment flowchart. Group 1 students engaged in biomechanics learning without metacognitive monitoring feedback in the location-based AR environment. Group 2 students experienced the biomechanics module in AR environments with metacognitive monitoring feedback. The difference between groups 1 and 2 was the metacognitive monitoring feedback in the feedback screen.

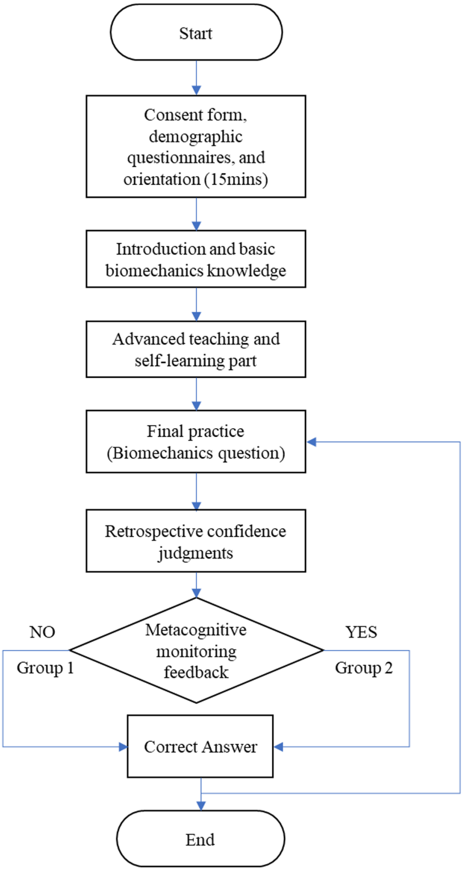

Start
Consent form, demographic questionnaires, and orientation (15mins)
Introduction and basic biomechanics knowledge
Advanced teaching and self-learning part
Final practice (Biomechanics question)
Retrospective confidence judgments
Metacognitive monitoring feedback
NO
Group 1
YES
Group 2
Correct Answer
End

Fig. 5. Experiment flowchart

4 Results

Workload differences between two groups were analyzed using ANOVA. No significant difference in workload was found among Group 1 and Group 2. Figure 6 shows the interval plot of workload on six dimensions among the two groups. The results show no significant difference for the overall workload [F (1, 41) = 0.34, p = 0.563], mental dimension [F (1, 41) = 0.56, p = 0.458], physical dimension [F (1, 41) = 0.03, p = 0.865], temporal dimension [F (1, 41) = 0.62, p = 0.436], performance dimension [F (1, 41) = 0.31, p = 0.579], effort dimension [F (1, 41) = 0.28, p = 0.600], and frustration dimension [F (1, 41) < 0.01, p = 0.994].

5 Discussion and Conclusion

According to the results, this study found that MCMF did not increase students' workload. It means that this feedback screen can be used without workload increment in

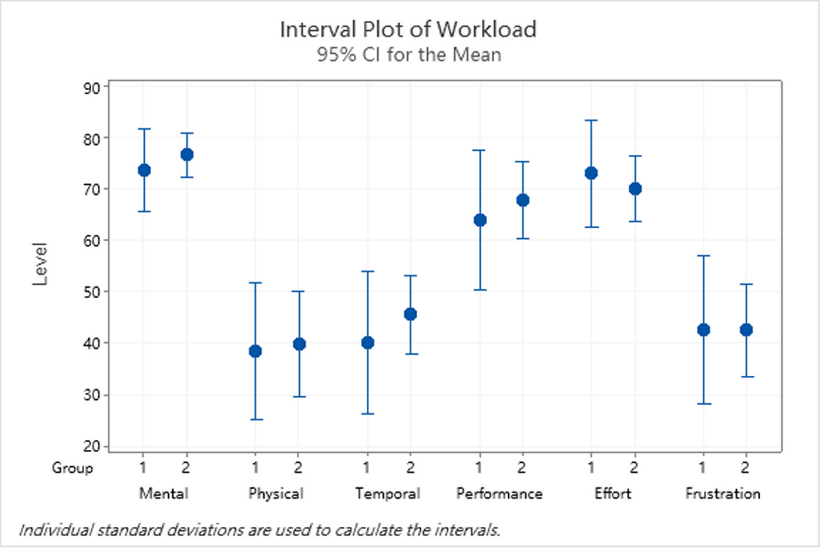


Fig. 6. The workload in six dimensions

an AR learning environment. The study advanced our understanding of metacognitive strategies on subject interaction in a location-based AR environment. The current outcomes could also be used to develop better metacognitive strategies without increasing learners' workload in the AR environment.

As we mentioned, mental demand measures perceptual activity such as thinking, deciding, and cognitive task demanding. In our study, retrospective confidence judgments support students to be aware of which learning contents need to be restudied and what problems they should practice more. The students who experienced MCMF were willing to restudy some of the AR modules based on the confidence level. They were more likely to restudy the modules when their retrospective confidence levels were low, which helped them improve their performance. Retrospective confidence judgment is one of the metacognitive prompting strategies detecting students' confidence level. According to Fiorella and Vogel-Walcutt [22], metacognitive prompting does not affect the acquisition of declarative and general knowledge, and it does not affect students' workload. On the contrary to this, procedural knowledge might increase workload in learning [25]. MCMF compares the confidence levels to actual test performance instead of influencing the procedural steps during AR learning. Therefore, there was no workload increment on the mental dimension caused by MCMF in a location-based AR environment.

Temporal demand measures learners' time pressure due to the experiment. MCMF might influence students' learning time and reviewing answers time because it continuously affects student metacognitive states (i.e., over or under-confident states). During the experiment, students altered their metacognitive states and the amount of time they spent on studying after receiving MCMF. However, a low degree of procedural knowledge was involved in the AR learning modules. For that reason, MCMF did not influence learners' workload in the temporal dimension.

The performance dimension in NASA-TLX assesses the level of success in completing a given task. Based on the results, we found that there was no significant impact on the performance dimension of students who experienced MCMF compared to the control group.

Physical demand in NASA-TLX measures the amount of physical activities required to complete the experiment. The results demonstrated that MCMF did not generate any additional physical intensity during the experiment. The effort dimension measures how hard the students have to work to maintain their level of performance. The results also showed that MCMF did not influence the effort level. Lastly, the frustration dimension measures how insecure and discouraged students felt during the experiment. The results presented that MCMF did not increase the frustration. Overall, MCMF did not increase the students' perceived workload in the location-based AR learning environment.

6 Limitations and Future Work

Although metacognition has already shown the potential for improvements in various fields, many issues should be resolved before MCMF apply to the AR learning environment and optimize existing AR applications. Accordingly, metacognitive skills need to be introduced to students to improve their learning processes. This study only tested MCMF in engineering learning content. It proved that the feedback screen helped to improve learning performance and to calibrate students' confidence levels without increasing their workload.

In the future, the AR learning system can be improved further with a better MCMF. This advanced feedback screen indicates the current confidence level and the cumulative confidence level. The continuous visual feedback offers a more apparent change in students' confidence levels. The parameters, such as the average confidence levels, actual performance comparisons, and descriptive statistics, can be provided to students and let them know better where they are now in terms of learning. Based on the feedback, the system might predict students who are highly motivated learners. Further, students will be able to regulate better their strategies to learn new knowledge in a location-based AR environment.

References

1. Hacker, D.J., Dunlosky, J., Graesser, A.C.: Handbook of Metacognition in Education. Routledge, London (2009)
2. Dunlosky, J., Tauber, S.U.K.: The Oxford Handbook of Metamemory. Oxford University Press, Oxford (2016)
3. Shaughnessy, M.F., Veenman, M., Kennedy, C.K.: Meta-Cognition: A Recent Review of Research, Theory, and Perspectives. Nova Publishers, New York (2008)
4. Evans, J.R., Fisher, R.P.: Eyewitness memory: balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. Appl. Cogn. Psychol. **25**(3), 501–508 (2011)
5. Schraw, G.: A conceptual analysis of five measures of metacognitive monitoring. Metacogn. Learn. **4**(1), 33–45 (2009)

6. Kim, J.H.: The effect of metacognitive monitoring feedback on performance in a computer-based training simulation. Appl. Ergon. **67**, 193–202 (2018)
7. Kim, J.H.: The impact of metacognitive monitoring feedback on mental workload and situational awareness. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 32–41. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_3
8. Galy, E., Cariou, M., Mélan, C.: What is the relationship between mental workload factors and cognitive load types? Int. J. Psychophysiol. **83**(3), 269–275 (2012)
9. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Sage Publications (2006)
10. Cao, A., Chintamani, K.K., Pandya, A.K., Ellis, R.D.: NASA TLX: software for assessing subjective mental workload. Behav. Res. Methods **41**(1), 113–117 (2009). https://doi.org/10.3758/BRM.41.1.113
11. Caldwell, J.A.: Fatigue in aviation. Travel Med. Infect. Dis. **3**(2), 85–96 (2005)
12. Wiegmann, D.A., Shappell, S.A.: Human error analysis of commercial aviation accidents: application of the Human Factors Analysis and Classification System (HFACS). Aviat. Space Environ. Med. **72**(11), 1006–1016 (2001)
13. Palinko, O., et al.: Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. ACM (2010)
14. Lauer, T.R., et al.: The Masses of Nuclear Black Holes in Luminous Elliptical Galaxies and Implications for the Space Density of the Most Massive Black Holes Based on observations made with the NASA/ESA Hubble Space Telescope, obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555. These observations are associated with GO and GTO proposals 5236, 5446, 5454, 5512, 5943, 5990, 5999, 6099, 6386, 6554, 6587, 6633, 7468, 8683, and 9107. Astrophys. J. **662**(2), 808 (2007)
15. Hwang, S.-L., et al.: Application control chart concepts of designing a pre-alarm system in the nuclear power plant control room. Nucl. Eng. Des. **238**(12), 3522–3527 (2008)
16. Erzberger, H.: Automated conflict resolution for air traffic control (2005)
17. Guo, W., Kim, J.H.: How augmented reality influences student workload in engineering education. In: Stephanidis, C., et al. (eds.) HCII 2020. LNCS, vol. 12425, pp. 388–396. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60128-7_29
18. Flavell, J.H.: Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. Am. Psychol. **34**(10), 906 (1979)
19. Nelson, T.O.: Metamemory: a theoretical framework and new findings. In: Psychology of Learning and Motivation, pp. 125–173. Elsevier (1990)
20. Livingston, J.A.: Metacognition: An Overview (2003)
21. Dunlosky, J., Metcalfe, J.: Metacognition. Sage Publications, Thousand Oaks (2008)
22. Fiorella, L., Vogel-Walcutt, J.J.: Metacognitive prompting as a generalizable instructional tool in simulation-based training. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications (2011)
23. Rhem, J.: Using Reflection and Metacognition to Improve Student Learning: Across the Disciplines, Across the Academy. Stylus Publishing, LLC (2013)
24. Bannert, M., et al.: Short-and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. Comput. Hum. Behav. **52**, 293–306 (2015)
25. Schraw, G., Moshman, D.: Metacognitive theories. Educ. Psychol. Rev. **7**(4), 351–371 (1995)

Effect of Height Perception on State Self-Esteem and Cognitive Performance in Virtual Reality

Grace Y. S. Leung[1,2](✉), Adrian K. T. Ng[1], and Henry Y. K. Lau[1]

[1] Department of Industrial and Manufacturing Systems Engineering,
The University of Hong Kong, Pokfulam, Hong Kong
{u3546662,adriang}@connect.hku.hk, hyklau@hku.hk

[2] Department of Psychology, The University of Hong Kong, Pokfulam, Hong Kong

Abstract. Tall stature has long been regarded as a socially desirable asset. Past literature has uncovered a positive correlation between height and cognitive ability, with evidence indicating that subjective evaluations of self-worth are highly susceptible to changes in body image. In this study, we aimed to investigate whether implementing a deliberate increase in height perception can affect individuals' state self-esteem, working memory, and visuospatial competency. Adapting the theoretical model of height-career success relationship, it is hypothesized that height alteration would lead to higher levels of positive self-appraisals, as well as better objective and subjective cognitive performance. In this pilot study, we leveraged an immersive virtual reality system to manipulate individuals' height. Participants conducted a series of letter recall and mental rotation tests, estimated their cognitive performance, and completed the State Self-Esteem Scale in a virtual environment. While improvements in working memory and spatial skills were observed in the increased height condition, ratings of self-perceived performance did not significantly differ. Among all facets of self-esteem, only appearance self-esteem showed a significant increase under the influence of height enhancement. This may suggest that benefits of taller height perceptions are sensitive to stature-related self-evaluations, even in the absence of external appraisals.

Keywords: Height · Virtual reality · Spatial ability · Mental rotation · Working memory · Self-esteem

1 Introduction

Across human history, societal bias in favor of tall stature has transcended cultural boundaries and become imperceptibly normalized in regular judgement and decision-making processes. Numerous studies have indicated that height is positively associated with leadership emergence, which can be further mediated by interpersonal dominance, vitality, and intelligence [2, 11]. From the evolutionary perspective, physically formidable organisms tend to be perceived as more resourceful in holding power, which in turn leads to higher self-esteem and contributes to the success in achieving a higher social rank within groups [13].

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 172–184, 2021.
https://doi.org/10.1007/978-3-030-77932-0_15

The impact of height-associated benefits is also well-manifested in people's socioeconomic status. Past literature has suggested that teenage height may serve as a promising predictor of success in adulthood, as taller individuals are typically endowed with better educational attainment, positive employment prospects, and higher earnings [8, 22]. According to Persico et al. [19], even in the same occupational field, the tallest quarter of the population may find themselves in more superior job positions. Moreover, taller workers in general enjoy an extra thirteen percent salary when compared to those in the 25th percentile, demonstrating a conspicuous wage disparity. It is clear that height outweighs a considerable variety of individual characteristics and serves as an important indicator for evaluation of capabilities and success.

Despite the strong preference for tall statures in societal settings, the reasoning behind this height premium remains a widely debated controversy. A number of studies have argued that tall children are on average more intelligent and thus more likely to outperform their peers in cognitive assessments [27]. The proven association between height and cognitive ability is evident throughout the lifespan. In particular, it was found that taller populations, who presumably possess stronger cognitive skills, tend to pursue more intellectually stimulating professions and receive higher earnings, rather than lending themselves to physically strenuous jobs that are lower-paying in general [7].

On the other hand, Judge and Cable's [12] theoretical model of the height–career success relationship examined a sequence of mediators that may account for the advantages of being tall. To start with, the model emphasizes the impact of height on self-esteem, which is defined as individual's overall sense of self-worth and competence [5]. Past literature has uncovered a profound relationship between subjective evaluations of height and self-worth. This association was found to be mediated by the effect of self-consciousness, which illustrates the significance of perception and reflection in formation of body image disturbance [3, 19]. Self-appraisals owing to height influences can also be mediated by social esteem, which refers to judgmental inferences directed from others to oneself, or in the case of height-related research, a construct operationalized as perceived stature [21]. While self-esteem is generally regarded as an endurable trait, its susceptibility to environmental stimuli often leads to momentary fluctuations in daily interactions.

Under the influence of rapid technological advancements in recent decades, mass media has played a critical role in transmitting societal perceptions of idealized body image, reinforcing the notion that physical attractiveness is linked to tall stature [9]. In particular, interactive digital platforms have further intensified public engagement in maladaptive self-evaluation processes. Body image dissatisfaction is directly correlated to appearance-related social comparisons [26]. Previous work has shown that shorter individuals are susceptible to suffering from low self-esteem, which would sequentially exert a negative influence on their social competence and psychological well-being [25]. Furthermore, self-esteem influenced by body morphology ideals would affect both objective and subjective performance, thus mediating the relationship between appearance and workplace success in terms of earnings and leadership emergence [12].

While theories accounting for benefits of being tall have been well established, limited attention has been paid to the effect of height-related changes on an individual level. If tall stature is associated with psychological and social advantages, would doses of height alteration experiences serve as a potential intervention for individuals suffering from unfavorable consequences of having a short stature?

One safe and convenient way of answering this question would be to utilize the novelty of immersive virtual reality (VR) for inducing altered body perceptions through lifelike immersions [32]. The integration of this advanced technology into clinical interventions and experimentation has received rising popularity in the field of mental health and clinical psychology (e.g., [24]). To date, increasing number of studies on body morphology manipulation have utilized digitalized virtual avatars to examine body-representation and its associated well-being effects. In particular, past literature has found that self-perceptions of virtual representations are correlated with self-esteem and body esteem in adolescence [29]. Many considered variables other than stature, for instance, weight and body shape, to develop preventive interventions and treatments for patients suffering from obesity and eating disorders [20, 23]. Manipulation of height in VR was tested in research that focused on height rather than other body ideations, which suggest that deliberate reduction in a person's height would lead to more negative self-evaluations and increased levels of paranoia [6]. Implementing direct alteration of an individuals' stature can ensure strict control for potential confounding effects of other indicators of body morphology, which may be prevalent in the creation of self-avatars.

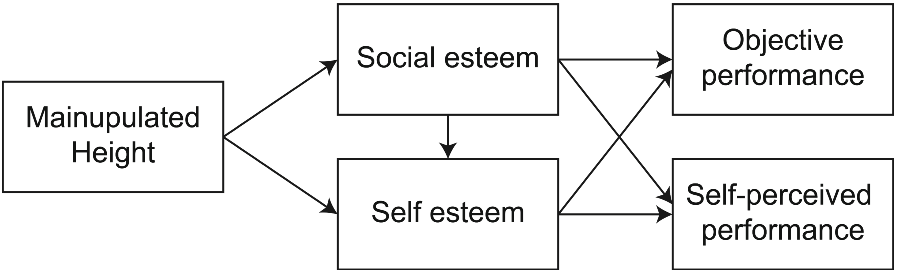


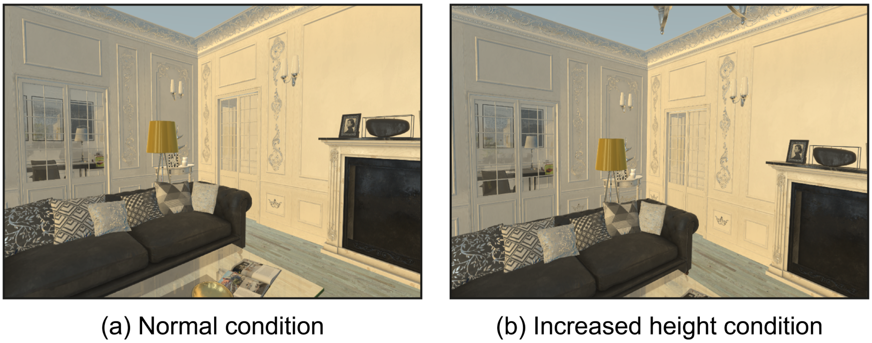
Fig. 1. Height-performance relationship theoretical model, adapted from Judge and Cable [12]

Here, we aim to investigate whether deliberate manipulation of a person's height in VR would influence one's self-evaluation and competence in cognitive assessments. Following the general theoretical basis of Judge and Cable's [12] model of the physical height and career success, we developed a hypothesized model of height-performance relationship. The model suggests that virtual height alternation would enhance individual's esteem, which would in turn lead to better objective and self-perceived performance (Fig. 1). (Note that adapted from Judge and Cable [12] model, social esteem and self-esteem are both mediating factors, however, we are not investigating social esteem in the current pilot study.) It is hypothesized that when compared to the initial stature, individuals with increased height would elicit higher levels of self-esteem, in particular, height-related appearance appraisals. In addition, we predict that when given a taller virtual stature, participants would demonstrate improved memory and spatial skills and hence better cognitive performance in both objective and subjective measures. To test these hypotheses, we invited healthy volunteers to experience an immersive VR system under differing height perceptions in order to compare their levels of self-esteem, memory skills and spatial ability.

2 Methods

2.1 Participants

Twelve healthy volunteers were recruited at the University of Hong Kong to take part in this pilot study. Two participants were excluded from the data analyses due to technical complications. Thus, the final sample size consisted of ten Chinese participants (7 females, $M_{age} = 26.6$ years, $SD_{age} = 5.62$ years, age range: 20–35 years), with body height ranging from 154 cm to 178 cm ($M = 161.9$ cm, $SD = 8.80$ cm). All participants had adequate stereopsis, normal or corrected-to-normal vision, and no cognitive impairment. Participants gave written informed consent and received payment for their participation. They were free to withdraw at any point of the study without negative consequences. The study was approved by the local Research Ethics Committee of the University of Hong Kong.


(a) Normal condition

(b) Increased height condition

Fig. 2. Sample view of the virtual scene in (a) normal condition (control) and (b) increased height condition (30 cm increased)

2.2 Apparatus

The experiment was conducted using a head mounted display (HMD). The HMD (VIVE Pro; HTC, New Taipei City, Taiwan) included separate 3.5″ displays for each eye (with measured luminance of 143 cd/m^2), each with a resolution of 1440 × 1600 pixels at a refresh rate of 90 Hz, yielding a 110° horizontal field of view. Fresnel lens is located in front of the displays to render visual content at optical infinity. Head tracking was facilitated by combining internal accelerometer and gyroscope, and external infrared tracking data. The display was connected to a 64-bit Windows 10 PC with an Intel® Core™ i7-7700K CPU @4.20 Ghz processor, 16 GB Ram, and a NVIDIA® GeForce® RTX

2080 Ti graphics card. We presented a virtual apartment with realistic furniture (Fig. 2a) developed in Unity (version 2018.4.6f1; Unity Technologies, San Francisco, CA, United States) and rendered using SteamVR (Valve, Bellevue, WA, United States). The viewpoint tracking system updated the location and orientation of the virtual stereoscopic camera to provide lifelike VR rendering [14].

In the experimental condition, we increased participants' height by 30 cm (Fig. 2b), which was the approximate size of a head, with regard to all aspects of the virtual experience. Except the perceptual manipulation, an identical virtual scene was leveraged for both conditions. Throughout the experiment, when required, digitized virtual screen displays (70 cm × 70 cm) were shown in front of the participants at eye level in order to present the cognitive task stimuli and instructions in the virtual apartment.

2.3 Experimental Design

The independent variable of interest was virtual height manipulation. Performance in letter recall and mental rotation tasks, denoted by the percentage of correct response, as well as participants' level of self-esteem serve as the outcome variables of the study.

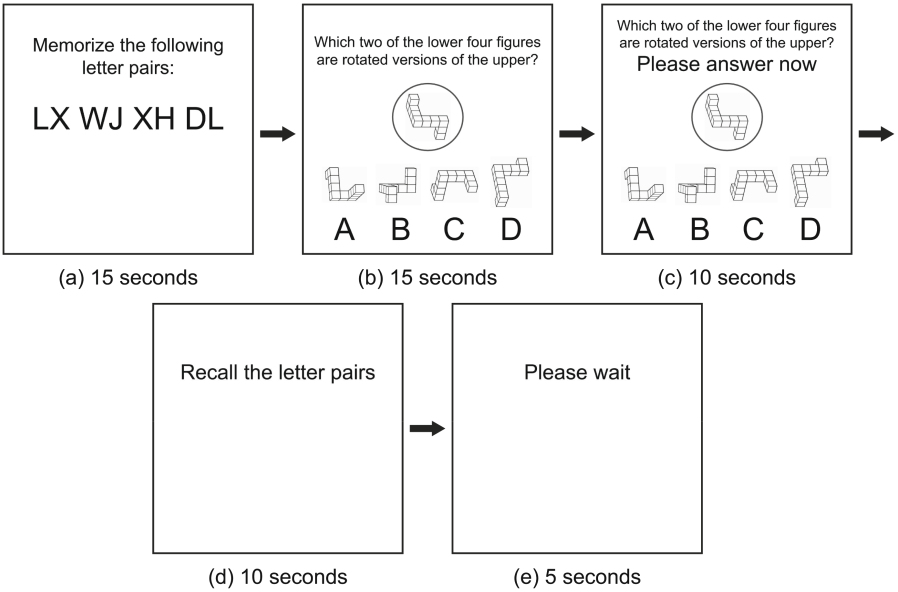
2.4 Procedure

Upon arrival to the laboratory, participants provided written informed consent to participate in the research study. The experimenter explained the cognitive test by presenting an example of the letter pairs and spatial rotation problem on a computer monitor. The participant did not practice the given task but were asked if they understood the procedures involved. In addition, they were asked to complete a short demographic survey and a baseline SSES questionnaire. The experimenter then introduced the apparatus and assisted the participants in putting on the HMD. In the virtual environment, a practice trial of the cognitive test was utilized to ensure that all participants understood the timing of each experimental stage. Feedback on response accuracy was not provided. However, participants could ask clarification questions and were prompted to respond if they did not do so in the designated task phase.

Each participant took part in two blocks of trials based on random assignment: the normal height condition and the increased height condition. Each block started off with a 2-min VR exposure, which featured an interactive object search task to elicit self-conscious comparison between the given environmental cues and one's physical height, so as to consolidate the novel height perception in this illusion of reality. They were also asked to walk around and search for objects hidden in the virtual scene. This interactive task was immediately followed by 10 randomized trials of cognitive tests, which included both letter recall and mental rotation tasks. Then, participants were asked to complete the self-esteem questionnaire and estimate their performance. There was a 5-min break between conditions where the HMD would be removed from the participants for reset of VR height manipulation. The duration of time required to complete all assessments was approximately 35 to 40 min. Throughout the experimental process, participants were not informed of the height alteration in the VR system and remained naïve to the aim of the study. We manipulated the height perception in the VR program before participants put on their HMD to avoid a sudden shift of visual perspective.

2.5 Assessments

Cognitive Tests. To evaluate the objective performance of participants under normal and heightened perceptions, a series of cognitive tests were adopted based on Steed et al.'s [28] experimental framework. We integrated letter recall and mental rotation tasks to assess individuals' cognitive abilities in terms of memory skills and spatial functioning respectively. In this study, a total of 20 independent trials were equally divided into two blocks through random assignment. The task stimuli and instructions were presented on the aforementioned virtual screen display, which was positioned in front of the participants in the immersive virtual environment. All stages were subjected to fixed durations of time and conducted in a continuous sequence automatically (see Fig. 3).


Memorize the following letter pairs:
LX WJ XH DL
Which two of the lower four figures are rotated versions of the upper?
A B C D
Which two of the lower four figures are rotated versions of the upper?
Please answer now
A B C D
Recall the letter pairs
Please wait

(a) 15 seconds (b) 15 seconds (c) 10 seconds (d) 10 seconds (e) 5 seconds

Fig. 3. Sequential demonstration of a sample trial of the cognitive test (a) exposure to letter pairs, (b) exposure to 3D figures, (c) completion of mental rotation task, (d) recollection of letter sequence, and (e) between-task short break

To compare participants' self-perceived subjective performance in the normal and increased height conditions, upon completion of all cognitive test in each block, we asked participants to estimate their accuracy rates in both the letter recall task and mental rotation task.

Letter Recall Task. Letter recall task established on Goldin-Meadow et al.'s [7] cognitive test was leveraged to measure participants' working memory capacity. In each trial, we presented four distinctive letter pairs to facilitate the memorization and recollection process. In accordance with prior work, it was found that relative to letter sequence consisting of both vowels and consonants, those formed solely by consonants can yield a more optimum level of difficulty for testing [7]. Hence, each trial of the letter recall tasks adopted in this study was developed based on eighty randomized consonant letters.

In the initial phase of each trial (Fig. 3a), participants were asked to memorize the visually presented letter sequence within a 15 s exposure period. Subsequent to stimuli exposure, participants experienced a 25 s retention interval. In this phase, the mental rotation task was conducted without the presence of any visual prompt with respect to the letter pairs (Fig. 3b, c). This was immediately followed by a 10 s recollection period in order to facilitate serial recall and verbal report of the memorized letter pairs (Fig. 3d). Participants' responses were recorded and scored by the experimenter, and reported as percentage correct in later analysis. Each correct letter pair was awarded with one mark, thus a maximum of forty marks could be attained in each block. No marks were given for letter pairs with noted transposition errors and incorrect combination of letters.

Mental Rotation Task. To examine participants' spatial ability, spatial tasks based on the Vandenberg and Kuse Mental Rotation Test were formulated using 3D figures in the Library of Shepard and Metzler-type Mental Rotation Stimuli [17, 18]. In this task, participants were first shown five figures in a 15 s exposure period (Fig. 3b). All stimuli were in rotation around the horizontal axis and presented in an identical white frame against a white background. A reference stimulus was positioned on top of four potential matching blocks, which were positioned in various orientations and labelled as option "A", "B", "C" and "D" respectively. Participants were asked to orient mental representations of the stimuli for dynamic comparisons, then verbally report two figures that shared the same configuration with the reference stimuli in the subsequent 10 s response period (Fig. 3c). Participants' responses, represented by two different alphabets, were recorded and scored by the experimenter, and reported as percentage correct in later analysis. One score was awarded for each correct response, which can accumulate to a maximum of twenty marks per condition. In addition to being a stand-alone assessment of visuospatial ability, the mental rotation task also serves the purpose of inducing cognitive load in the letter recall process.

State Self-Esteem Scale (SSES). We used the State Self-Esteem Scale, a well-validated and psychometrically sound measurement, to measure momentary fluctuations of individual self-esteem subsequent to height manipulation [10]. It comprises twenty self-report items, each rated on a 5-point Likert scale (1 = Not at all, 2 = A little bit, 3 = Somewhat, 4 = Very much, and 5 = Extremely). Three subscales were used to conduct multidimensional assessment in the specific facets of appearance, performance, and social self-esteem. In the current study, participants were asked to rate the SSES for how they perceive themselves in relation to their experience in the virtual environment.

2.6 Data Analysis

Following a within-group experimental design, paired-sample t-test analyses were carried out using jamovi (version 1.2.27). Outcome variables were compared between the normal and increased height conditions. The order of conditions was counter-balanced in order to minimize potential confounding influence of the sequence, for instance, cognitive fatigue and learning effects. We checked for normality using Shapiro-Wilk test.

3 Results

3.1 Letter Recall Task

Figure 4a shows the letter recall task performance in the normal and increased height conditions. The within-group experimental design was tested with a paired-sample t-test to examine the effect of height manipulation on working memory capacity with regard to serial recall performance. Participants under the influence of heightened virtual perception remembered 23.78% more letter pairs when compared to the normal height condition. The difference in mean proportion of correctly recalled letter pairs across conditions was statistically significant, $t(9) = 2.56$, $p = 0.031$, Cohen's d = 0.809. In terms of subjective performance, the self-report estimation of correctly remembered letter pairs in the increased height condition was 8% higher than that of the control condition. However, the results were not confirmed statistically, $t(9) = 1.58$, $p = 0.149$, Cohen's d = 0.499.

3.2 Mental Rotation Task

Figure 4b shows the mental rotation task performance in the normal and increased height conditions. We conducted a paired-sample t-test to compare the mean proportion of correctly solved spatial tasks under the effect of differing height perceptions. Our results indicated that participants identified 17.48% more correct mental rotation figures when they became taller in the virtual environment. The difference in performance across conditions was confirmed statistically, $t(9) = 2.68$, $p = 0.025$, Cohen's d = 0.846. However, participants' subjective estimations of correctly solved mental rotation problems did not differ significantly across the two conditions, $t(9) = -0.345$, $p = 0.738$, Cohen's d = −0.109.

3.3 State Self-Esteem

Figure 5 shows participants' overall level of SSES, as well as the appearance, performance, and social facets in the normal and increased height condition. We conducted a paired-sample t-test to examine the potential fluctuations in self-esteem induced by virtual height alteration. Results revealed that the SSES total score was higher in the increased height condition when compared to the normal height condition. However, the mean scores were not significantly different across conditions, $t(9) = 1.28$, $p = 0.233$, Cohen's d = 0.405.

To further investigate the effects of height manipulation on different domains of self-esteem, paired-sample t-tests were conducted to compare the SSES scores in each of the three subscales. We found that participants' appearance self-esteem was significantly higher under height manipulation when compared to the control scenario, $t(9) = 2.50$, $p = 0.034$, Cohen's d = 0.789. However, we found no evidence for significant changes across conditions in performance self-esteem, $t(9) = 0.77$, $p = 0.462$, Cohen's d = 0.243, and social self-esteem, $t(9) = 1.15$, $p = 0.279$, Cohen's d = 0.364.

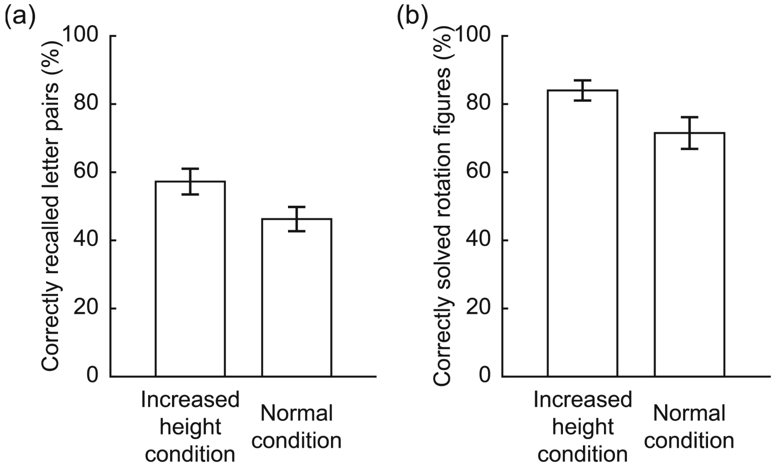


Fig. 4. Performance in cognitive tests (a) letter recall task and (b) mental rotation task in the normal and increased height condition expressed in terms of percentage correct. Error bars indicate SEM across participants.

4 Discussion

Over time, extensive literature has been devoted to investigating the underlying causes and deep-rooted effects of height-associated benefits in societal settings [12]. While the impact of stature on self-esteem and cognitive ability has garnered scholarly interest, few studies have yet attempted to address both constructs in a collective manner. In this pilot study, we manipulated participants' virtual height in an immersive VR system to examine their self-esteem, memory skills, and spatial ability under different situations.

In accordance with our hypotheses, the present findings revealed that increasing a person's height in virtual environment can significantly lead to better objective performance in cognitive tests. The results were illustrated by the increase in proportion of correctly remembered letter pairs and correctly solved mental rotation tasks. While the manipulation of height was moderately subtle and remained unnoticed throughout the experimental procedures, participants demonstrated stronger memory and spatial skills when they became taller in the virtual environment. In contrary to our hypotheses, the self-perceived performance of cognitive tests did not differ significantly across conditions. In fact, during our debriefing process, we casually explain and discuss the experiment with the participants. It is interesting to note that height-induced improvement in cognitive abilities was not consciously acknowledged by the majority of the participants. We speculate that this observation might be due to the varied changes in different domains of self-esteem. As height constitute a major component of appearance-related appraisals, rationales underlying the effect of height manipulation on increased appearance self-esteem is pronounced and straightforward. In spite of this, self-esteem remained largely unchanged in other subtypes. In particular, the insignificant findings with regard to performance self-esteem may shed light on the relatively consistent self-report ratings of cognitive tasks, indicating a close-knitted relationship between self-esteem and self-perceived performance. Nonetheless, the results confirmed that the effects of height

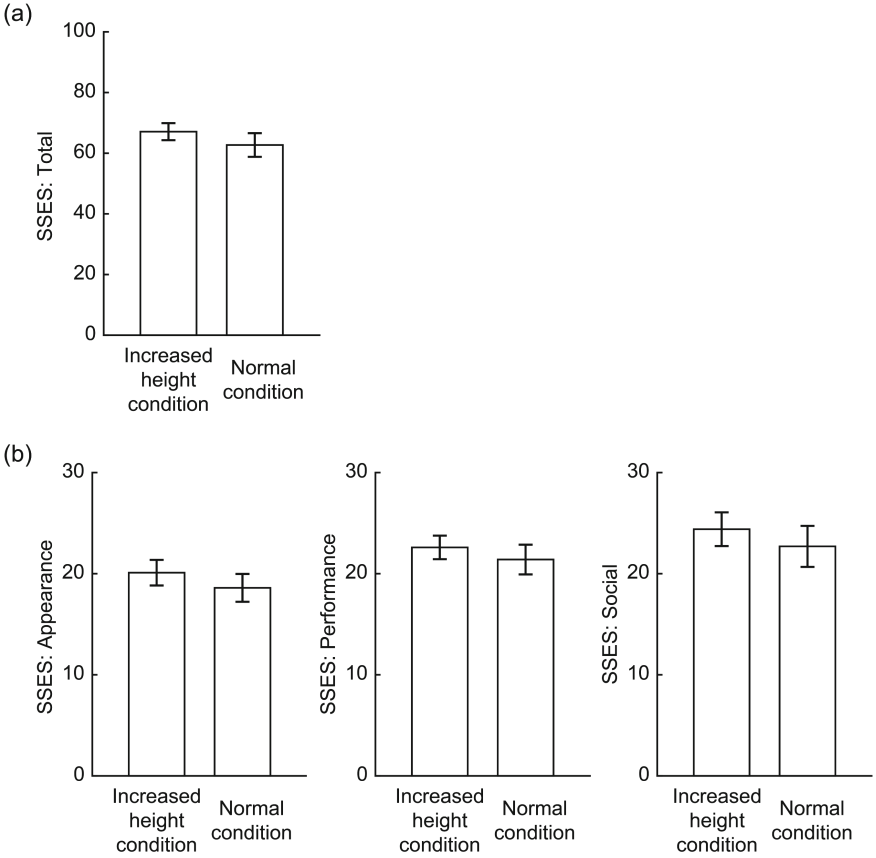


Fig. 5. (a) Total score and (b) specific facets of State Self-Esteem Score (SSES) in the normal and increased height condition. Error bars indicate SEM across participants.

alterations were sensitive to stature-related self-appraisals, even in the absence of external evaluations.

To the best of our knowledge, this is the first study to examine self-esteem, visuospatial ability, and memory skills under direct height manipulation in a virtual environment. The advantages of leveraging VR systems to implement our experiment comes in threefold. Firstly, immersive VR disengages participants from the external world to provide realistic experience in cognitive, affective, and behavioral domains, ensuring that the test conducted can achieve high ecological validity [22]. Secondly, studies have indicated that the traditionally prominent gender differences in mental rotation skills are predominantly diminished when tasks are conducted in virtual settings [15]. Therefore, the confounding sex effect in relation to visuospatial ability should not influence the findings in this study. (Note, however, that we did not formally test this in the pilot study). Thirdly, VR provides a convenient and accessible medium to induce effective height alterations. Yee and Bailenson [31] found that individuals assigned with taller avatar as digital representation of themselves were more confident and outperformed their shorter counterparts in a virtual negotiation task. Their findings confirmed the proteus effect and suggested that representational manipulations in VR are capable of yielding dynamic and rapid behavioral changes. However, it is to note that rather than rendering digitized self-avatars, the current study performed direct manipulation on individuals' height in VR. This is to eliminate the potential confounding influence of social esteem in relation to perceived representations, which may mediate the effect of height adjustment on self-esteem.

One limitation of the present findings inheres in the relatively small sample size of this pilot study. It is advised that future studies should aim to replicate results with a larger sample to facilitate mediation analysis. However, we implemented a within-group experimental design and found sufficient effect sizes for our results, which may mitigate concerns in relation to power and sample size. On this basis, future research could explore the potential mediating effect of other height-related variables, such as social dominance and leadership emergence, as well as cognitive abilities in other domains. Studies could also consider the potential of utilizing VR-based height alteration in improving people's cognitive abilities. In this study, we mainly focused on working memory and spatial ability as both are important indicators of academic achievements and social functioning [1, 16]. It is of interest to further examine how the incorporation of height manipulation into evidence-based VR interventions could benefit a variety of populations. For instance, it may hold the potential to maximize treatment outcomes for patients suffering from cognitive deficits. Promising prospects may also be found in enhancing self-perceived psychosocial functioning of children with idiopathic short stature [30]. Furthermore, the study can be further developed by employing advanced technology to refine the height alteration process. Full-body tracking equipment and eye-trackers are promising add-ons to elicit real-time realistic, synchronized, and individualized immersion experiences [4]. Alternative ways of manipulating a person's height in VR can also be investigated and compared against the current approach. Considering the impact of social esteem, researchers could display 3D avatars in virtual environments to stimulate users' visualization of their perceived height. The cave automatic virtual environment (CAVE) would also be an interesting replacement for VR systems facilitated by the HMD.

5 Conclusion

In this pilot study, we utilized an immersive VR system to compare individuals' state self-esteem, working memory, and visuospatial competence in differing virtual body heights. We adapted the height-career success theoretical model to predict height-induced enhancement in subjective self-evaluation and cognitive abilities. Our results indicated that increasing individuals' height in VR can elicit higher levels of appearance self-esteem, and improve objective performance in letter recall and mental rotation tests. However, we found no evidence for significant changes in self-perceived performance and overall self-esteem. Future research is needed to explore the potential of incorporating VR-based direct height manipulation in clinical trials and practical settings.

Acknowledgement. Adrian K. T. Ng holds a Hong Kong PhD Fellowship, funded by the Hong Kong Research Grants Council. This work was also partially supported by Hong Kong Research

Grants Council's General Research Fund awarded to Henry Y. K. Lau (Grant number: 17203620). We would like to express our gratitude to Prof. Michael Peters and Dr. Christian Battista, lead researchers of the Library of Shepard and Metzler-type Mental Rotation Stimuli, for their generous endorsement to utilize the mental rotation figures in our study. We would also like to thank Esther Y. Y. Chiu and Cheryl H. Y. Leung for their technical assistance.

References

1. Aronen, E.T., Vuontela, V., Steenari, M.-R., Salmi, J., Carlson, S.: Working memory, psychiatric symptoms, and academic performance at school. Neurobiol. Learn. Mem. **83**(1), 33–42 (2005). https://doi.org/10.1016/j.nlm.2004.06.010
2. Blaker, N.M., Rompa, I., Dessing, I.H., Vriend, A.F., Herschberg, C., Van Vugt, M.: The height leadership advantage in men and women: testing evolutionary psychology predictions about the perceptions of tall leaders. Group Process Intergroup Relat. **16**(1), 17–27 (2013). https://doi.org/10.1177/1368430212437211
3. Booth, N.D.: The relationship between height and self-esteem and the mediating effect of self-consciousness. J. Soc. Psychol. **130**(5), 609–617 (1990). https://doi.org/10.1080/00224545.1990.9922952
4. Caserman, P., Garcia-Agundez, A., Konrad, R., Göbel, S., Steinmetz, R.: Real-time body tracking in virtual reality using a Vive tracker. Virtual Reality **23**(2), 155–168 (2018). https://doi.org/10.1007/s10055-018-0374-z
5. Cast, A.D., Burke, P.J.: A theory of self-esteem. Soc. Forces **80**(3), 1041–1068 (2002). https://doi.org/10.1353/sof.2002.0003
6. Freeman, D., Evans, N., Lister, R., Antley, A., Dunn, G., Slater, M.: Height, social comparison, and paranoia: an immersive virtual reality experimental study. Psychiatry Res. **218**(3), 348–352 (2014). https://doi.org/10.1016/j.psychres.2013.12.014
7. Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., Wagner, S.: Explaining math: gesturing lightens the load. Psychol. Sci. **12**(6), 516–522 (2001). https://doi.org/10.1111/1467-9280.00395
8. Gorry, D.: The influence of height on academic outcomes. Econ. Educ. Rev. **56**, 1–8 (2017). https://doi.org/10.1016/j.econedurev.2016.11.002
9. Grogan, S.: Body Image: Understanding Body Dissatisfaction in Men, Women and Children. Routledge, New York (2017)
10. Heatherton, T.F., Polivy, J.: Development and validation of a scale for measuring state self-esteem. J. Pers. Soc. Psychol. **60**(6), 895–910 (1991). https://doi.org/10.1037/0022-3514.60.6.895
11. Hermanussen, M., Scheffler, C.: Stature signals status: the association of stature, status and perceived dominance-a thought experiment. Anthropol. Anz. **73**(4), 265–274 (2016). https://doi.org/10.1127/anthranz/2016/0698
12. Judge, T.A., Cable, D.M.: The effect of physical height on workplace success and income: preliminary test of a theoretical model. J. Appl. Psychol. **89**(3), 428–441 (2004). https://doi.org/10.1037/0021-9010.89.3.428
13. Just, W., Morris, M.R.: The Napoleon complex: why smaller males pick fights. Evol. Ecol. Res. **17**(5), 509–522 (2003). https://doi.org/10.1023/b:evec.0000005629.54152.83
14. Ng, A.K.T., Chan, L.K.Y., Lau, H.Y.K.: A low-cost lighthouse-based virtual reality head tracking system. In: 2017 International Conference on 3D Immersion (IC3D), pp. 1–5. IEEE Press, New York (2017). https://doi.org/10.1109/IC3D.2017.8251910

15. Parsons, T.D., et al.: Sex differences in mental rotation and spatial rotation in a virtual environment. Neuropsychologia **42**(4), 555–562 (2004). https://doi.org/10.1016/j.neuropsychologia.2003.08.014
16. Pellegrino, J.W., Alderton, D.L., Shute, V.J.: Understanding spatial ability. Educ. Psychol. **19**(4), 239–253 (1984). https://doi.org/10.1080/00461528409529300
17. Peters, M., Battista, C.: Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. Brain Cogn. **66**(3), 260–264 (2008). https://doi.org/10.1016/j.bandc.2007.09.003
18. Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., Richardson, C.: A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. Brain Cogn. **28**(1), 39–58 (1995). https://doi.org/10.1006/brcg.1995.1032
19. Prieto, A.G., Robbins, M.C.: Perceptions of height and self-esteem. Percept. Mot. Skills **40**(2), 395–398 (1975). https://doi.org/10.2466/pms.1975.40.2.395
20. Riva, G., Gaggioli, A., Dakanalis, A.: From body dissatisfaction to obesity: how virtual reality may improve obesity prevention and treatment in adolescents. Stud. Health Technol. Inform. **184**, 356–362 (2013). https://doi.org/10.3233/978-1-61499-209-7-356
21. Salice, A.: Self-esteem, social esteem, and pride. Emot. Rev. **12**(3), 193–205 (2020). https://doi.org/10.1177/1754073920930788
22. Sanchez-Vives, M.V., Slater, M.: From presence to consciousness through virtual reality. Nat. Rev. Neurosci. **6**(4), 332–339 (2005). https://doi.org/10.1038/nrn1651
23. Scarpina, F., Serino, S., Keizer, A., Chirico, A., Scacchi, M., Castelnuovo, G., et al.: The effect of a virtual-reality full-body illusion on body representation in obesity. J. Clin. Med. **8**(9), 1330 (2019). https://doi.org/10.3390/jcm8091330
24. Schuemie, M.J., Van Der Straaten, P., Krijn, M., Van Der Mast, C.A.: Research on presence in virtual reality: a survey. Cyberpsychol. Behav. Soc. **4**(2), 183–201 (2001). https://doi.org/10.1089/109493101300117884
25. Siegel, P.T., Clopper, R., Stabler, B.: Psychological impact of significantly short stature. Acta Paediatr. **377**, 14–18 (1991). https://doi.org/10.1111/apa.1991.80.s377.14
26. Shahyad, S., Pakdaman, S., Shokri, O.: Prediction of body image dissatisfaction from self-esteem, thin-ideal internalization and appearance-related social comparison. Int. J. Travel Med. Glob. Health **3**(2), 65–69 (2015). https://doi.org/10.20286/ijtmgh-030299
27. Spears, D.: Height and cognitive achievement among Indian children. Econ. Hum. Biol. **10**(2), 210–219 (2012). https://doi.org/10.1016/j.ehb.2011.08.005
28. Steed, A., Pan, Y., Zisch, F., Steptoe, W.: The impact of a self-avatar on cognitive load in immersive virtual reality. In: 2016 IEEE Virtual Reality (VR), pp. 67–76. IEEE Press, New York (2016). https://doi.org/10.1109/VR.2016.7504689
29. Villani, D., Gatti, E., Triberti, S., Confalonieri, E., Riva, G.: Exploration of virtual body-representation in adolescence: the role of age and sex in avatar customization. Springerplus **5**(1), 1–13 (2016). https://doi.org/10.1186/s40064-016-2520-y
30. Visser-van Balen, H., Sinnema, G., Geenen, R.: Growing up with idiopathic short stature: psychosocial development and hormone treatment: a critical review. Arch. Dis. Child. **91**(5), 433–439 (2006). https://doi.org/10.1136/adc.2005.086942
31. Yee, N., Bailenson, J.: The Proteus effect: the effect of transformed self-representation on behavior. Hum. Commun. Res. **33**(3), 271–290 (2007). https://doi.org/10.1111/j.1468-2958.2007.00299.x
32. Zyda, M.: From visual simulation to virtual reality to games. Computer **38**(9), 25–32 (2005). https://doi.org/10.1109/MC.2005.297

The Influence of Gender on Human's Cognitive Ability and the Correlation Research of Different Cognitive Dimensions

Zhongqi Liu[1,2], Le Dong[1,2], and Qianxiang Zhou[1,2](✉)

[1] Key Laboratory for Biomechanics and Mechanobiology of the Ministry of Education, School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China
zqxg@buaa.edu.cn

[2] Beijing Advanced Innovation Centre for Biomedical Engineering,
Beihang University, Beijing 102402, China

Abstract. Human's ability and characteristic are important content in ergonomics research. Only when machine and environment adapt to person's ability and characteristic, can human work safely, healthily, comfortably and efficiently. Four cognitive characteristic indexes were studied in this experiment which they are: visual response speed, attention span, spatial location memory span and short-time memory span. The aim was to study the influence of gender on four indicators and the correlation between them. Thirty-one postgraduates of Beihang University took part in the experiment. The subjects included 17 female and 14 male who aged from 22 to 25 years old, with an average age of 23. They were healthy and their vision or corrected vision were 1.5 or above. The results showed that there were no significant differences in the reaction time, attention span, short term memory span and spatial location memory span between males and females ($P > 0.05$). According to the comparison of the comprehensive value of cognitive ability, the comprehensive value of comprehensive ability of girls and boys was about 0.49 and 0.48 respectively, and the cognitive ability of girls was slightly higher than boys. The results of independent T-test showed that the $P > 0.05$, so there was no significant difference in cognitive ability between boys and girls. In terms of the correlation of four indicators in different cognitive dimensions, the Pearson correlation coefficient of the short-time memory span value and the attention span value was 0.445, and the significance level was 0.05. There was a certain correlation between the two indicators, while there was no obvious correlation among other indicators. This research can provide reference for information product design and some job recruitment.

Keywords: Ergonomics · Cognition · Gender · Correlation · Attention · Memory

1 Introduction

Human's ability and characteristics are an important part of ergonomics research, machines and environment must adapt to human abilities and characteristics, can human

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 185–196, 2021.
https://doi.org/10.1007/978-3-030-77932-0_16

beings work safely, healthy, comfortable and efficiently. Man is a complex system with all aspects of ability and characteristics that must be taken into account when designing a machine, such as the size of the human body, the size of the human force, the visual, auditory, human attention, memory, learning, reaction speed and other psychophysical and cognitive characteristics. This study used experiments to explore the effects of gender differences on people's cognitive abilities and the correlation between different cognitive dimensions. People have many cognitive dimensions, and this experiment only focused on four of them: human response speed or the reaction time (RT), attention breadth (AB), spatial location memory breadth (SLMB), and short-term memory breadth (STMB).

The speed of human response refers to the ability of a person to respond quickly to various external stimuli, reflecting the coordination and rapid response ability of the nerve and muscle system. Reaction speed is an important part of the evaluation index of physical fitness, which reflects the efficiency of muscle work and the function of human motor system. The faster the reaction speed, the faster the body's response to stimulation, with the increase of exercise fatigue, the reaction speed shows a downward trend, therefore, the reaction speed is also an indicator of fatigue level. In the field of psychology, the reaction time is considered to be a reliable psychological indicator, which can measure the level of excitability and inhibition of the brain and analyze various psychological activities such as human perception, feeling, attention, thinking, creation and learning [1]. It is concluded that there are two basic cause variables in the reaction time experiment, namely speed and accuracy. One of the prominent problems in the reaction time experiment is to tradeoff the relationship between the two. When subjects join a sport, they will consider the law of speed accuracy trade-offs: if the action is the faster, it will be less accurate; on the contrary, the action is the accurater, the speed will be slower. In recent years, the relevant research on reaction time has become more and more in-depth. Many scholars have studied the reaction time from different angles, such as psychology, medicine and physical education, and have made some important findings. Zhao Runshuan et al. found that the body fat rate was related to the reaction rate, and concluded that the reaction time increased by 0.0014s and 0.0015s when the body fat rate of men and women exceeded 1%. Excess body fat rate can lead to a decrease in physical fitness [2]. Zhang Liwei et al. made a analysis on the research of reaction time, and he put forward that the reaction of the short-distance speed project was better than that of the long-distance speed project, and the reaction time of the adversary open sports project was better than that of the non-confrontational and closed sports [3].

Attention breadth refers to the pointing and concentration of psychological activities or consciousness to a certain object, that is, people point to and concentrate their own perception, memory, thinking and other activities on the selected object. Attention has the function of selection, maintenance, integration and adjustment, the input of information to make choices, thus ensuring a clearer understanding of things and more accurate response, therefore, it is an important psychological quality for people to acquire knowledge, master skills, complete various intelligence operations and practical operations, and it is also the basis of information processing and other psychological activities [4]. There has been studies on the effects of emotional induction on attention breadth in humans themselves [5]. The earliest studies have shown that things with emotions have the advantage of gaining access to and focusing on attention resources compared to emotionless things. Rowe et al. looked at the effects of positive emotions on attention breadth. It was found that positive emotions did increase the breadth of visual attention compared to negative and neutral emotions, resulting in a more significant flanker effect [6].

Spatial location memory breadth refers to an individual's ability to identify, encode, store, characterize, decompose/combine and abstract/summarize objects or spatial graphics in the mind, including spatial observation, spatial memory, spatial imagination and spatial thinking. Researchers have carried out some analysis from the perspective of individual sex and age. There were great differences in the size, importance and development process of gender differences in visual spatial ability and interpretation theory, some studies have found that girls lag behind in tasks with spatial factors, and the differences between boys and girls increased with age; another study conducted a visual test on college students, and found that there was no gender difference, while many previous studies have found that visual ability was the biggest gender difference in cognitive field. There are relatively few studies on the age difference of visual spatial ability, but the results were more consistent: young people have a greater advantage in spatial ability. In addition to gender and age factors, the researchers also studied the visual spatial abilities of different types of subjects. For example, Chinese scholar Zou Jinli discussed the cognitive characteristics of Chinese children's visual space at different literacy levels. Zhao Wei conducted experimental research on the speech consciousness and visual spatial cognition of Chinese students with difficulty reading. The research on the visual spatial ability of different groups of people in China is still limited to one aspect of spatial ability, which lacks the whole discussion of spatial ability, and the selection range of the subject groups is also narrow [7].

The task of short-term memory breadth is an important index to evaluate short-term memory ability. It has been found that the memory breadth of different materials varies greatly, for example, the breadth of numbers is generally greater than the breadth of words. Another short-term memory task, the Sternberg recognition task, had similar characteristics. Sternberg recognition task presented a string of numbers to the subjects, called reaction set, for the subjects to remember, followed by a test stimulus that asked the subjects to determine whether they were part of the reaction set. With the increase of reaction set items, the time for the subjects to do the correct reaction increased linearly. According to Sternberg, this reflected the fact that searching for items in short-term memory was a series of end-of-life scans. The search speed (that is, the scan time for each item) could be estimated by the slope of the linear regression equation for reaction time and reaction set size [8]. Many studies have shown that materials with a wide memory have shorter scanning times. Cavanagh collected literature on the memory breadth and search speed of seven different types of stimulus materials and found an inverse relationship between the two indicators of short-term memory [9]. There are many factors that affect the breadth of short-term memory. The size, complexity, and familiarity of the blocks all affect the capacity of short-term memories. Studies such as Yu Bolin and Zhang Wutian showed that the short-term memory capacity of two-tone synthesized words and four-word idioms was not as large as that of single-tone words, while four-word idioms were not as many as two-tone words; with the increase of block complexity, short-term memory capacity tends to decrease gradually; high-frequency words had more short-term memory capacity than low-frequency words. Zhang Wutian et al. found that the stroke complexity of Chinese characters had a significant effect on short-term memory hold, and the recognition time of words with different stroke complexity had a significant negative correlation with their memory hold [10].

2 Method

2.1 Subjects

Thirty-one postgraduate students from Beihang University took part in the experiment, aged between 22 and 25, with an average age of 23. They were in good health, had no specific disease, and had naked or corrected visual acuity of 1.0 or above.

2.2 Experiment Content

Reaction Time Measurement. The time between the presentation of the stimulus and the response is called the response time. Response time measurements include visual response time measurements and auditory response time measurements. This experiment measured visual response time. The visual stimulus was a green circle (Fig. 1). All the subjects reacted by pressing the green key. The test was performed for 30 times, and the stimulation was presented at an interval of 2 s after each preparation. If there was preemptive action in the preparation stage, the result was invalid, and the computer would reject the subject and warned him. In addition, 2S of blank detection stimulus was randomly added in a group every 5 times. If any subject was forestall at pressing key, he would be warned, and this group of experiments would be re-conducted. Finally, the mean of the effective result was taken as the reaction time.

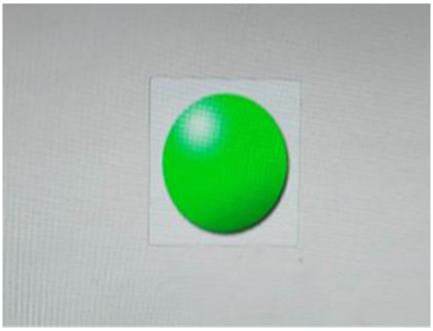


Fig. 1. Experimental stimulation for reactions time

Attention Span Measurement. Attention span usually refers to the scope of visual attention, that is, the number of objects perceived at the same time. In this experiment, the attention span of the subjects with different number of red dots arranged randomly was calculated by psychophysical method.

During the experiment, the number of red dots randomly presented on the computer screen (As shown in Fig. 2) ranged from 5 to 12, 10 of each type, 80 in total. Each display time was 0.25 s. The subject was asked to press the corresponding number key to type the answer (or click the input box with the mouse and then enter the answer through the selection menu). The percentage of correct responses to different dots was counted. Starting from five dots, the number of dots with the first 50% correct responses was calculated by linear interpolation method as the attention span value.

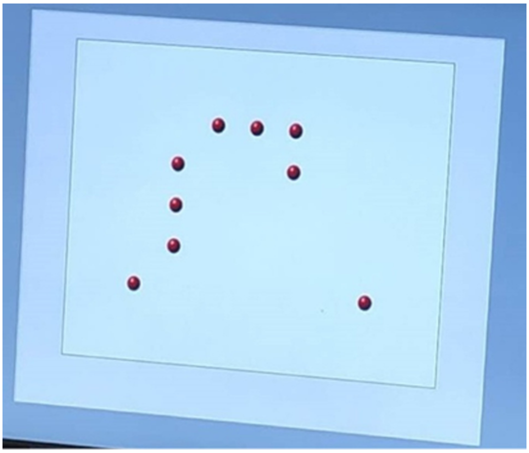


Fig. 2. Experimental stimulation of attention span

Short Term Memory Span Measurement. Memory span refers to the length of a series of stimuli that can be reproduced immediately after a sequence of stimuli is presented in a fixed order. The intervals between the stimuli presented must be equal, and the results reproduced must correspond to the order in which they were presented. Similar to the concept of sensory threshold, the so-called "just able to reappear immediately" means that 50% of the times can be reappeared immediately.

Digital memory span is a simple and easy method to measure short-term memory ability. The stimulus presented in this experiment was a series of numbers (Fig. 3).

The subjects read the instructions carefully, figuring out how to remember and how to enter the answers. When entering the answer, there should be no Spaces between the numbers. If there was any error, the subject could press Back Space to delete and re-enter. After typing, pressed enter to confirm. The interval between the numbers was 750 ms, each number presented 250 ms, starting with 3 digits, then 4, 5, 6… Until the same number series was wrong three times or up to 12 digits.

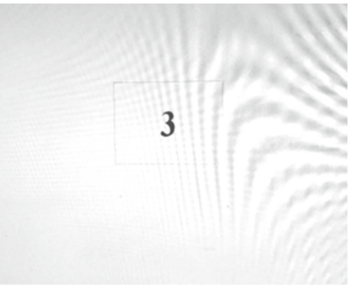


Fig. 3. Experimental stimulation of short-term memory span

The method of calculating memory span was as follows:

If each series is presented three times in a row, the longest series that can pass the three times is taken as the cardinal number, and then the length of other stimulation series that can pass is added to the cardinal number by 1/3 or 2/3, and the sum of the two is counted as memory span. This method was used to measure and calculate the digit memory span in this experiment. For example, if the longest series that could pass three times was 7 digits, the base number was 7. If the 8-digit series passed twice, the 9-digit series passed once, and the 10 digit series failed once, the memory span was $7 + 2/3 + 1/3 = 8$.

Spatial Location Memory Span Measurement. Spatial location memory span refers to the length of the spatial location series that the subject can reproduce immediately after presenting a series of positions in a fixed order. Participants must reproduce the order in which they were originally presented.

During the experiment, a 5 × 3 green table (Fig. 4) was presented on the computer screen, and the bright pink dots were randomly presented in some of the 15 grids (starting with 3 consecutive grids at a time). The subjects were asked to try their best to remember the location and order of the dots. After the dots appeared, the subjects were asked to click the table in the order in which the dots appeared just now, and then input by clicking the "OK" button with the mouse. After a span has been done three times, if it is not all wrong, the span is multiplied by 1 and continues until a span has been wrong three times in a row or has completed 12 tasks. Each time the participant entered an answer, if there were any errors, he or she could click on the last table to modify.

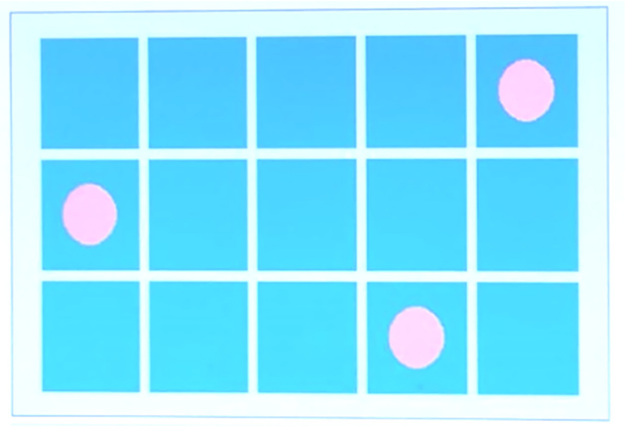


Fig. 4. Experimental stimulation of spatial location memory span (Color figure online)

2.3 Experimental Process

In this experiment, the reaction time, attention span, short-term memory span and spatial location memory span were measured in the order. The subjects first practiced the experiment by themselves until they were proficient, and then they carried out the formal measurement experiment. In the process of the experiment, they needed to rest and relax for five minutes to complete an index, and then they carried out the experimental measurement of the next index. The experimental scene was shown in Fig. 5.

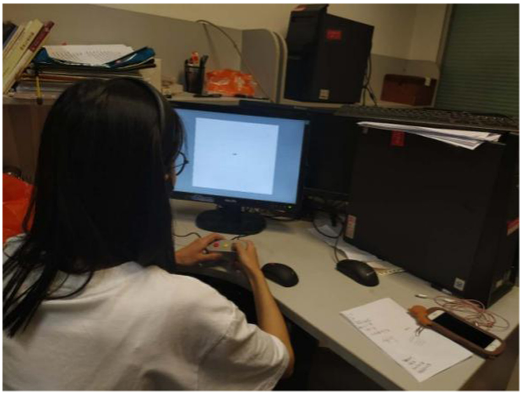


Fig. 5. Experiment scene

3 Results

3.1 Verification and Elimination of Abnormal Data

Box-plot, also known as box-whisker plot, box-plot or boxplot, is a statistical graph showing the dispersion of a group of data. It gets its name from the shape of a box, as shown in Fig. 6. The standard of box-plot to determine outliers is based on quartiles and quartile distance. Quartiles have a certain resistance, and up to 25% of the data can become arbitrary and far away without greatly disturbing the quartiles. Therefore, outliers cannot affect this standard, and the results of box-plot to identify outliers are more objective. There is a box in the middle of the box line diagram, that is, the pink part. There are lines on the left, middle and right of the box. The left is the lower quartile (Q1), the right is the upper quartile (Q3), and the middle is the median. The difference between the upper and lower quartiles is the interquartile range (IQR). Q1–1.5IQR is used to get the lower edge (minimum) and Q3+1.5IQR is used to get the upper edge (maximum). The data outside the upper edge is the maximum outlier, and the data outside the lower edge is the minimum outlier. In short, the data outside the upper and lower edge is the outlier.

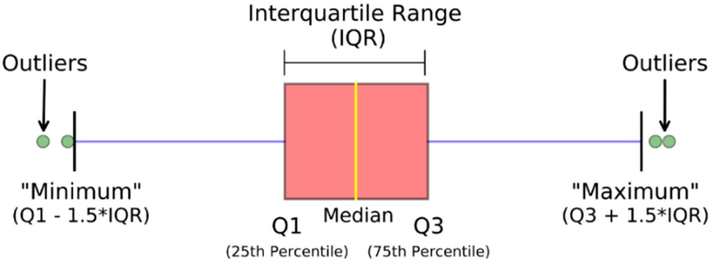


Fig.6. Box line diagram

According to the results of boxplot, in the memory span value of spatial position, data No. 15 was small and was an outlier, which can be eliminated. In the value of short-term memory span, data No. 30 was a small outlier, which can be eliminated. There was no outlier in the span of attention value. In the simple reaction time, there were 23, 30, 24, 29 which were outliers, which could be removed. Since the data corresponding to the experiment number was the experimental result of the same person, based on this, if there was a certain data deviation, it might be caused by the status of the subject or other factors. In order to ensure the reliability of the experimental data, the data obtained by the subject will be excluded here. In conclusion, for the collected data, data No. 15, 23, 24, 29 and 30 were excluded here for further statistical analysis of the remaining data.

3.2 The Influence of Gender on Cognitive Ability

The Influence of Gender on Different Dimensions of Indicators. Independent T test was used to determine whether there were significant gender differences in the four indicators. The results were shown in Table 1:

Table 1. Statistical results of the effect of gender on cognition

Cognitive indicators	Gender	Mean	F	t	Sig.(P)
SLMB	M	6.1	0.49	1.271	0.216
	F	5.8			
STMB	M	7.7	1.029	−0.966	0.344
	F	8.2			
AB	M	8.3	0.306	−1.357	0.187
	F	9.2			
RT /ms	M	643	0.368	0.963	0.345
	F	551			

The results showed that: the average value of spatial location memory span of boys was 6.132, which was slightly higher than that of girls' 5.770. The average value of short-term memory span of boys was 7.666, and that of girls was 8.166, which was slightly higher than that of boys' data. The average value of attention span of boys was 8.308, and that of girls is 9.176. The average value of simple reaction time of boys is 643.40, the average value of girls is 551.438. It could be seen that there was a slight difference in the average value of men and women of the four indicators. The independent T test results showed that there were no significant gender differences($P > 0.05$) response time, attention span, short term memory span and spatial location memory span.

The Influence of Gender on Comprehensive Cognitive Ability. Does gender have any effect on the combined value of the four indicators? The weighted calculation method could be used to carry out mathematical calculation of the four indexes to get the comprehensive value. Since the dimension of each index was different, each index needed to be normalized. The normalization algorithm was shown in Eq. (1).

$$\mathrm{Y} = \frac{\mathrm{X} - \mathrm{MIN}}{\mathrm{MAX} - \mathrm{MIN}} \tag{1}$$

Where: Y represented the normalized value. X represented the experimental results before normalization. MIN represented the minimum value in the experimental results. MAX represented the maximum value of the experimental result.

The four indexes were calculated according to the weight to get a comprehensive value. Weight distribution method: the weight of simple reaction was 1.9; the other three indicators were 2.7 each. The weight sum of the four indexes was 1. The comprehensive value of cognitive ability was calculated according to Eq. (2).

$$\mathrm{Y} = 1.9{*}X_1 + 2.7{*}(X_2 + X_3 + X_4) \tag{2}$$

In formula (2), X1, X2, X3 and X4 represented the normalized simple reaction time, attention span, short-term memory span and spatial location memory span respectively; Y represented the comprehensive value of cognitive ability.

Table 2. The influence of gender on comprehensive cognitive ability

Gender	Mean	F	t	Sig.(P)
M	0.481	0.358	−0.239	0.813
F	0.493			

Independent T-test was conducted on the comprehensive value of cognitive ability, and the results were shown in Table 2:

The average comprehensive value of boys was about 0.481, and that of girls was about 0.493. The comprehensive value of boys was slightly lower than that of girls. The results of independent t-test showed that $P > 0.05$, so there was no significant gender difference in the comprehensive value of cognitive ability.

3.3 Correlation of Different Cognitive Dimensions

The calculation formula of Pearson correlation coefficient ρ_{xy} of two n-dimensional vectors X and Y is shown in Eq. (3):

$$\boldsymbol{\rho_{xy}} = \frac{\sum_{i=1}^{n}\left[(\boldsymbol{X_i} - \overline{\boldsymbol{X}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})\right]}{\left(\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i} - \overline{\boldsymbol{X}})^2}\right)\left(\sqrt{\sum_{i=1}^{n}(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^2}\right)} \tag{3}$$

Where: X_i and Y_i represent the ith element value of n-dimensional vector X and Y; $\overline{X}$ and $\overline{Y}$ represent the average value of the n elements of the n-dimensional vectors X and Y.

The correlation coefficient between 4 indicators were shown in Table 3:

Table 3. The correlation results among 4 cognitive indicators

Cognitive indicators	SLMB	STMB	AB	RT /ms
SLMB	\	−0.133	−0.247	−0.039
STMB	\	\	.445*	−0.09
AB	\	\	\	0.078
RT /ms	\	\	\	\

*At the 0.05 level (double tails), the correlation was significant.

The results showed that the Pearson correlation coefficient of attention span and short-term memory span was 0.445, and there was a certain correlation at the significance level of 0.05. There was no significant correlation between other indexes.

4 Discussion

There are significant physical differences between women and men, but are there also significant psychological differences between them? Especially in terms of cognitive ability. Some occupations require higher comprehensive cognitive ability, such as fighter pilots. Understanding the cognitive differences between different genders will be of practical significance for the selection and training of some professional personnel. The researchers thought there were no significant differences between men and women in psychological traits overall, but there were differences in some areas, such as women's ability to imagine, think visually and speak, and their emotional nature. There are many studies that show that women have better willpower, endurance, adaptability and sense of responsibility [10, 11].

On gender differences, many studies have inconsistent results. Zhou Zhen et al. studied the gender differences in the development of middle school students' spatial figure folding, expansion, rotation and figure reasoning ability. The results showed that there was a significant difference in male dominance under certain conditions. However, in which age group male students were significantly better than female students in which spatial indicators, further research is needed [12]. Liu Kehui, Yan Xiaojun et al. [13] obtained memory span data of 160 college students in the study of memory span experiment based on regression analysis. The feasibility of regression analysis of the data was verified, which showed that it was statistically significant. The linear equation between short-term presentation time and memory span of numbers and letters was established. Combined with mapping analysis, it was found that there was no significant gender difference in memory span. In the study of college students' digital short-term memory span and processing strategy by Liu Wanlun [14], the research results showed that there was no gender difference in college students' digital short term memory span, which was consistent with this study. Wang Jiatong et al. Explored the differences between male and female flight cadets in basic flight cognitive ability by using four indexes: number retrieval, character recognition, hidden figure and word matching. The results showed that female flight cadets were better than male flight cadets in perceptual discrimination, short-term memory and initiative and agility of thinking exploration. There was no significant difference between male and female students in other cognitive abilities, and there was no significant difference in the comprehensive scores of the four cognitive indexes. There are some similarities and differences with this study [15].

However, it must be realized that in addition to the gender difference, even if the individual difference of the same gender is very large, so in the selection and training of career should not be targeted at gender, but should be more targeted at the measurement of individual cognitive ability.

5 Conclusion

After excluding the outliers of the data of each indicator, independent T test was conducted for each indicator, and the results showed that there was no significant gender difference in the simple response time, attention span, short term memory span, and spatial location memory span. The mean spatial location memory span of male students is slightly higher than that of female students, while the mean short-term memory span and attention span of female students are slightly higher than that of male students. The mean value of boys is slightly higher than girls in simple reaction. Among the four cognitive dimensions, only short-term memory span and attention span are correlated, and they are closely correlated, while other cognitive dimensions are not correlated.

References

1. Xu, J.J.: The evaluation standard research of college students' response speed and movement speed. Nanjing Normal University (2016)
2. Ping, Z., Zhao, R.S., Bai, X.Q., Ou, Y.W., Feng, L.: A Comparison study on the effects of excessive body fat rate on body reaction speed among adults. Zhejiang Prevent. Med. **25**(7), 1–3 (2013)
3. Li, J.L., Zhang, L.W.: The study on the reaction time of athletes. J. Beijing Univ. Phys. Educ. **18**(3), 31–35 (1995)
4. Li, X.Y., Wu, X.Y., Han, L.P., Wei, Y.B., Wang, T.: The effects of acute moderate hypoxia on human performance of attention span and attention shift. J. Fourth Mil. Med. Univ. **20**(1), 71–73 (1991)
5. Zhang, C.: The effect of two-way-emotional condition on attention breadth. Zhejiang University (2012)
6. Rowe, G., Hirsh, B., Anderson, A.K.: Positive affect increases the breadth of attentional selection. Panas **104**(1), 383–388 (2007)
7. Lv, C.C.: An experimental study on the visual spatial cognitive ability and working memory span of college students. Northwestern University (2008)
8. Foss, D.J., Dowell, B.E.: High-speed memory retrieval with auditorily presented stimuli. Percep. Psychophys. **9**(6), 465–468 (1971). https://doi.org/10.3758/BF03208953
9. Cavanagh, J.P.: Relation between the immediate memory span and the memory search rate. Psychol. Rev. **79**(6), 525–530 (1972)
10. Wang, T., Hu, W.D., Li, X.J., Miao, D.M.: Design and reliability analysis of hidden figure test. J Fourth Mil. Med. Univ. **10**(supplement), 46–48 (1998)
11. King, R.E., Mcglohn, S.E., Retzlqff, P.D.: Female United States air force pilot personality: the new right stuff. Mil. Med. **162**(10), 695–701 (1997)
12. Zhou, Z., Lian, S.Q., Zhou, C.L.: Sex difference of students' ability to recognize spatial pattern. J. Math. Educ. **10**(4), 93–95 (2001)
13. Yan, X.J., Liu, K.H., Zou, B.H., Xiong, Y.B.: Experimental study of human depth perception based SPSS. J. Chongqing Univ. Sci. Technol. **15**(supplement), 49–52 (2013)
14. Liu, W.L.: Research on numeral short-memory ranges and processing strategies in college. J. Huinan Normal Univ. **6**(28), 89–91 (2004)
15. Wang, J.T., Hu, Q.H., Lv, J., Yan, Q.L., Su, H., Ma, J.: Comparative study of flight cognition ability between male and female flight cadets. J. Fourth Mil. Med. Univ. **25**(22), 2035–2037 (2004)

Neurophysiological Visual Classification Indicators in the Brain-Computer Interface

Sergey Lytaev[1,2](✉)

[1] St. Petersburg State Pediatric Medical University, St. Petersburg 194100, Russia
[2] St. Petersburg Federal Research Center of the RAS, St. Petersburg 199178, Russia

Abstract. The article presents the results of original research in the context of discussion of modern studies of the well-known psychological phenomenon of P300 evoked potentials in Brain Computer Interaction (BCI) systems. The aim of this research was to study the invariant processes of perception of the model "human-computer interaction" when classifying visual images with an incomplete set of features based on the analysis of the early, middle, late and slow components (up to 1000 ms) of event-related potentials (ERP). 26 healthy subjects (men) aged 20–26 years were investigated. Visual evoked potentials (VEPs) in 19 monopolar sites from the head surface according to the 10/20 system were recorded. The stimuli were a number of visual images with an incomplete set of features used in neuropsychological research. ERPs were analyzed at a time interval of 1000.0 ms from the moment of stimulation, using data from topographic brain mapping, as well as an assessment of the spatiotemporal characteristics of ERPs. Stepwise discriminant and factor analysis to establish the stability of ERPs parameters were applied. The results made it possible to establish that component N_{450} is the most specialized indicator of the perception of unrecognizable (oddball) visual images. The amplitude of the ultra-late components N_{750} and N_{900} is also higher under conditions of presentation of the oddball image, regardless of the location of the registration points.

Keywords: Event-related potentials · Visual evoked potentials · Wave P_{300} · Brain-computer interface · Oddball paradigm · Categorization of images

1 Introduction

Neuroimaging and neurophysiological methods enable the identification of neural substrates in the visual system that is invisible for behavioral studies. This has led to progress in understanding how extra-logical thinking (insight) emerges from basic cognitive mechanisms. Technologies to stimulate insight thinking are becoming available, including even direct brain stimulation interventions. One of the techniques that has both scientific and applied value is the BCI system. Modern BCI systems use a number of electrophysiological signals. Among them are visual evoked potentials, super-slow physiological processes (SSPP), alpha-, beta-, gamma-band EEG activity and the P_{300} component of evoked potentials [1, 4, 6, 8].

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 197–211, 2021.
https://doi.org/10.1007/978-3-030-77932-0_17

VEPs reflect the neurophysiological mechanisms underlying the processing of visual information and are the result of multiple summations (50–100) of visual stimuli on the screen, appearing at different frequencies.

SSPP—slow changes in electrical potentials generated in the cortex. A person can learn to control them, although, as a rule, this requires a lot of training. Once they achieve high levels of accuracy, they can use language support programs that allow users to select letters using a series of two choices. SSPP are in the same frequency range (less than 1 Hz) with the electro-dermal activity. The study of the SSPP dynamics showed that they reflect the level of stable functioning and are a physiological indicator that determines the state of brain structures and the course of a number of other bioelectric processes. To date, numerous studies have shown that SSPP are an adequate physiological method for studying the cerebral system for providing emotions and mental activity. By their nature, SSPP are close to electro-dermal reactions that make up the essence of a lie detector [9, 14].

EEG alpha and beta rhythms are activity in the primary sensory or motor areas of the cortex when they are not involved in sensory processing or motor output. A decrease in the amplitude of alpha and beta rhythms is associated with movement or preparation for movement, while an increase in the amplitude of these rhythms is associated with relaxation. The reduction is referred to as “event-related desynchronization” (ERD), while the increase is referred to as “event-related synchronization” (ERS). Interestingly, they also meet with motor images (with mental rehearsal of movement), which makes them relevant in BCI systems [10, 11, 13].

Finally, the P_{300} component of the evoked potentials can be used as control signals for the BCI. In a number of studies, P_{300} BCI systems have been tested in people with disabilities.

A 2D ball/cursor control system was tested with five disabled people. Up, right, down, and left arrows flashed randomly on the screen, and the user had to pay attention to an arrow pointing the desired direction. The target stimulus triggers the P_{300}, which can be recognized by the system and causes the ball to move in the desired direction. The results showed that the P_{300} is a suitable telltale signal for people with disabilities [23].

In another research, the authors [28] used the oddball paradigm with four choices (yes, no, entry, end) for visual, auditory or combined modalities. The subjects were required to focus on “yes” or “no” during each series of flashes. The results showed that P_{300}-based communication is possible for subjects with amyotrophic lateral sclerosis for both modalities. However, the average communication speed in both studies was relatively low (7.67 bit/min and 1.80 bit/selection on average, respectively).

However, it was shown that BCIs can be used based on the P_{300} with higher transmission rates [8]. The authors used six different images (TV, telephone, lamp, door, window, and radio) in their study with five people with disabilities. The subject had to silently calculate how often the prescribed image flashed. The results showed that the bitrate (speed) is in the range of 10–25 bit/min. This indicates that the P_{300} BCI can be effectively used by people with disabilities.

Event related potentials represent the brain activity recorded in response to events. ERP can be divided into two classes. Exogenous ERPs are the result of early automatic stimulus processing, while endogenous ERPs are the result of later, conscious stimulus processing. Conscious processing occurs with a delay of about 100 ms, when the visual signal is sent to the secondary projection zone of the visual analyzer – the parietal and temporal cortex [9, 15]. Although within 30–40 ms after the appearance of the stimulus, the visual signal already reaches the primary visual cortex the behavioral response is based on unconscious processing (reflexes) – in a period of 100 ms.

Endogenous ERP components are used for BCI systems, as they are associated with the recognition and attention of the subject, and can also reflect the motivation of the subject. Wave P_{300} is an ERP that has received a lot of attention in BCI research. Users (test subjects, patients) can manipulate the amplitude of the P300, paying more attention to a specific event. Therefore, the P_{300} is used in many BCI systems to reveal intentions or information hidden in the EEG.

There is a regularity – the lower the probability of an event occurring, the greater the P_{300} amplitude. To induce a reliable P_{300} wave, the probability of a target stimulus is usually set at around 10%. In addition, when many off-target stimuli precede the target stimulus, the P_{300} amplitude is higher than if a small number of off-target stimuli preceded the target stimulus. Also, if the time interval between two stimuli is large (long interval between stimuli), the amplitude of P_{300} is higher, while short intervals between stimuli lead to lower amplitudes.

On the one hand, the P_{300} is a typical answer to a given task, which is true for every person. But on the other hand, there are individual differences in the latency and amplitude of P_{300}, which influence the interpretation of the results.

In a line of studies of BCIs with P_{300} registration, a slight decrease in the productivity of information processing during sessions was found, which may be associated with an addictive effect [28]. Thus, multiple presentation of events leads to a decrease in the amplitude of the P_{300} and therefore to a decrease in performance.

During unexpected experimental sessions, the P_{300} amplitude spontaneously returns to the level of the first session [28]. On the other hand, it has been shown that the individual parameters of the P_{300} are relatively stable [24].

An important application for the P_{300}-based BCI system is the spelling device (P_{300}-speller), which allows patients with disabilities to communicate. BCI can provide an opportunity to answer simple “yes” or “no” questions at a speed of up to 20 bpm. In addition, BCI can provide slow text processing with a bit rate of about 2 wpm [33]. Despite the low transmission rate, it improves the quality of life of patients with amyotrophic lateral sclerosis.

BCI can be used to control a wheelchair. A prototype brain-guided wheelchair has been developed that uses an EEG/P_{300} signal and a motion control strategy to navigate safely and efficiently in a building [26]. With the P_{300} BCI system, the user can select a destination in the menu by counting the number of flashes of the destination. Further, the wheelchair moves to the selected and desired destination along a predetermined path. It is important, however, that the BCI system has localization information in order to find the correct path.

In addition to the wheelchair, an important application for people with severe motor impairments is the control of neuroprosthetic devices. BCIs can be used to control limb movements, for example, a robotic arm. It has been shown that BCIs based on the activity of cortical neurons are able to control three-dimensional movements of the robot arm [21, 31].

In theory all electrical devices connected to the BCI system can be controlled to one degree or another. In a home environment, it will be relevant to control light sources, music devices, TV, etc. The BCI system (environment controller) based on visual EPs in a stationary state, as well as control of a virtual apartment based on the P_{300} evoked potential, is described. The P_{300} can be used in a variety of environments, incl. virtual reality. While the goal of BCI technology is to provide users with disabilities with a communications or device management system, it also opens up new ideas for the consumer market, enabling future developments [5, 7, 12].

Some multimedia applications are developed and used by healthy people. Gamers are probably an important target group for using BCI systems. They are involved in new futuristic technologies [1] and often live in their own world when they are in the game. They can easily wear hats and invest in peripherals. The development of games integrated into BCI devices would open up a new gaming market.

BCI systems can also be of interest as an auxiliary interface used for military purposes, and even for specialized users such as surgeons, machinists or aircraft technicians. But before such applications can be implemented, BCI systems must be proven to be reliable, safe and useful.

The aim of this research was to study the invariant processes of perception of the model “human-computer interaction” when classifying visual images with an incomplete set of features based on the analysis of the early, middle, late and slow components (up to 1000 ms) of event-related potentials.

2 Materials and Methods

The research on 26 healthy subjects (men) aged 20–26 years were performed. Event-related potentials were recorded using a computer neuromapper in 19 monopolar sites from the surface of the head according to the 10/20 system with a reference electrode on the earlobe. The frequency bandwidth of the amplifier is 5–70 Hz. VEPs over 30 accumulations were averaged.

The investigated person was in a darkened screened room in a state of calm wakefulness, sitting in a comfortable chair. Stimulation was carried out by short-term (duration 0.3 ms, frequency 0.5 Hz) presentation of test images on the display screen, synchronized with the neuromapper computer. The distance from the observer to the screen is 2 m, the size of the images is 0.5×0.5 m.

The stimuli were two figures with an incomplete set of features (key and glasses). Our experience in testing both healthy and patients made it possible to use fragments of a key as a model of an identifiable image, the image of which usually does not cause difficulties (all 26 people were identified). On the contrary, the short-term presentation of image fragments of glasses, as a rule, is above the recognition threshold, even for healthy people with normal vision, which made it possible to use it as a model of an unrecognizable image (23 people were not identified). The image of key fragments was viewed as a simple cognitive stimulus. Fragments of glasses were considered as “oddball” visual image.

ERPs at a time interval of 1000.0 ms from the moment of stimulus delivery (see Fig. 1) using data from topographic brain mapping (in 19 sites) were analyzed, as well as an assessment of the spatio-temporal characteristics of evoked potentials (in 7 sites). The averaged ERP measured the amplitude (relative to the isoline) and peak latencies (PL) of all negative components (above the isoline) and one positive (below the isoline). The components have been designated N_{70}, N_{150}, P_{250}, N_{350}, N_{450}, N_{750} and N_{900}.

The amplitude-time characteristics of evoked potentials were processed using two main types of multivariate analyzes – stepwise discriminant (to establish the main differences in amplitudes and peak latencies) and factorial (to assess the stability of these indicators). The final result for stepwise discriminant analysis is the F-criterion. Differences were considered significant at $F > 4.0$. Factor analysis was chosen as the second method for determining the stable and unstable characteristics of the evoked potentials, where the criterion for determining the serial number of a factor is the stability of the parameter. Such an approach in physiological studies is justified from the point of view that dispersion, being an indicator of stability, is a characteristic of the quality of regulation in biological systems and, especially, in brain structures [14].

3 Results

VEPs in response to recognizable and unrecognizable stimuli in healthy subjects with subsequent statistical comparison of the results were recorded. Evaluation of the amplitude-temporal characteristics of the VEP components and the data of topographic brain mapping during EP registration in response to stimulation by visual structures with an incomplete set of features shows a certain type of potential deployment in time. In Fig. 1 shows a topomap of a 19-channel evoked potential with a waveform marking after 140 ms (N_{150}) from the moment of signaling.

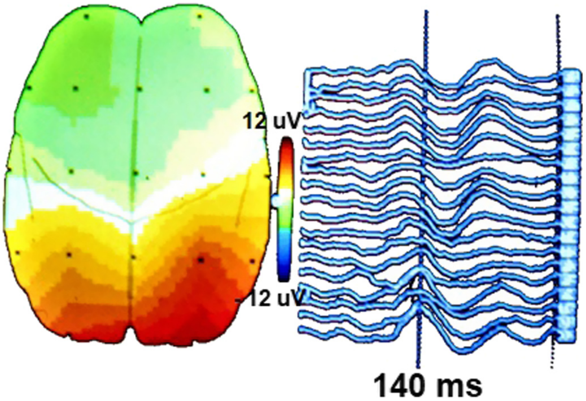


Fig. 1. Map of the visual evoked potential (at marking on 140 ms) and VEP distribution according to the 10–20 system in the perception of an identifiable visual image. Analysis time – 400 ms.

The maximum activation is recorded mainly in the posterior parts of the brain. In addition, we analyzed early negative waves with PL of 50–100 ms (N_{70}), intermediate components having a maximum amplitude at PL values of 120–200 ms (N_{150}), as well as late negative fluctuations in the interval of the analyzed epoch of 300–450 ms (N_{350} and N_{450}). After that, in the dynamics of VEP, there was certain "lull" corresponding to a time interval of 450–700 ms and representing an averaged electroencephalogram (see Fig. 2).

Following this "quiet" area, VEP waves also were noted, sometimes comparable in amplitude with the intermediate components and named by us N_{750} and N_{900} (see Fig. 2). Of the positive waves, only an oscillation with a PL of 200–300 ms (P_{250}) was analyzed, which, among other components located below the isoline had maximum amplitude and a more stable manifestation.

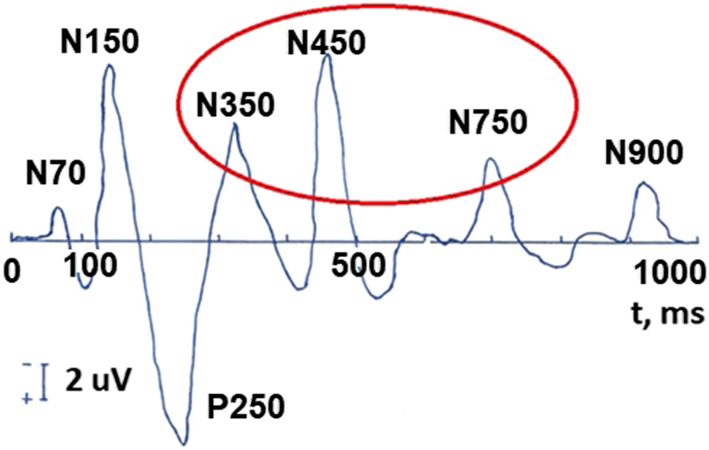


Fig. 2. Visual evoked potential (numbers – time, ms) according to the 10–20 system during the perception of an unrecognizable (oddball) visual image. Analysis time – 1000 ms.

The amplitude-time parameters of N_{70} are characterized by a stable manifestation only in the occipital leads. In other parts of the cerebral cortex, they are less pronounced, both in amplitude and in functional stability, and in the frontal leads sometimes they are absent altogether. At the same time, attention is drawn to the fact that the N_{70} amplitude in the central regions of the brain ($F > 4.0$) somewhat lightened during the perception of an unrecognizable image.

Brain mapping and processing of the amplitude-temporal parameters of the N_{150} component (see Fig. 3) shows the maximum activation of the anterior parts of the brain, as well as some time lag and a decrease in the amplitude (see Fig. 3) of this wave in the parietal and occipital leads. The amplitude characteristics of N_{150} are significantly higher in the central leads (C_4-C_3) during the perception of an identifiable object ($F > 10.0$), and the spatio-temporal parameters in this case have a stable manifestation. In the occipital and frontal areas of the cerebral cortex, the N_{150} amplitude slightly differs in both tasks ($F < 4.0$), while the presentation of an unrecognizable image is accompanied by more stable temporal parameters.

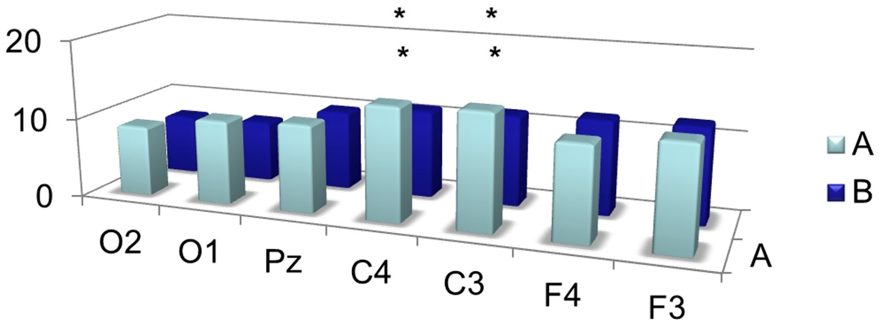


Fig. 3. Averaged values of amplitudes of the N_{150} ERP component during the perception of recognizable (A) and unrecognizable (oddball) (B) images. On the abscissa – points according to the 10–20 system, on the ordinate – amplitude, uV. Note - * - F > 4.0; ** - F > 10.0; *** - F > 20.0; in others cases – not significant (F < 4.0).

Regarding P_{250}, one can trace, on the one hand, a fairly high stability of temporal parameters, especially in tests with stimulation by a recognizable object, and on the other hand, there is almost complete absence of differences in amplitude characteristics (F < 4.0). The brain maps in both cases reflect the synchronous activation of all parts of the brain.

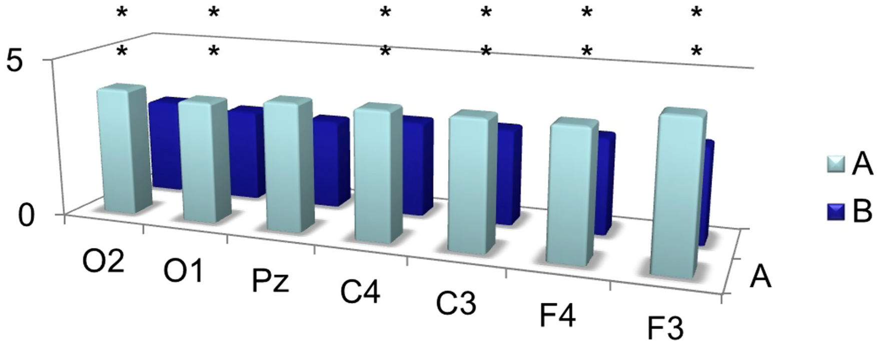


Fig. 4. Averaged values of amplitudes of the N_{350} ERP component during the perception of recognizable (A) and unrecognizable (oddball) (B) images. On the abscissa – points according to the 10–20 system, on the ordinate – amplitude, uV. Note – see Fig. 3.

The most significant differences, according to our data, are noted after 250–300 ms from the moment of stimulus presentation. Moreover, according to the degree of stability for all subsequent EP components (N_{350}, N_{450}, N_{750}, N_{900}), there is a high variability of peak latencies (low factor serial number) and at the same time, relative stability of amplitude characteristics (maximum factor serial number).

Thus, the amplitude of N_{350} (see Fig. 4) is significantly higher in the case of VEP registration in response to the identified signal. Attention is drawn to the fact of an increase in its amplitude from the occipital (O_2-O_1) to the central (C_4-C_3) and frontal leads (F_4-F_3). It is characteristic that a similar tendency towards an increase in amplitude from the occipital regions to the antero-central regions is also noted in the dynamics of N_{450} (see Fig. 5). If in the occipital-parietal leads (O_2, O_1, Pz) the values of the N_{450} amplitudes in both tests practically do not differ ($F < 4.0$), then in the central (C_4-C_3) and frontal leads they significantly increase ($F > 10.0$) with perception of an unrecognizable signal.

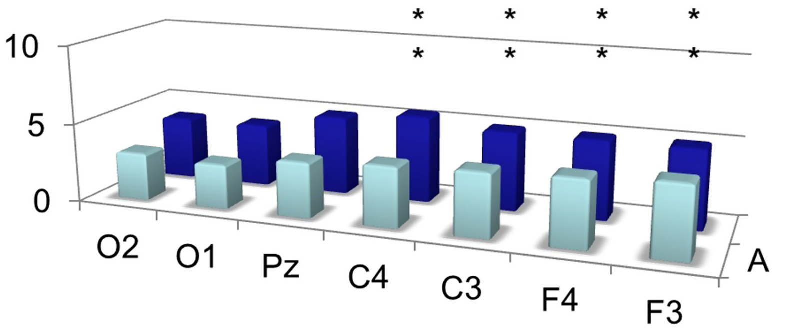


Fig. 5. Averaged values of amplitudes of the N_{450} ERP component during the perception of recognizable (A) and unrecognizable (oddball) (B) images. On the abscissa – points according to the 10–20 system, on the ordinate – amplitude, uV. Note – see Fig. 3.

The amplitude of the ultra-late components N_{750} (see Fig. 6) and N_{900} (see Fig. 7) is also higher under the conditions of presentation of an unrecognizable image, and regardless of the location of the registration points. At the same time, it is noted that if, upon perception of an identifiable object, the amplitudes of N_{750} and N_{900} practically do not differ from each other in different leads or even approach the isoline (see Fig. 6), then in the case of stimulation by an unrecognizable object, the amplitude of N_{750} is facilitated from the occiput to the forehead (see Fig. 7), and the dynamics of N_{900} is reversed (the amplitude is higher in the occipital electrodes).

4 Discussion

The history of studies of the P_{300} wave of evoked potentials dates back to the mid-60s of the last century. There are a huge number of publications on this topic. Summarizing these studies, it can be noted, firstly, that the P_{300} wave is an endogenous indicator of cognitive activity associated with decision-making, and secondly, this wave in real time is a complex of waves with a peak latency of 300 ms or more. And third, three paradigms are currently used to generate the P_{300} ERPs wave: one-stimulus, oddball, and three-stimulus. In each case, the subject is asked to monitor the achievement of the goal by pressing a button or making logical inferences followed by a report [3, 4, 18, 24, 28].

There are a line of indicators for assessing sensory processes. These include, in particular, the threshold of awareness of an external stimulus, characteristics of the orienting reflex, as well as bioelectric, autonomic, behavioral, emotional and other reactions [17]. It is believed that the greatest theoretical and methodological difficulties are caused by

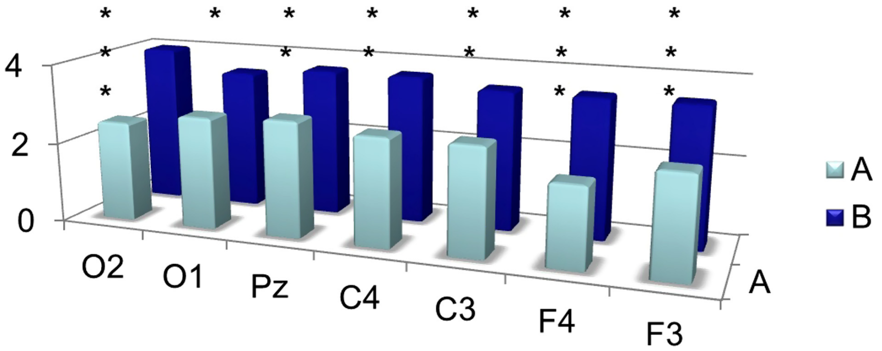


Fig. 6. Averaged values of amplitudes of the N_{750} ERP component during the perception of recognizable (A) and unrecognizable (oddball) (B) images. On the abscissa – points according to the 10–20 system, on the ordinate – amplitude, uV. Note – see Fig. 3.

the indication of the awareness of the stimulus, for the study of which there are a number of specialized tests associated with the registration of electrographic characteristics in response to conscious and unconscious stimuli [5, 13, 30].

Discussion of the obtained results is associated with the awareness of the presented stimuli with subsequent invariant recognition by the subjects. The use of images with an incomplete set of features for these purposes is a convenient model for assessing the studied mechanisms of visual perception, on the one hand, in BCI systems, and, on the other hand, in the study of the phenomenon of insight [1, 10, 11, 19].

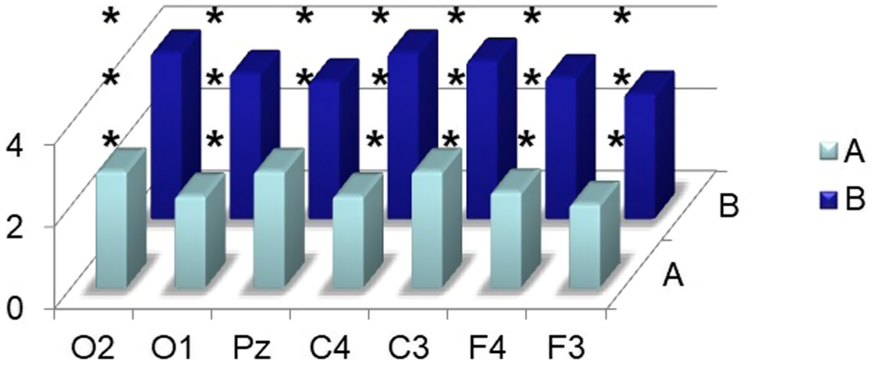


Fig. 7. Averaged values of amplitudes of the N_{900} ERP component during the perception of recognizable (A) and unrecognizable (oddball) (B) images. On the abscissa – points according to the 10–20 system, on the ordinate – amplitude, uV. Note – see Fig. 3.

The parameters of evoked bioelectric activity in time sequence have already been used in BCI systems for the past two decades. We proceed from the thesis that by origin, ERP components are divided into two types. Exogenous ERPs are an indicator of early automatic processing of the physical characteristics of external stimuli – the amplitude of stimuli (brightness, color, contrast, sound pressure level, etc.). Endogenous ERPs are the result of later processing of cognitive signals. Subconscious (reflex) and conscious processing begins to be carried out after 100–150 ms, when the visual signal is sent to the secondary projection zone of the visual system - the posterior parietal and inferior temporal cortex. Although the visual signal reaches the primary visual cortex 50–70 ms after the appearance of the stimulus, the behavioral response during this period is based on unconscious processing [5, 24].

Thus, early ERP components (50–100 ms) serve as indicators of the primary cortical response, reflecting the physical properties of the stimulus [16]. The analysis of the amplitude-time characteristics of the N_{70} in our study correlates with this thesis. With the same energy intensity of both stimuli, the N_{70} amplitude in the occipital electrodes slightly differs depending on the situational task. At the same time, the enhancement of this component in the central regions of the brain is characteristic under conditions of stimulation in an "oddball" image. This fact is associated with the initial activation of selective attention mechanisms in the associative areas of the brain.

The greatest degree of correlation between the configurations of evoked potentials in different areas of the cerebral cortex is observed at a time interval of 100–200 ms from the moment of signal delivery (see Fig. 3). This is associated with the universal optimal conditions for the exchange of neural information between these areas, regardless of the cognitive task. However, unrecognition of patterns is accompanied by significant suppression of the N_{150} amplitude, especially at the central registration points. Based on this, it is assumed that the initial increased activation of selective attention mechanisms during the perception of an oddball object is subsequently replaced by a less pronounced excitation of the associative neural network during the exchange of incoming and stored information.

The spatial-temporal distribution of the P_{250} and N_{350} VEP waves in both tests generally covers all scalp electrodes (see Fig. 4). The temporal characteristics of the P_{250} component are quite stable, in contrast to N_{350}, which obviously indicates identical mechanisms of image categorization. The perception of the identified image in comparison with the oddball way is regarded as the final invariant image recognition, in contrast to the opposite task, accompanied by a decrease in the amplitude of N_{350} with a subsequent increase in the amplitude of the later components.

The analysis of research shows that among the ERP components with PL more than 300 ms the most studied are oscillations in the interval 400–500 ms, called either N_{400} or Nx-Ny [21, 24]. ERP waves with a PL of about 300 ms are by now a paradigm for the indication of cognitive loads. The N_{400} component is considered a more specific answer. An increase in the N_{400} amplitude is associated with the perception of new cognitive information [18, 23]. In some studies, data show an increase in N_{400} amplitude upon activation of visual memory. There are publications where the N_{400} pattern was registered at the first presentation of a picture or a word, and, conversely, the amplitude was reduced upon repeated presentation of the object. On the basis of a series of studies that compared evoked responses upon stimulation with "new" and "old" stimuli, a number of authors hypothesized that N_{400}, in response to a new semantically significant stimulus, reflects the processes of "exhaustive search" of images in long-term memory for identification purposes [17].

The results of this research show that a wave with PL 400–500 ms (N_{450}) has a clear formation regardless of the perceptual task (see Fig. 5). The most pronounced amplitude is observed in the median and frontal leads when the "oddball" image is perceived. Similar data were also observed in studies with non-rhythmic presentation of words, where, in contrast to rhythmic presentation, the amplitude of N_{450} increased. Based on the foregoing, we can conclude about the specificity of negative waves with PL 400–500 ms reflecting the so-called "detailed process" or "exhaustive search" when identifying new information.

The cognitive significance of the later VEP components (more 600 ms) is more controversial in the authors' views. Some researchers believe that waves with a PL of 600–800 ms are an indicator of known (repetitive) images, scenes, faces [1, 18]. Sometimes the components of "late negative" ERP, like N_{400}, are associated with the processing of unfamiliar information [15]. The results of studies of the so-called reactive selective potential are of particular interest here. In these experiments, the answer was absent for simple types of the presented problems (response time up to 450 ms) and, conversely, increased in the time interval 600–800 ms with increasing test load complexity.

Our research data show that waves with PL of 700–800 ms (N_{750}) have a more pronounced manifestation in response to unrecognizable images in the associative cortex. It can be assumed that, by analogy with the N_{450}, the N_{750} component also reflects the "exhaustive search" processes for the subsequent categorization of the image.

The formation of VEP waves after 800 ms is also usually associated with the receipt of new information [17, 24]. This was especially clearly demonstrated in a series of studies with consistently presented stimuli [15]. With the presentation of sequential individual features of objects, first the components increased with PL 600–700 ms, and then later ones in the interval 800–1200 ms. In the present study, the N900 amplitude is more pronounced in response to "unusual" images, however, compared to the N_{450} and N_{750}, the N_{900} amplitude is greater at the occipital-parietal electrodes (see Fig. 6–7).

Thus, all ERP components that are formed 100–150 ms after the presentation of a visual image are considered endogenous by origin with the inclusion of conscious processes, categorization, decision-making, as well as under the influence of emotions and motivations. In this case, the endogenous components of ERP can act as indicators for BCI systems, since they are associated with the recognition and attention of the subject, as well as indicators of subconscious processes of extra-logical thinking. Besides concentration, fatigue and stressful conditions also affect the use of the BCI system [1, 8, 13]. However, if all scalp electrodes are used for BCI systems, the implementation of the insight processes has a more specific neural organization [2, 20, 25].

It is the P_{300} wave that is the ERP that has attracted a lot of attention in BCI research [28, 33]. Users (test subjects) can manipulate the amplitude of the P_{300}, paying more attention to a specific event. Therefore, the P_{300} is used in many BCI systems to reveal intentions or information hidden in the EEG [13]. There is the following regularity – the less the probability of an event (identification, categorization, decision making), the greater the amplitude of P_{300}. To induce a reliable P_{300} wave, the probability of the target stimulus is usually set at about 10% [26, 28].

The BCI system based on the P_{300} has a very interesting and important property in that it does not require initial user training in order to register the P_{300} in response to the desired goal. Thus, the user can immediately start using it. This BCI property has been developed in systems for wheelchair control and neuroprostheses [13, 26, 33]. There are a number of prototypes for driving a wheelchair, however, movement is rather limited. A prototype of a brain-guided wheelchair has been developed that uses an EEG/P_{300} signal and a motion control strategy to navigate safely and efficiently in a building. With the P_{300} BCI system, the user can select a destination in the menu by counting the number of flashes of the destination.

An important application for the P_{300} BCI system is the spelling device [26], which allows patients with disabilities to communicate. BCI can provide an opportunity to answer simple “yes” or “no” questions at a speed of up to 20 bpm. In addition, BCI can provide slow text processing with a bit rate of about 2 wpm [28]. Despite the low transmission rate, it improves the quality of life of ALS patients.

Thus, on the one side, the P_{300} is a typical answer to the task at hand, which is true for every person. But on the other side, there are individual differences in the latency and amplitude of P_{300}, which influence the interpretation of the results.

Insight studies have shown the activation of multicomponent processes and corresponding neural substrates, and some of them are subject to minor changes in attention, emotion, motivation and other factors. One can only guess what the future will open up studies of insight-related individual differences in neuroanatomy, cytoarchitectonics and genetics [27, 29, 32]. The psychopharmacology of insight and creativity, currently largely unexplored, promises to contribute both to our scientific understanding of insight and to methods for improving it.

Insight or extra-logical thinking occurs when a person suddenly rethinks a stimulus, situation, or event in order to produce a non-obvious, non-dominant interpretation. This can take the form of solving a problem (“aha” moment), understanding a joke or metaphor, or recognizing ambiguous perceptions. Research into insight began more than 100 years ago, but neuroimaging and electrophysiological methods have only been applied to their study in the past two decades [2, 10, 11]. Recent work has highlighted insight-related coarse semantic coding in the right hemisphere and internally focused attention before and during problem solving. Individual differences in propensity to problem-solving out-of-logic, rather than consciously and analytically, are associated with different patterns of resting brain activity [11]. In short, the cognitive neuroscience of understanding is an exciting new area of research related to fundamental neurocognitive processes.

Neuroimaging insight studies highlight neural correlates from both fMRI and EEG data [10, 22]. The EEG has excellent temporal resolution, but limited spatial resolution. In contrast, fMRI has excellent spatial resolution, but limited temporal resolution, so it is best suited for localizing a neural event in space. At the same time, these methods were able to identify neural correlates of insight both in space and in time.

When solving the problem intuitively by a specific subject, compared with solving identical problems by analytical processing, the EEG shows a burst of high-frequency (gamma-range) activity over the right temporal lobe, and fMRI shows the corresponding change in blood flow in the right medial anterior-superior temporal gyrus [2, 11]. In the original fMRI experiment, this right temporal region was the only area exceeding strict statistical thresholds, but weak activity was found in other areas, including the hippocampus and parahippocampal gyri, anterior and posterior cingulate cortex. In a later study with more participants and stronger imaging modalities [30], the same network of regions far exceeded the critical statistical threshold, with the right anterior temporal region being the strongest again.

The spatial and temporal consistency of fMRI and EEG researches [11] suggested that these results were caused by the same underlying brain activation. The response of the right temporal brain was identified as the main neural correlate of insight indication, because (a) it occurred at about the moment when the participants realized the solution to each of these problems, (b) the same area is involved in other tasks requiring semantic analysis and integration; and (c) gamma-band activity has been assessed as a binding mechanism for conscious information.

The results of the present study on ERP registration data show the similarity of the tests with the correct recognition of fragments of glasses (oddball images) and dual images. At the intermediate stage of perception (100–200 ms), in both cases, the activity of the central and frontal cortex decreases, mainly in the left hemisphere. At the later stages of information processing (300–500 ms), the temporal-parietal and occipital parts of the brain on the right are activated, with the difference that when dual objects are perceived, this process is extended to 700–800 ms with the activation of the central and occipital fields of the right hemisphere.

Other electrophysiological data suggest that even from two opposite ERP results, it is possible to trace the features of internal and external insight. Internal insight is associated with positive ERP components after stimulus onset ($P_{200–600}$) above the superior temporal gyrus. External insight is associated with a negative component of ERP (N_{320}). These results already show that external and internal insights differ at the behavioral and neurophysiological levels. Various neurobiological insights suggest that presenting a solution or a decision clue leads to the same “aha” moment as attempting an inner decision [19, 22, 27, 29].

5 Conclusion

For the amplitude-temporal characteristics of ERP in tasks with the correct recognition of fragments of glasses (oddball images) and double images that simulate insight similar results were noted. At the intermediate stage of perception (100–200 ms), in both cases, the activity of the central and frontal cortex decreases, mainly in the left hemisphere. At the later stages of information processing (300–500 ms), the temporo-parietal and occipital parts of the brain on the right are activated, with the difference that when double objects are perceived, this process expands to 700–800 ms with the activation of the central and occipital electrodes.

A negative wave with a peak latency of 400–500 ms (N_{450}) of ERP is the most specific indicator among all the components of the evoked response when perceiving “oddball” visual images.

The N_{350} component of ERP has large amplitude in the case of correct image recognition, which is regarded as the completion of the processes of invariant signal estimation at this stage.

Late negative ERP waves with a peak latency of 600–1000 ms are more significantly pronounced in the perception of an unrecognizable image and are a reflection of the mechanisms of "exhaustive search" in order to categorize the incoming information.

References

1. Allison, B., Graimann, B., Gräser, A.: Why use a BCI if you are healthy? BRAINPLAY 07 Brain-Computer Interfaces and Games Workshop at ACE. In: Advances in Computer Entertainment, Salzburg, Austria, pp. 7–11 (2007)
2. Bowden, E.M., Jung-Beeman, M., Fleck, J., Kounios, J.: New approaches to demystifying insight. Trends Cogn. Sci. **9**(7), 322–328 (2005)
3. Dietrich, A., Kanso, R.: A review of EEG, ERP, and neuroimaging studies of creativity and insight. Psychol. Bull. **136**(5), 822–848 (2010)
4. Donchin, E., Spencer, K.M., Wijesinghe, R.: The mental prosthesis: assessing the speed of a p300-based brain-computer interface. IEEE Trans. Rehabil. Eng. **8**(2), 174–179 (2000)
5. Farwell, L.A., Smith, S.S.: Using brain MERMER testing to detect knowledge despite efforts to conceal. J Forensic Sci. **46**(1), 135–143 (2001)
6. Haider, A, Fazel-Rezai, R.: Application of P300 event-related potential in brain computer interface. In: Event-Related Potentials and Evoked Potentials, Phakkharawat Sittiprapaporn, IntechOpen, London (2017)
7. Hill, N.J., Lal, T.N., Bierig, K., Birbaumer, N., Schölkopf, B.: An auditory paradigm for brain-computer interfaces. In: Advances in neural information processing systems. MIT Press, Cambridge, MA, USA, pp. 569–576 (2005)
8. Hoffmann, U., Vesin, J., Ebrahimi, T., Diserens, K.: An efficient P300-based brain–computer interface for disabled subjects. J. Neurosci. Methods **167**(1), 115–125 (2008)
9. Khil'ko, V., Shostak, V., Khlunovskiĭ, A., et al.: The topographic mapping of evoked bioelectrical activity and other methods for the functional neural visualization of the brain. Vestn. Ross. Akad. Med. Nauk **48**(3), 36–41 (1993)
10. Kounios, J., Beeman, M.: The Aha! moment: the cognitive neuroscience of insight. Curr. Dir. Psychol. Sci. **18**(4), 210–216 (2009)
11. Kounios, J., Beeman, M.: The cognitive neuroscience of insight. Annu. Rev. Psychol. **65**(1), 71–93 (2014)
12. Kuss, D.J., Griffiths, M.D.: Internet and gaming addiction: a systematic literature review of neuroimaging studies. Brain Sci. **2**(3), 347–374 (2012)
13. Levi-Aharoni, H., Shriki, O., Tishby, N.: Surprise response as a probe for compressed memory states. PLoS Comput. Biol. **16**(2), e1007065 (2020)
14. Lytaev, S.: Modeling and estimation of physiological, psychological and sensory indicators for working capacity. Adv. Intell. Syst. Comput. **1201**, 207–213 (2021)
15. Lytaev, S., Shevchenko, S.: VEPs and AEPs mapping of occlusive lesions in cerebral vessels. Ann. NY Acad. Sci. **821**(1), 524–528 (1997)
16. Lytaev, S., Aleksandrov, M., Popovich, T., Lytaev, M.: Auditory evoked potentials and PET scan: early and late mechanisms of selective attention. Adv. Intell. Syst. Comput. **775**, 169–178 (2019)
17. Lytaev, S., Aleksandrov, M., Lytaev, M.: Estimation of emotional processes in regulation of the structural afferentation of varying contrast by means of visual evoked potentials. Adv. Intell. Syst. Comput. **953**, 288–298 (2020)
18. Lytaev, S., Vatamaniuk, I.: Physiological and medico-social research trends of the wave P300 and more late components of visual event-related potentials. Brain Sci. **11**(1), 125 (2021)

19. Luo, J., Niki, K.: Function of hippocampus in "insight" of problem solving. Hippocampus **13**(3), 316–323 (2003)
20. Metuki, N., Sela, T., Lavidor, M.: Enhancing cognitive control components of insight problems solving by anodal tDCS of the left dorsolateral prefrontal cortex. Brain Stimul. **5**(2), 110–115 (2012)
21. Nijboer, F., Sellers, E.W., Mellinger, J., et al.: A P300-based brain–computer interface for people with amyotrophic lateral sclerosis. Clin. Neurophysiol. **119**(8), 1909–1916 (2008)
22. Qiu, J., et al.: The neural basis of insight problem solving: an event-related potential study. Brain Cogn. **68**(1), 100–106 (2008)
23. Piccione, F., Giorgi, F., Tonin, P., et al.: P300-based brain computer interface: reliability and performance in healthy and paralysed participants. Clin. Neurophysiol. **117**(3), 531–537 (2006)
24. Polich, J.: Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. **118**(10), 2128–2148 (2007)
25. Reber, P.J., Kounios, J.: Neural activity when people solve verbal problems with insight. PLoS Biol. **2**(4), 500–510 (2004)
26. Rebsamen, B., et al.: Controlling a wheelchair indoors using thought. IEEE Intell. Syst. **22**(2), 18–24 (2007)
27. Rothmaler, K., Nigburb, R., Ivanova, G.: New insights into insight: Neurophysiological correlates of the difference between the intrinsic "aha" and the extrinsic "oh yes" moment. Neuropsychologia **95**(1), 204–214 (2017)
28. Sellers, E., Donchin, E.: A P300-based brain–computer interface: initial tests by ALS patients. Clin. Neurophysiol. **117**(3), 538–548 (2006)
29. Shen, W., et al.: Right hemispheric dominance of creative insight: an event-related potential study. Creat. Res. J. **25**(1), 48–58 (2013)
30. Subramaniam, K., Kounios, J., Parrish, T.B., Jung-Beeman, M.: A brain mechanism for facilitation of insight by positive affect. J. Cogn. Neurosci. **21**(3), 415–432 (2009)
31. Van Dinteren, R., Arns, M., Jongsma, M.L.A., Kessels, R.P.C.: P300 Development across the Lifespan: a Systematic review and meta-analysis. PLoS ONE **9**, 0087347 (2014)
32. Wagner, U., Gais, S., Haider, H., Verleger, R., Born, J.: Sleep inspires insight. Nature **427**, 352–355 (2004)
33. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., et al.: Brain–computer interfaces for communication and control. Clin. Neurophysiol. **113**(6), 767–791 (2002)

An Evaluation of Two-Dimensional Digital Input Models for Mathematical Structure: Effects on Working Memory, Cognitive Load, and Efficiency

Francis Quinby[1](✉), Seyeon Kim[2], Marco Pollanen[1], Wesley S. Burr[1], and Michael G. Reynolds[3]

[1] Department of Mathematics, Trent University, Peterborough, ON, Canada
francisquinby@trentu.ca

[2] Department of Psychology, Carleton University, Ottawa, ON, Canada

[3] Department of Psychology, Trent University, Peterborough, ON, Canada

Abstract. Typesetting 2-dimensional mathematical notation can present challenges to users who rely upon WYSIWYG (what you see is what you get) word processing editors, which allow for direct manipulation of text. These editors use various models to represent 2-dimensional mathematical structure within the 1-dimensional word processing environment. The 2-dimensional nature of mathematical notations manifests itself in a variety of ways, and we hypothesize that two distinct models, structure-based (Microsoft Word Equation Editor) and free-form (MC^2: Mathematics Classroom Collaborator) handle different types of mathematical structure with varying degrees of success. To test this hypothesis, an eye-tracking study was conducted to compare how these two models affect task efficiency for mathematical "expert" and "novice" users, as well as working memory interference, and cognitive load. The study required users to transcribe mathematical expressions containing three types of structure: linear (1-dimensional), exponential, and rational (fractions). Handwriting was used as a control. Results showed superior performance by the structure-based model for the transcription of fractions, while the free-form model displayed ameliorated performance for the transcription of exponents. Handwriting was found to be significantly more efficient, but cognitive effects were inconclusive. Few differences were found with respect to user mathematical experience level. These findings show evidence that neither of these models is superior for the typesetting of all mathematics, but rather that features of each model are better equipped to handle different mathematical structures. Therefore, word processing editors can improve the facility of typesetting 2-dimensional mathematics by incorporating elements of both tested models to improve the overall user experience. Such optimization will ultimately facilitate the digital learning and communication of mathematical content.

Keywords: Eye-tracking · Working memory · Mathematical software · Software user interfaces

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 212–222, 2021.
https://doi.org/10.1007/978-3-030-77932-0_18

1 Introduction

The digital typesetting of mathematical content poses a large challenge to many individuals, especially novice-level users. While a number of software applications have the ability to do so they are not without problems in efficiency and effectiveness, or they present users with a steep learning curve. For these reasons, handwritten submissions to mathematics assignments are generally accepted in academic settings at levels up to and including undergraduate university classes [14]. While this presents minimal problems for in-person classes, it poses a greater challenge for online educational formats. The saliency of this issue has been compounded by the COVID-19 pandemic which has forced many nations to migrate education to an online format. As a result, many educators and students have been forced to re-imagine traditional mathematics classes [5], and typesetting (more complex) mathematical content digitally has proved challenging to many [3]. The issues raised by the difficulty of digital typesetting are unlikely to go away once the current pandemic has ended, as online education has been increasing in popularity worldwide and advances in internet technology has greatly increased access to education, a trend that is unlikely to cease. Thus, it is vital that the user-experience of software applications designed for the typesetting of 2-dimensional mathematics be optimized in order to maximize the digital learning and communication of mathematical content.

The challenges mentioned above arise largely due to the fact that mathematical notation is often composed of 2-dimensional structures which can be difficult to represent in word processing software applications. Solutions to this issue can be conceptualized as falling into two categories: LaTeX (or similar), in which 1-dimensional input strings are compiled into 2-dimensional mathematical structures; and WYSIWYG (What You See Is What You Get) editors which allow users to directly edit mathematical output. Due to the steep learning curve of languages and interfaces like LaTeX, most beginners rely solely upon WYSIWYG editors despite limitations of available features. WYSIWYG editors can be classified based on the model chosen to represent 2-dimensional mathematical structure within the word processing environment [2]. One such model is a structure-based model, in which 2-dimensional "shells" are inserted into the workspace, e.g., a fraction would be inserted as a horizontal line with empty squares above and below (the model of Microsoft Word Equation Editor, MWEE) such as is shown in Fig. 1. An alternative is the free-form model that allows users to drag and drop mathematical symbols and structures into a 2-dimensional workspace. In this model structures do not have to be inserted by the user. Instead they are created by manipulation of the spatial configuration of the symbols in the expression, e.g., dragging a 2 into superscript position squares an element. Due to the widespread use of WYSIWYG editors for academic [10] and other applications, it is important to maximize the user-experience of such applications for typesetting mathematical content by optimizing the elements which reduce their overall efficiency and effectiveness. However, to do so these problematic elements must first be identified.

Fig. 1. Fraction structure initially inserted into Microsoft Word Equation Editor workspace (left) and populated with 1 as numerator and x as denominator in the expression $\frac{1}{x}$ (right).

Limited research has examined how the use of these different models affect the ease of typesetting mathematics [1,2,8]. Furthermore, to our knowledge no studies have examined if and how these models might affect the ease of transcription of different types of mathematical structures, as these can be represented symbolically in markedly different manners (e.g., fractions and exponents). Past research has shown that expressions containing 2-dimensional structure are entered into the user workspace in a notably different sequential order when using a free-form model versus a structure-based model based on the types of structures within the expression [12], which leads us to believe that there may also exist a difference in the ease of transcribing different mathematical structures with these two models. The answer to such an inquiry is crucial to improving the user-experience of applications that make use of digital input models for mathematical structure and should be considered by software developers when creating and improving these applications.

In order to address this issue, we report an eye-tracking experiment that compares elements of efficiency and ease of use of the structure-based and free-form models when typesetting mathematical expressions containing different types of 2-dimensional mathematical structure. MWEE was used to represent the structure-based model while MC^2 [4], an online mathematics communication application, was used to represent drag and drop (free-form) models. Handwriting was used as a control. The use of eye-tracking technology allowed for collection of metrics related to user cognitive load and working memory interference, two important factors for software applications used in educational settings. Past eye-tracking research has shown that mathematical expertise affects how individuals read and process mathematical material [6,7,9], so participant mathematical expertise was also considered.

Based on previous findings from Quinby *et al.* [12], we hypothesized that the structure-based model and the free-form model would show differences in performance when being used to transcribe different types of 2-dimensional mathematical structure. We also hypothesized, based on findings from Kim *et al.* [6,7], that individuals with greater mathematical expertise would show improved performance when transcribing mathematical content, regardless of medium, due to increased familiarity with mathematical symbology.

2 Method

The resulting experiment was a mixed-design factorial experiment with mathematical expression type and transcription method within-subjects factors and participant mathematical expertise as a between-subjects factor. Mathematical expression type consisted of three factor levels chosen to represent structure commonly found in mathematics: linear, exponent, and fraction. Transcription method was composed of three factor levels chosen to represent two common structural models for the digital representation of 2-dimensional mathematics as described above, MWEE (structure-based model) and MC^2 (free-form model), as well as pen and paper (handwriting) as a control. Finally, participant mathematical expertise was a two-level factor, with participants divided into mathematical "experts" and "novices". Experts were defined as individuals who were majoring in either mathematics or physics, or had taken at least four undergraduate or graduate mathematics courses. Anyone who did not meet this criteria was defined as a novice. The participants ($n = 26$) consisted of undergraduate and graduate students from Trent University in Peterborough, Ontario, Canada with 12 classified as "experts" ($n = 12$) and 14 as "novices" ($n = 14$).

The experimental task consisted of transcribing 36 mathematical expressions, 12 while using each of the three transcription methods. The stimuli consisted of 12 mathematical expressions for each of three different structure classes and, therefore, participants transcribed four expressions of each structure class with each of the three transcription methods. Examples of each expression structure class are shown in Table 1. All stimuli contained four variables, represented by Roman letters. One variable was positioned on the left-hand side of the equation, and three on the right-hand side of the equation with coefficients $\neq 1$. Therefore, linear expressions contained 10–11 symbols, fractions contained 10–12 symbols and exponents contained 13–14 symbols. The powers of the exponents were restricted to the digits 4–9 in order to reduce possible memory effects from individuals cognitively encoding the exponent in components such as x^2 and x^3 as "squared" and "cubed" which might reduce working memory demands due to familiarity and grouping. Individual stimuli were randomly generated for each participant and were presented in a psuedo-random order to control for possible learning effects.

Table 1. Examples of the three classes of mathematical structure within stimuli used in experiment.

Structure	Example
Linear	$k = 2l - 7h - 3g$
Exponent	$x = -4j^9 + 7i^4 + 9m^6$
Fraction	$l = -\frac{4q-3c}{7i}$

Prior to the commencement of experimental trials, participants received a detailed description of the experimental procedure, were shown an instructional video on how to transcribe the three classes of expressions using the two digital transcription methods, and completed four practice trials (2 with each digital method). The experimental stimuli were shown to participants on a flat screen desktop monitor which they were then instructed to transcribe digitally on a laptop, or by hand with pen and paper. Upon completion of each trial, participants focused their gaze on a coloured sticker for approximately three seconds and gave a verbal indication that they were indeed finished.

Participants wore a ViewPoint PC-60 EyeTracker® (Arrington Research, Inc.) head mounted eye-tracker throughout the experiment, allowing for the recording of eye movements during experimental tasks from which metrics for working memory interference and cognitive load were extracted. Eye movements are divided into two types of events: *fixations* (when the eye fixates on a given location) and *saccades* (rapid eye movement from one fixation location to another) [17]. The duration and number of fixations has been shown to correspond with cognitive processing, and eye-tracking technology has been used extensively in mathematical education research [18].

Inferences about cognitive load were based on the mean duration of fixations on the user workspace during the transcription task, as past research has shown fixations of longer duration occur during periods of increased cognitive load [11]. A link between working memory and mathematical performance has been established [16] and if interference is occurring as a result of the use of a certain typesetting application, it may have an effect on the encoding of mathematical material. Therefore, the number of eye-gaze transitions from the presented stimulus to the user workspace during the experimental task was used as a measure of working memory interference. This metric provided a count of the number of times that the participant referred to the stimulus during the replication task, or the number of memory encoding events necessary to replicate the expression. Furthermore, to provide additional insight into working memory effects we measured the mean eye-gaze dwell time on the stimulus. If a participant required fewer eye-gaze transitions but showed an increased mean dwell time on the stimulus, then it could be assumed that they were encoding more information into working memory during each group of fixations on the stimulus which could confound effects on working memory interference. However, if variability was observed within the number of transitions, but mean dwell time on the stimulus was consistent between factors, then working memory interference can be attributed with greater confidence.

3 Results

In a manner similar to that used in Quinby *et al.* [13], eye-tracking output video recordings and corresponding data had to be prepared for quantitative analysis. This consisted of the categorization of gaze locations throughout the experimental tasks, by manually coding fixation locations as occurring in one of the following locations: workspace, stimulus (presented mathematical expression), or indeterminate. The precise moment that the stimulus was presented for transcription on the monitor was recorded as the start time. A trial ended when participants focused their gaze at the prespecified sticker location. All analyses were completed using R 3.6.0 [15] with figures created with ggPlot2 [19]. Differences at $p < 0.05$ were considered to be statistically significant.

3.1 Task Efficiency

Task efficiency was assessed by examining the time to complete each transcription (Table 2). There was an interaction between transcription method and expression type on the time to transcribe ($F_{4,22} = 179.5, p < 0.001$). Follow-up contrasts showed that transcribing of exponents was more efficient for the free-form model than the structure-based model ($t_{25} = 15.46, p < 0.001$). However, the opposite was true for fractions, where the structure-based model was more efficient than the free-form model ($t_{25} = 8.24, p < 0.001$). Consistent with findings from Quinby *et al.* [13], handwriting mathematical expressions was found to be significantly more efficient than transcribing them digitally ($t_{25} = 13.99, p < 0.001$). Participant expertise was not found to have a significant effect on mean completion time. Effects on task efficiency (Fig. 2) were found to be relatively consistent with those observed for working memory interference (Fig. 4) and cognitive load (Fig. 3).

Table 2. Mean completion time in experimental trials (seconds).

Structure class	Transcription method	Mean	SE
Linear	Free-form (MC^2)	12.52	3.411
	Structure-based (MWEE)	12.90	5.520
	Handwritten	9.66	3.091
Linear	Free-form (MC^2)	24.51	9.048
	Structure-based (MWEE)	39.21	10.720
	Handwritten	12.96	4.287
Fraction	Free-form (MC^2)	31.98	12.69
	Structure-based (MWEE)	20.93	7.498
	Handwritten	11.65	3.735

3.2 Cognitive Load

Inferences about cognitive load were based on mean fixation duration. As can be seen in Fig. 3, mean fixation duration was largest in the pen and paper condition suggesting that participants were able to undertake more work during

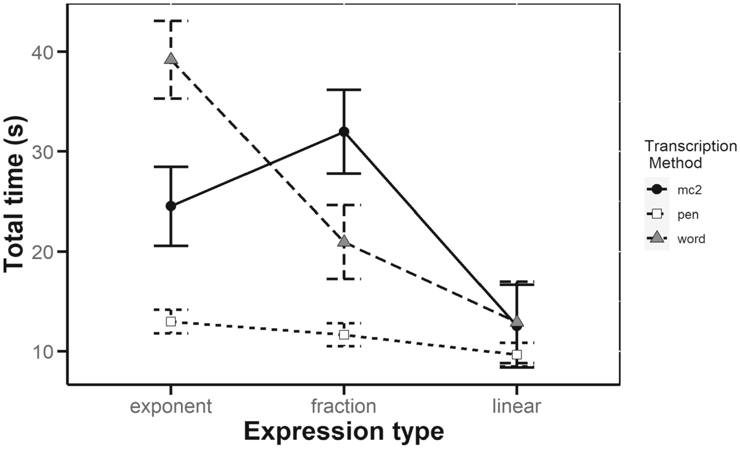


Fig. 2. Interaction plot: effect of expression type and transcription method, on total time with 95% CI.

each fixation than with the other two interfaces. Critically, there was a significant interaction between expression type and transcription method on mean fixation duration ($F_{4,22} = 18.56, p < 0.001$) consistent with both of these factors affecting cognitive load. Follow-up contrasts showed mean fixation durations were shorter for the free-form method than for the structure-based method ($t_{25} = 5.90, p < 0.001$) when transcribing exponents consistent with less cognitive load (less work being done) per fixation in the free-form condition. Interestingly, the opposite pattern was found for fractions ($t_{25} = 6.46, p < 0.001$), suggesting that fewer resources were available when using the free-form model compared to the structure-based model to encoding information for fractions. Mean fixation duration was affected by participant expertise suggesting that this factor affected the available cognitive load ($F_{1,25} = 7.169, p < 0.05$). As expected, experts had a greater mean fixation duration ($\bar{x} = 0.804s$, SE = 0.009) than novices ($\bar{x} = 0.611s$, SE = 0.013), suggesting that experts were able to do more work per fixation than novices.

3.3 Working Memory Interference

There was a main effect of transcription method on the number of transitions between the original expression and the writing location ($F_{2,24} = 4.782, p < 0.05$). This is consistent with transcription method interfering with working memory. In contrast, there was no effect of transcription method on eye-gaze dwell time on the stimulus ($F_{2,24} = 3.084, p = 0.055$). This suggests that the transcription method affected the size of the units of information encoded on each transition. These main effects were qualified by an interaction between transcription method and expression type for both number of transitions ($F_{4,22} = 16.74, p < 0.001$) and mean dwell time on the stimulus ($F_{4,22} = 4.90, p < 0.01$;

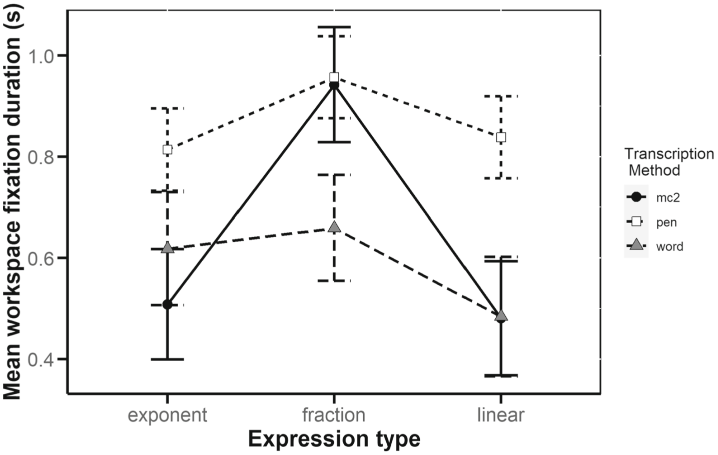


Fig. 3. Interaction plot: effect of expression type by transcription method on mean duration of fixation on workspace with 95% CI.

Fig. 4: right). As can be seen on the left hand side of Fig. 4, there was no effect of transcription method in the number of transitions for linear expressions or fractions, but there was for expressions with exponents. Importantly, there were fewer transitions between the stimulus and user workspace when transcribing the exponent expressions for the free-form model ($\bar{x} = 6.70$, SE $= 1.77$) than the structure-based model ($\bar{x} = 7.13$, SE $= 1.95$; $t_{25} = 2.928, p < 0.01$). This is consistent with the structure-based model interfering with working memory when transcribing exponents. Analysis of the dwell time on the source when transcribing the exponents revealed that there was no significant difference between the free-form model and the structure-based model ($t_{25} = 0.990, p = 0.332$). This pattern suggests that transcribing expressions containing exponents with the structure-based model resulted in more working memory interference than when transcribing similar expressions with the free form model. Together, the transition and dwell time data suggest that participants compensated for the working memory interference when using the structure based model by encoding smaller units of information into working memory. No significant differences in either working memory interference measures for the fraction expression type were observed. Finally, both the main effects of mathematical expertise as well as any interactions involving this factor were found not to be statistically significant.

In summary, we can see that the free-form model provides users with an environment which is more efficient for the transcription of exponents in mathematical expressions compared to the structure-based model while the opposite is true for the fraction structure. Altogether, these results imply that elements of the free-form model allow users to transcribe exponents with greater efficiency and reduced cognitive resources than when using a structure-based model, while the same structure-based model provides similar benefits for transcribing fractions compared to a free-from model.

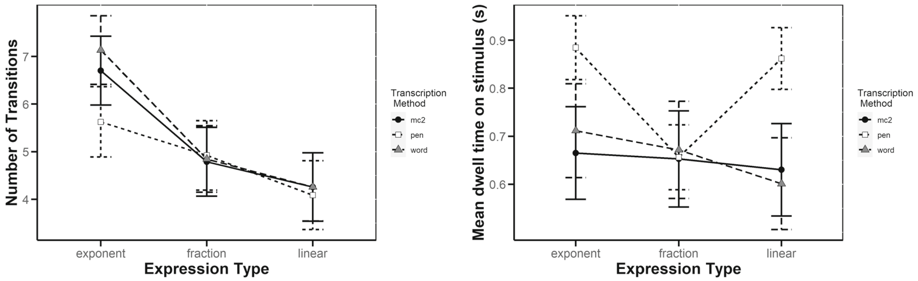


Fig. 4. Plots of interaction effects of expression type and transcription method on number of transitions (left) and mean dwell time on stimulus (right) with 95% CI

3.4 Handwriting (control)

Though handwriting the expressions (stimuli) showed a number of significant effects on all three outcome performance measures, we chose not to report these in great detail for a number of reasons. The cognitive effects relating to working memory interference and cognitive load may be less significant and show evidence of the employment of an alternate cognitive strategy by participants during handwritten trials, due to the following: the handwritten condition resulted in significantly greater mean dwell time on the presented stimulus compared to the other transcription methods. This would seem to imply that, while handwriting expressions, participants spent more time encoding the expression into working memory during each group of fixations on the stimulus because fewer cognitive resources were needed for transcription of the expression, compared to digital transcription. We do not discount that this may, in fact, have been the case for some participants. However, from experimenter observation it was noted that multiple participants were able to write out large parts of the presented expressions by hand, while their eye-gaze remained fixed on the stimulus. The result of this was an increased dwell time on the stimulus and a reduced number of eye-gaze transitions between the stimulus and workspace. This strategy would seem to employ fewer working memory resources and perhaps greater physical coordination by participants, and for this reason, comparisons drawn between the digital transcription conditions and the handwritten condition with respect to working memory interference are dependent on the cognitive load resources required for the transcription act itself.

4 Conclusion and Future Work

In conclusion, we have shown clear support for the hypothesis that the two different models for digital representation of 2-dimensional mathematical structure in software applications differ in their handling of different classifications of this structure. Our results show that neither model for representing mathematical structure is superior for all mathematical expressions: both handle different types of mathematical structure with varying degrees of success. Specifically, the free-form model is more efficient and effective for transcribing exponents while the structure-based model is more efficient for the transcription of fractions. However, our study failed to show any significant effects related to mathematical expertise on the ease of use of software applications for transcribing mathematics. Based on this finding, knowledge of the mathematical content to be transcribed neither increases efficiency of the process nor decreases cognitive interference effects during the process.

There were a number of limitations present in this study which are important to note. First of all, while this experiment was designed to test differences in performance between these two models, the models exist within unique user interfaces. Certain elements of these interfaces may differ, such as the size of items contained within menus that symbols are inserted from as well as the spatial configuration of these menus. These differences could have contributed to the effects observed. To truly isolate effects of the underlying models themselves, it would be necessary to make use of two programs with (nearly) identical user interfaces, but which use different models to represent the mathematical structure. Another possible limitation which may have contributed to the lack of effect of mathematical expertise on measures of performance during the use of these software applications is the criteria used to define an individual as an “expert”, and within the subset of students who met this criteria there was likely a wide range of true expertise. Finally, students likely had varying levels of familiarity with the software applications used within the experiment. It is likely that many students were quite familiar with the user interface of Microsoft Word due to its high level of market penetration, and a number of the participants likely had prior experience using the Equation Editor (structure-based model). On the other hand, few, if any participants were familiar with the interface of MC^2 (free-form model) as it is a program designed for more niche, academic purposes and is therefore not widely used. This prior experience could have had some effect on the observed outcomes of the experiment.

Overall, these findings are important for developers of software applications designed to digitally typeset mathematical content to consider, and show evidence that the best model for use in WYSIWYG editors may be a hybrid which incorporates elements of both tested.

References

1. Anthony, L., Yang, J., Koedinger, K.R.: Evaluation of multimodal input for entering mathematical equations on the computer. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1184–1187 (2005)
2. Gozli, D.G., Pollanen, M., Reynolds, M.: The characteristics of writing environments for mathematics: behavioral consequences and implications for software design and usability. In: Carette, J., Dixon, L., Coen, C.S., Watt, S.M. (eds.) CICM 2009. LNCS (LNAI), vol. 5625, pp. 310–324. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02614-0_26

3. Irfan, M., Kusumaningrum, B., Yulia, Y., Widodo, S.A.: Challenges during the pandemic: use of e-learning in mathematics learning in higher education. Infinity J. **9**(2), 147–158 (2020)
4. Kang, S., Pollanen, M., Damouras, S., Cater, B.: Mathematics classroom collaborator (MC2): technology for democratizing the classroom. In: Davenport, J.H., Kauers, M., Labahn, G., Urban, J. (eds.) ICMS 2018. LNCS, vol. 10931, pp. 280–288. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-96418-8_33
5. Khirwadkar, A., Khan, S.I., Mgombelo, J., Obradovic-Ratkovic, S., Forbes, W.A.: Reimagining mathematics education during the covid-19 pandemic. Brock Educ.: J. Educ. Res. Pract. **29**(2), 42–46 (2020)
6. Kim, S., Pollanen, M., Reynolds, M.G., Burr, W.S.: Identification of errors in mathematical symbolism and notation: implications for software design. In: Davenport, J.H., Kauers, M., Labahn, G., Urban, J. (eds.) ICMS 2018. LNCS, vol. 10931, pp. 297–304. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-96418-8_35
7. Kim, S., Pollanen, M., Reynolds, M.G., Burr, W.S.: Problem solving as a path to comprehension. Math. Comput. Sci. **14**(3), 607–621 (2020). https://doi.org/10.1007/s11786-020-00457-1
8. Knauff, M., Nejasmic, J.: An efficiency comparison of document preparation systems used in academic research and development. PlOS One **9**(12), 1–12 (2014). Article ID e115069
9. Kohlhase, A., Fürsich, M.: Understanding mathematical expressions: an eye-tracking study. In: FM4M/MathUI/ThEdu/DP/WIP@ CIKM, pp. 42–50 (2016)
10. Loch, B., Lowe, T.W, Mestel, B.D.: Master's students' perceptions of Microsoft Word for mathematical typesetting. Teach. Math. Appl.: Int. J. IMA **34**(2), 91–101 (2015)
11. Marandi, R.Z., Madeleine, P., Omland, Ø., Vuillerme, N., Samani, A.: Reliability of oculometrics during a mentally demanding task in young and old adults. IEEE Access **6**, 17500–17517 (2018)
12. Quinby, F., Kim, S., Kang, S., Pollanen, M., Reynolds, M.G., Burr, W.S.: Markov transition matrix analysis of mathematical expression input models. In: Bigatti, A.M., Carette, J., Davenport, J.H., Joswig, M., de Wolff, T. (eds.) ICMS 2020. LNCS, vol. 12097, pp. 451–461. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52200-1_45
13. Quinby, F., Pollanen, M., Reynolds, M.G., Burr, W.S.: Effects of digitally typesetting mathematics on working memory. In: Harris, D., Li, W.-C. (eds.) HCII 2020. LNCS (LNAI), vol. 12186, pp. 69–80. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49044-7_7
14. Quinlan, J., Tennenhouse, C.: Perceived utility of typesetting homework in post-calculus mathematics courses. PRIMUS **26**(1), 53–66 (2016)
15. R Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013)
16. Raghubar, K.P., Barnes, M.A., Hecht, S.A.: Working memory and mathematics: a review of developmental, individual difference, and cognitive approaches. Learn. Individ. Differ. **20**(2), 110–122 (2010)
17. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. **124**(3), 372 (1998)
18. Strohmaier, A.R., MacKay, K.J., Obersteiner, A., Reiss, K.M.: Eye-tracking methodology in mathematics education research: a systematic literature review. Educ. Stud. Math. **104**(2), 147–200 (2020). https://doi.org/10.1007/s10649-020-09948-1
19. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. Springer, New York (2016)

Evaluation of Relationship Quality Within Dyads Through the Performance in Dual-Player Cooperative Tasks

Haijiang Yan[1,2], Kai Yu[1,2], and Guozhen Zhao[1,2](✉)

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
{yanhj,yuk,zhaogz}@psych.ac.cn

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Human beings are influenced widely by relationship among individuals. However, there is still a lack of a systematic, objective and direct way to assess the quality of relationships within dyads. The current study proposed a task-specific method to evaluate relationship quality within friends and lovers. Thirteen-eight paired friends and 38 paired lovers were recruited. All of them performed the Pattern Matching and Go/No-go task, and their task performance were recorded. The friendship and love quality were measured respectively. Based on the results, there is significant influence relationship between the average step of initial level in Pattern Matching task and quality of friendship within friends. Furthermore, both average time of advanced level in Pattern Matching task and total RT difference in Go/No-go task were related to the relationship quality between lovers. The study demonstrates the validity of task-specific evaluation interface for different relationship types, and provides a method improving the understanding of human social relationships by machines, which may be effective for establishment of affective interaction between humans and computers.

Keywords: Relationship quality evaluation · Friendship · Love · Cooperative task

1 Introduction

Human beings are defined as social animals, and the need to belong is proposed to be a fundamental human motivation [1]. Both from the perspective of evolution and human society, social relationships play an extremely important role in human existence [2]. Exactly, relation bonds, to some extent, determine our thoughts, emotions, behaviors and so on, affecting every aspect of life, such as psychosocial wellbeing [3], work performance [4] and even physical health [5, 6]. Recent study from Pieh et al. (2020) demonstrates that relationship quality is related to mental health during Coronavirus Disease (COVID-19) lockdown, indicating that positive social bonds may undertake a protective role in mental health during catastrophes [7]. At the group level, social network

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 223–233, 2021.
https://doi.org/10.1007/978-3-030-77932-0_19

ties could predict more organizational citizenship behavior [8]. Wang et al. (2019) find that relationship has an indirect effect on megaproject success [9]. Moreover, adaptive social networks promote the wisdom of crowds, which means that the quality of social relationship is positively linked to collective intelligence [10]. In a word, where there are people, there are relationships at work.

While effecting our thoughts, emotions and behaviors a lot, social relationships in human have been overlooked in researches which aim at improving naturalness and effectiveness of human-computer interface. As inherent characteristic and key factor of human mind, social relationships, in essence, means an exchange or infection of intentions, affections, thoughts and emotions among people [1, 2]. In the interactions between humans and computers, as the humans' minds spreading and exchanging through the interface, relationship may occur naturally and be evaluated easily. Hence, the understanding of social relationships by machines or computers would improve affective interaction, which is intended to add affective factors in human-computer interaction.

1.1 Relationship Quality Evaluation

Social interactions exist in our daily lives, the quality of which affects both individuals and groups. Therefore, it is significantly meaningful to perform the relationship quality evaluation. As the basic unit of social relationship networks, relationship within dyads serves as the core concept and index when researchers try to evaluate relationship quality. Currently, the most common methods to perform the evaluation are self-reports, interviews and observations [11]. Self-reports provide multi-dimensional structural information about quality of relationship. However, the subjective relationship quality scale, such as Friendship Qualities Scale (FQS) [12], Network of Relationships Inventory (NRI) [13], and Sternberg's Love Scale [14], suffer from their inherent defects and shortcomings (i.e., subjectivity and social desirability). Generally, the latter two methods are used as the supplement to subjective scale, providing additional unstructured and objective behavioral information [11]. Thus, there is still a lack of a systematic, objective and direct way to access the quality of relationships within dyads.

1.2 Different Relationship Types, Different Context

Relationship quality is complex in its measurement not only because the multiple dimensions behind the concept, but also because that various social bonds develop and form during different social interactions, and there are great differences among the different relationship types, such as the social contexts and the goals of the relationships. Therefore, it should be considered in the relationship measurement that both the definition and the context of a specific relationship are important. Reblin et al. (2020) attempted to figure out the predictive effect between relationship quality and marital functioning in couples under multiple contexts (i.e., positive, neutral and so on), the results show that the inclusion of multiple dimensions of relationship quality across different contexts improved prediction of marital functioning for both women and men, and multicontextual relationship quality assessments are highly recommended [15]. That is to say, different contexts, as the inducing factors, could reflect different aspects related to different relationship qualities.

This study focuses on two common types of relationship within dyads, friendship and love. Friendship, as a dynamic and mutually beneficial relationship between two individuals, plays an important role in problem-solving in human evolution and social development [2, 16]. Friendship is defined by Berndt et al. (2002) as a type of interpersonal relationship with high intimacy, prosocial behavior, self-esteem, support, loyalty and low conflict from a multidimensional perspective [17]. Based on Sternberg's Triangular Theory of Love, love consists of three components: intimacy, passion and commitment [14]. Recent fMRI study reveals that love is primarily a motivation process, called "attraction system" which is based on the dopaminergic reward system [18]. In general, friendship and love are two completely different relationship, informally friendship is a kind of cooperation while love is a kind of attraction that is more of an emotional and reward component.

1.3 Dual-Player Cooperative Tasks

As a common experimental paradigm in researches of emotional domain, multi-player cooperative tasks are expected to provide a new interface to evaluate the quality of relationship. Unlike the traditional methods used in relationship evaluation, cooperative tasks provide a new perspective from which the qualities of relationships are evaluated through the performance in specific contexts and goals related to the relationship instead of the structural information about multiple dimensional experiences of the relationship. In essence, the two perspectives have great homogeneity. Anderson et al. (2004) proposed that there is a close relationship between the quality of interpersonal relationship and emotional cohesion, and this emotional cohesion effect plays an important role in the synchronization of team cooperation behaviors, cooperation consciousness and behavior rhythm, which determine the team performance in the cooperative tasks [19]. Therefore, the performance of cooperative tasks may reflect the quality of interpersonal relationship to a certain extent. However, no research till now discuss the value and utility of cooperative task performance in the evaluation of relationship quality.

Considering that the contexts of evaluation are sensitive to its objects, different cooperative tasks might be sensitive and applicable for different relationship patterns, and the key points may be in the cooperative patterns and goals of the tasks. In this study, we redesigned two distinct dual-player cooperative task, pattern matching task and go/no-go task, to explore the role of task performance in relationship quality evaluation. In brief, pattern matching task asks two participants to perform a decision-making in turn to achieve a goal eventually, and go/no-go task mainly reflect the synchronization of cue-reactivity within dyads. Thus, pattern matching task is related to collective decision-making and problem solving, and the latter one is related to the consistency of behavior rhythm. Based on the above background, we hypothesis that pattern matching task is able to reflect quality of friendship, and go/no-go task may be more sensitive in evaluation of love experiences.

The purposes of this study are to find out whether cooperative task performance related indicators could be effective on the evaluation of relationship quality within dyads, and whether there is a task-specific evaluation effect by comparing the evaluated effects of two dual-player cooperative task paradigms, the graph matching task and the go/no-go task, within two distinct relationship patterns (friendship and love).

2 Methods

2.1 Participant

Seventy-six pairs of participants (n = 152) with different relationships (Friends & Lovers) were recruited. Friends: Thirty-eight pairs of friends with same gender participated in the study. Lovers: Thirty-eight pairs of lovers were recruited. The inclusion criteria required that the relationships between the participants were maintained for at least 3 months and they would meet each other at least twice a week. All participants reported that they were in good physical and mental state, right-handed, and had normal or corrected vision, no color blindness or color weakness. Informed consent was signed after a detailed description of the study. The study was approved by the Ethics Committee of the Institute of Psychology, Chinese Academy of Sciences.

2.2 Materials and Tasks

Friendship Quality

The friendship quality was measured by Chinese Version of Friendship Quality Inventory (FQI-CV). The scale has 15 items, including 4 dimensions of Support, Conflict, Partnership, Intimacy. The scale shows good internal consistency and discriminative validity among Chinese college students [20]. The mean score of the paired participants was used to represent the friendship quality within dyads.

Love Quality

The relationship quality between lovers was measured by Experiences in Close Relationships Scale (ECRS) [21]. The scale consists of 36 items, including two dimensions of avoidance and anxiety, with 18 items in each dimension. The scale has good reliability and validity.

Dual-Player Pattern Matching Task

The process of the task referred to previous studies, which was used to evaluate performance of teamwork [22, 23]. The task includes two difficulty levels: Initial and Advanced. The initial level of task consisted of 6 trials. In each trial, a target pattern consisting of color stones was presented on a 5 × 5 lattice. The target pattern was composed of two different colors. The participants were required to regenerate the shape and color of target pattern on another 5 × 5 lattice without any pattern. In the advanced stage with 2 trials, the lattice was set to 7 × 7, and the target pattern increased accordingly. Other parameters remained the same as the initial level. The task consisted of 8 trials totally in which there were 8 different target patterns. Each pair was asked to complete all same 8 trials.

As shown in Fig. 1, the paired participants changed color or location of stones to match the target pattern by manipulating the movement of the cursor in turn, and only one operation was allowed per step. Besides, one of the participants can control move-up and move-down of the cursor, and color-shifting of stones, another can control move-left and move-right of the cursor, and color-shifting.

To evaluate the performance, mean completion time and steps at each level were calculated. Mean completion time were defined by average time of correct trials at each level. Mean completion steps were defined as the average number of moves in correct trials at each level.

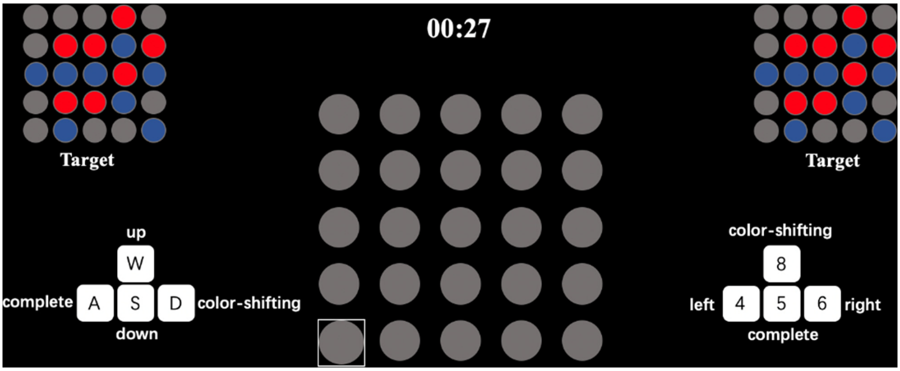


Fig. 1. Stimuli of pattern matching task in initial level

Dual-Player Go/No-Go Task

The task was adapted from the dual-player collaborative response paradigm used in a Near-Infrared Spectroscopy (NIRS) study by Cui et al. (2012) [24]. The original paradigm required paired participants to respond to specific target stimuli ("go stimulus"), then participants would receive a feedback. When the difference between the response time of the paired subjects was less than one-eighth of the total of their response time, it was recorded as the correct trials, and one point was given to the participants. Otherwise, the current trials failed, and no point was given.

Considering that it is relatively simple to respond to a single go stimulus, in this study we added no-go stimulus into the original paradigm. The adapted task included two blocks, each with 30 trials, 60 trials in total, including 42 go trials and 18 no-go trials. In each trial, firstly a white circle was present on the screen as a clue of beginning, and then go stimulus (green circle) or no-go stimulus (red circle) occurred. The SOA between the clue and stimulus ranged from 600 ms to 1500 ms. Once the go stimulus was presented, two subjects need to press the button at the same time, whether the trial was correct was up to the difference between the response time of the paired subjects as introduced before. The total score will be displayed in the upper right corner of the screen, and the symbols " + " and "−" would be displayed on the left and right sides respectively, the display of " +" means that in the current trial the participant is faster than his/her partner, and "−" means the opposite. Once there was a no-go stimulus, two subjects were required not to press the button, and if they did so, it was recorded as a correct trial. Otherwise, the current trials failed. The task interface is shown in Fig. 2. We collected the indicators of task performance including total score and total difference of response time within lovers.

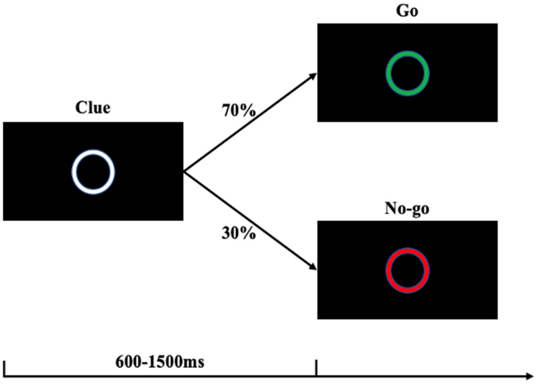


Fig. 2. The task scene and materials of Go/No-go task

2.3 Procedure of Experiment

Upon arrival, each pair of participants confirmed the registration information and completed the informed consent form, then took a five-minute break and started to fill out the scale independently (Friends filled in the FQI-CV, Lovers completed the ECRS). After filling in the scale, they began to complete the two tasks, the sequence of which was counter-balanced in all participants in order to get rid of sequence effect. As shown in Fig. 3, the experiment programs were run on a 27-in. Dell computer (resolution 1920*1080, refresh rate 60 Hz), with an external keyboard to record the response. The study required each pair participants to sit 60 cm away from the monitor and 20 cm apart.

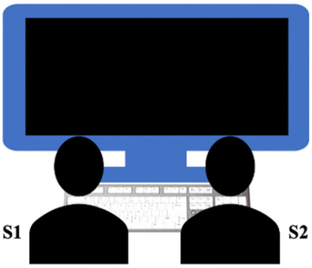


Fig. 3. The task scene and materials of the experiment

The tasks included practice trials and formal trials. In order to ensure the quality of the study, we started the formal trials after ensuring that the participants were familiar with the task requirements. They were encouraged to finish the task seriously. The study lasted for about half an hour and the participants would receive ¥50 as reward.

2.4 Data Analysis

The data analysis was performed by SPSS25. In both relationship type (Friends & Lovers), the task performance indicators were used as the independent variable, and the scores of each dimension of relationship quality were used as the dependent variable to establish multi-variable regression model respectively.

3 Results

3.1 Descriptive Statistics

Seventy-six paired participants with different relationship (Friends & Lovers) were recruited. Friends: 38 pairs of friends participated in the study, among which 15 pairs were males and 23 pairs were females. Mean age of them was 21.91 ± 1.95 years. Lovers: 38 pairs of lovers were recruited, and mean age of them was 22.08 ± 2.23 years. All detailed information was shown in Table 1.

Table 1. Descriptive statistics (Mean ± SD)

Variables	Lovers (Npairs = 38)	Friends (Npairs = 38)
Demographic variables		
Age	22.08 ± 2.23	21.91 ± 1.95
Gender (M/F)	–	15/23
Task performance		
Average Time ^{a}/(s)	77.00 ± 14.46	83.52 ± 22.74
Average Time ^{b}/(s)	127.87 ± 30.11	129.04 ± 37.22
Average steps ^{a}	33.51 ± 3.68	33.61 ± 2.15
Average steps ^{b}	63.08 ± 5.83	62.17 ± 6.23
Total score	243.11 ± 9.51	241.87 ± 10.53
Total RT difference	75.92 ± 24.63	72.80 ± 19.93
Love Quality (ECRS)		
Avoidance	2.69 ± 0.49	–
Anxiety	3.48 ± 0.79	–
Total score	3.08 ± 0.52	–
Friendship quality (FQI-CV)		
Support	–	3.79 ± 0.36
Conflict	–	1.98 ± 0.68
Partnership	–	3.83 ± 0.55
Intimacy	–	3.54 ± 0.67
Total score	–	3.79 ± 0.35

Note: ^{a} Initial trials; ^{b} Advanced trials; *P < 0.05, ** P < 0.01; RT, reaction time.

3.2 Multi-variable Regression Model

To establish the predictive relationship between task performances and relationship quality, the stepwise regression method was used to clarify whether indicators of task performance in the two tasks could reflect the quality of either friendship or love.

Friendship

As shown in Table 2, average steps of initial level in Pattern Matching task could negatively predict quality of friendship (F = 5.148, R2 = 0.125, P = 0.029), which was the only one indicator reaching the significant level among all indicators of performance.

Love

The total score of ECR did not show any significant effect in regression model with the task performance. However, as shown in Table 3, average time of advanced level in Pattern Matching task and total RT difference in Go/No-go task could together predict anxiety in love (F = 5.677, $R^2 = 0.245$, P = 0.007), while other indicators of performance were excluded from the model.

Table 2. Linear regression model (FQI-CV)

Variables	Beta	T	P
Demographic variables			
Age	.102	.563	.581
Gender (M/F)	−0.121	−.645	.523
Task performance			
Average Step [a]	−.354	−2.269	.029*
Average Time [a]/(s)	−.124	−.653	.518
Average Time [b]/(s)	−.096	−.543	.591
Average steps [b]	−.013	−.073	.942
Total score	−.083	−.526	.602
Total RT difference	−.048	−.295	.770

Note: [a] Initial trials; [b] Advanced trials; *P < 0.05, ** P < 0.01; RT, reaction time.

Table 3. Linear regression model (ECR-Anxiety)

Variables	Beta	T	P
Demographic variables			
Age	−.088	−.397	.706
Task performance			
Total RT difference	.347	2.361	.024*
Average Time ^{b}/(s)	−.330	−2.245	.031*
Average Time ^{a}/(s)	.244	1.086	.285
Average Step ^{a}	.043	.291	.773
Average steps ^{b}	−.002	−.009	.992
Total score	−.071	−.363	.719

Note: ^{a} Initial trials; ^{b} Advanced trials; *P < 0.05, ** P < 0.01; RT, reaction time.

4 Discussion

The current work established a task-specific method to evaluate the relationship quality of dyads with specific type of relationship, namely friends and lovers. Based on the results, we found that performance in Pattern Matching task (average step in initial level) was relatively sensitive to friendship quality between paired friends, however, both performances in Pattern Matching task (average time in advanced level) and Go/No-go task (total RT difference) were related to the relationship quality between lovers.

Less average step in initial level of Pattern Matching task predicted better quality of friendship. Brendgen et al. (2001) pointed out that friendship is far beyond a kind of emotional experience, it is related to individual's social cognition and behavior interaction pattern with others, which plays an important role in cooperative problem solving [25]. In fact, as a core factor of collective intelligence [10], relationship quality reflects how effective and accurate the interaction is between friends in cooperative problem solving. Average step means how many operations in completion of the task, indicating the quality of strategy used in the task. As a result, friends with better quality of friendship would make more clever decision-making and perform better in the Pattern Matching task.

Higher RT difference in Go/No-go task and less time spend in advanced level of Pattern Matching task was associated with poor quality of relationship quality between lovers. Recent fMRI study reveals that love is a kind of motivation process, called "attraction system" which is based on the dopaminergic reward system [18]. Obviously, unlike friendship, love is a relationship more likely to share, experience and infatuate with each other. Therefore, love is more prone to be a kind of relationship with logic of understanding instead of rationality. Compared with Pattern Matching task, the point of Go/No-go task is how consistent of cognition and action within lovers, which is related to the consistency of RT in Go/No-go task. However, the result showed that less average time spend in advanced level of Pattern Matching task was also associated with poor quality of relationship quality between lovers. Aloni et al. (2004) pointed out that love is blind, revealing that love is associated with biased and inaccurate perception within lovers [26], which means better quality of love may lead to poor decision-making in problem solving. As a result, average time in Pattern Matching task showed positive relationship with love quality.

The study demonstrates the task-specific evaluation of different relationship types, and shows the prospect of cooperative task in evaluation of relationship within dyads in the future. In the future work, more relationship types and cooperative tasks can be included and the task-specific evaluation method would be further validation. Furthermore, the study provides a method improving the understanding of human social relationships by machines, which may be effective for establishment of affective interaction between humans and computers. Moreover, as the relationship between humans and computers being more complex, the probability of evaluation and even formation of relationship between humans and computers should be considered and discussed.

Acknowledgments. This work was supported by the National Key Research and Development Plan (2018YFC0831001) and Innovation Capacity Support Plan of Shaanxi Province (2020KJXX-024).

References

1. Baumeister, R.F., Leary, M.R.: The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol. Bull. **117**(3), 497–529 (1995)
2. Lewis, D.M.G., Conroy-Beam, D., Al-Shawaf, L., Raja, A., Dekay, T., Buss, D.M.: Friends with benefits: the evolved psychology of same- and opposite-sex friendship. Evol. Psychol. **9**(4), 543–563 (2011)
3. McPheters, J.K., Sandberg, J.G.: The relationship among couple relationship quality, physical functioning, and depression in multiple sclerosis patients and partners. Fam. Syst. Health **28**(1), 48–68 (2010)
4. Alejandro, T.B., Souza, D.V., Boles, J.S., Ribeiro, Á.H.P., Monteiro, P.R.R.: The outcome of company and account manager relationship quality on loyalty, relationship value and performance. Ind. Mark. Manage. **40**(1), 36–43 (2011)
5. Reblin, M., Donaldson, G., Ellington, L., Mooney, K., Caserta, M., Lund, D.: Spouse cancer caregivers' burden and distress at entry to home hospice: the role of relationship quality. J Soc Pers Relat **33**(5), 666–686 (2016)
6. Skoyen, J.A., Rentscher, K.E., Butler, E.A.: Relationship quality and couples' unhealthy behaviors predict body mass index in women. J. Soc. Pers. Relat. **35**(2), 224–245 (2018)
7. Pieh, C., O' Rourke, T., Budimir, S., Probst, T.: Relationship quality and mental health during COVID-19 lockdown. PLoS One **15**(9), e0238906 (2020)
8. Scott, K.L., Zagenczyk, T.J., Li, S., Gardner, W.L., Cogliser, C., Laverie, D.: Social network ties and organizational citizenship behavior: evidence of a curvilinear relationship. Eur. J. Work Organ. Psychol. **27**(6), 1–12 (2018).
9. Wang, D., Lu, Y., Fang, S.: Connection between relationship quality and megaproject success: moderating role of contractual functions. Adv. Civil Eng. **2019**, 1–13 (2019)
10. Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P.M., Moussaid, M., Pentland, A.: Adaptive social networks promote the wisdom of crowds. Proc. Natl. Acad. Sci. U S A **117**(21), 11379–11386 (2020)

11. Bagwell, C.L., Bender, S.E., Andreassi, C.L., Kinoshita, T.L., Montarello, S.A., Muller, J.G.: Friendship quality and perceived relationship changes predict psychosocial adjustment in early adulthood. J. Soc. Pers. Relat. **22**(2), 235–254 (2016)
12. Bukowski, W.M., Hoza, B., Boivin, M.: Measuring friendship quality during pre- and early adolescence: the development and psychometric properties of the friendship qualities scale. J. Soc. Pers. Relat. **11**(3), 471–484 (1994)
13. Furman, W.: The measurement of friendship perceptions: conceptual and methodological issues. Cambridge University Press, New York, NY, pp. 41–65 (1998)
14. Sternberg, R.J.: A triangular theory of love. Psychol. Rev. **93**(2), 119–135 (1986)
15. Reblin, M., Vaughn, A.A., Birmingham, W.C., Smith, T.W., Uchino, B.N., Spahr, C.M.: Complex assessment of relationship quality within dyads. J. Commun. Psychol. **48**(7), 2221–2237 (2020)
16. Burk, D.I.: Understanding friendship and social interaction. Childhood Educ. **72**(5), 282-285 (1996)
17. Berndt, T.J.: Friendship quality and social development. Curr. Dir. Psychol. ENCE **11**(1), 7–10 (2002)
18. Fisher, H., Aron, A., Brown, L.L.: Romantic love: an fMRI study of a neural mechanism for mate choice. J. Comp. Neurol. **493**(1), 58–62 (2005)
19. Anderson, C., Keltner, D.: The emotional convergence hypothesis: Implications for individuals, relationships, and cultures. The social life of emotions, pp. 144–163 (2004)
20. Fan, X., Fang, X.: Reliability and validity of the Chinese version of friendship quality inventory. Chin. J. Clin. Psychol. **12**, 133–146 (2004)
21. Li, T., Kato, K.: Measuring adult attachment: Chinese adaptation of the ECR scale. Acta Psychol. Sin. **38**, 399–406 (2006)
22. Hwang, S.T., Whang, M.: Research on emotional convergence technology based on noncontact sensing and social interaction. J. Korea Converg. Soc. **8**, 163–170 (2017)
23. Lee, J.N., Whang, M., Park, S., Hwang, S.T.: Engagement classification algorithm based on ECG (electrocardiogram) response in competition and cooperation games. J. Korea Game Soc. **17**, 17–26 (2017)
24. Cui, X., Bryant, D.M., Reiss, A.L.: Nirs-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. Neuroimage **59**, 2430–2437 (2012)
25. Brendgen, M., Markiewicz, D., Doyle, A.-B., Bukowski, W.M.: The relations between friendship quality, ranked-friendship preference, and adolescents' behavior with their friends. Merrill-Palmer Q. **47**(3), 395–415 (2001)
26. Aloni, M.B., Frank, J.: Is love blind? The effect of experience and infatuation on the perception of love. J. Nonverbal Behav. **28**(4), 287–295 (2004)

Cognitive Activity Recognition Based on Self-supervised Learning from EEG Signals

Yifeng Yang, Yingjie Zhao, Yanyu Lu(✉), and Shan Fu

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, People's Republic of China
luyanyu@sjtu.edu.cn

Abstract. Identifying the cognitive activities of the human brain is a daunting task due to the fact that those cognitive activities were not observable directly by any known sensing technology. Electroencephalogram (EEG) provided a very useful information that associated with the cognitive activities very closely and was used widely for understanding and recognizing human cognitive activity. In this paper, an experiment was designed to abstractively present the process of "target search" and "action execution" which were among the most common types of cognitive activities during human-computer interaction. EEG signals of the applicants of the experiment were collected and used for subsequent cognitive activity analysis which included a novel temporal-based self-supervised learning approach using BERT to pre-train the data for feature embedding. These encoded points generated by the proposed algorithm could be assigned to the corresponding categories by k-means to capture hidden information about the dominate type of the cognitive activities in any period of 20 ms. The results showed that this approach can distinguish the cognitive activities between the target searching and action executing. And the results suggest that in such a task, subjects' cognitive activities are relatively pure in the initial search moments, but later in the task, multiple activities may be mixed.

Keywords: Cognitive activity recognition · Self-supervised learning · Time series analysis · BERT

1 Introduction

In recent years, with the development of technology, more and more machines and equipment have appeared in people's lives, aiding or replacing people in many fields to complete essential tasks. The human-computer interface provides an important way for operators to achieve human-computer interaction to ensure the machine's accurate operation. Taking aviation transportation as an example, the airplane has become a common choice for people to travel, and compared with other forms of transportation is also the safest way. Airlines spend a great deal of effort on hardware equipment to fly safely, such as using more sophisticated sensor devices or developing safer aircraft software systems. In the aircraft system, the flight deck also plays a vital role in ensuring

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 234–247, 2021.
https://doi.org/10.1007/978-3-030-77932-0_20

aviation safety because it is the only interface between the pilot and the aircraft. Various sensor devices and controls on the plane are connected. A flight deck is a complex human-computer interaction system [1]. Human factors have gradually become a major factor affecting flight safety. Whether a flight deck can be designed to match the pilot's cognitive ability can directly affect the aircraft's flight safety. Excessive pilot workload, inattentiveness, and other conditions can pose a risk to flight safety. Therefore, detecting human physiological information to obtain the pilot's cognitive activity can be ensure flight safety from another perspective. Compared with other physiological signals, EEG, which can more directly reflect the human brain's cognitive activity during the task, is considered as one of the most effective detection means.

In the human-computer interaction of complex systems, humans need to identify goals, make decisions, and perform appropriate operations on complex interfaces based on the situation. The clarity of the interface design affects human cognitive activity. Existing behavioral measures can only provide indirect measures of human cognitive activities, and it is difficult to directly observe human cognitive activities such as comprehension and decision making. EEG provides a good approach. The non-invasive EEG acquisition method has many advantages, as the information from the brain is recorded through a wearable device attached to the subject's scalp, which reduces the risk of surgery and is more comfortable and cheaper to use than the invasive method of acquisition. EEG has a higher temporal resolution than Magnetoencephalography (MEG) and can be more easily deployed in a broader range of sites. Through non-invasive EEG detection techniques, many studies have investigated human fatigue while driving and the presence of inattention and inefficient operation [2, 3]. EEG signals can also be used to probe people's intentions. Some researchers, who built a driver braking intention detection system, detect drivers trying to apply the brakes when danger appears by capturing EEG signals while driving [4, 5].

Event-related potential (ERP) is a way to study human cognitive activity, which is an EEG signal evoked by infrequently occurring sensory stimuli, often appearing at parietal cortex locations in the brain. P300 is a type of ERP, and its peak occurs approximately 300ms after the event of interest [6]. The classical oddball paradigm is the commonly used ERP experimental paradigm. However, the above studies are based on EEG signals evoked by external stimuli. Getting data in this way makes it easy to locate the target data segment and tag it with the appropriate label. But in the actual human-computer interaction process, the change of human cognitive activity is endogenous or top-down. The identification of this endogenous cognitive activity is vital for understanding the HCI process. So, in this paper, we designed a target searching-execution task to study the subjects' cognitive activity changes through using the EEG signals collected during the task.

In previous studies about EEGs processing, the extraction of effective features from EEG signals to identify cognitive activity under different stimuli has been one of the main problems. The methods used in past studies generally include the time series modeling of EEG signals, using the model's coefficients as features, and the application of Fourier transform or wavelet transform to extract the features from EEG signals in the frequency domain and the time-frequency domain [7, 8]. With the development of deep learning, many researchers have proposed to use CNN to extract features from original EEG data

[9, 10]. The LSTM network is also well suited for processing temporal information [11, 12]. In this paper, although the changes in subjects' cognitive activity were endogenous and the data did not correspond to stimulus labels, the EEG signal generated at a specific moment during the task is related to the preceding and following moments. Therefore, this paper proposes to construct a time-series based self-supervised learning task to pre-train the pre-processed data as feature extraction using the BERT model based on the attention mechanism in the absence of labels.

Here is the rest of the paper's general organization: Sect. 2 provides an introduction to subjects and equipment and details how the experiments were designed and implemented. Section 3 focuses on how to pre-process the acquired raw data. Section 4 details the method proposed in this paper for extracting useful features to reflect different cognitive activities based on a self-supervised learning task in the absence of data labels. Section 5 focuses on the results of the experiment and the analysis of these results. Section 6 summarizes the conclusion of this paper. In such a task, each trial subjects repeated such an orderly process of search to click.

2 Experiment

2.1 Subjects

Three healthy participants were invited to participate in the experiment, all of whom were around 24 years of age and had a normal or corrected vision. Before the experiment, all subjects had been informed of the entire process and were competent enough to complete the trials following the experiment's basic requirements. The participants voluntarily signed an informed consent form.

2.2 Equipment

EEG Signal Acquisition Equipment

A 30-channel Cognioniocs system was used to collect the EEG signals during the experiment, and the electrode locations conformed to international standards 10–20 (see in Fig. 1 and Fig. 2) [13]. The device's sampling frequency was set at 500 Hz, and the impedance value of the dry electrodes was kept at about 600 kilohms. The device is connected to computer A, which displays brain waves in real-time.

Touch Screen

In our experiment, subjects were required to tap the screen directly with their fingers, so we used a touch screen. The screen displays the program designed for the experiment. When the subject clicks on the screen, the computer B connected to this touch screen sends a mark to computer A, recording the EEG signals, which would be used to segment the trials.

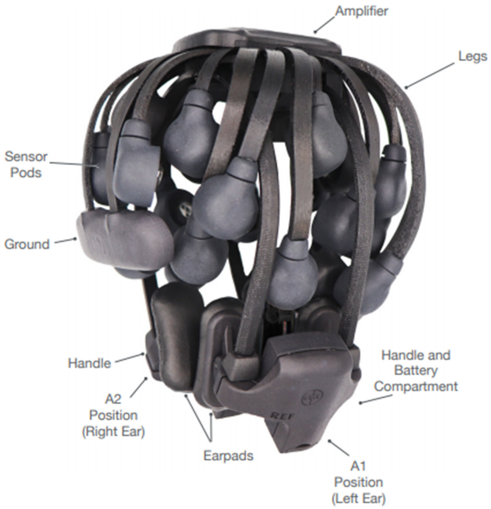


Fig. 1. Quick-30 30-channel dry EEG headset

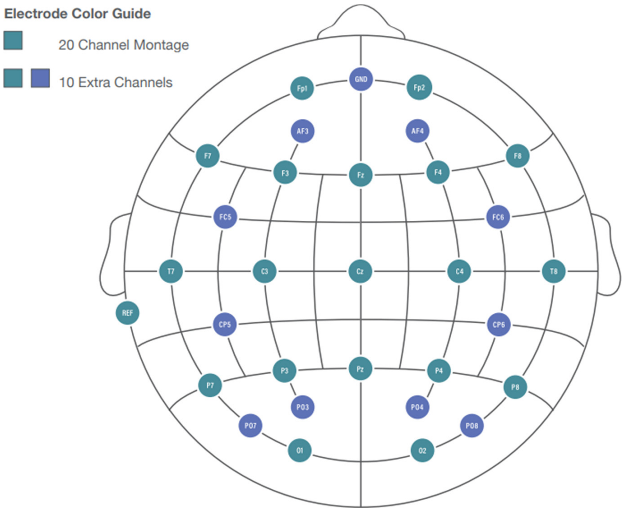


Fig. 2. 30 dry electrodes sensor location

2.3 Experiment Procedure

The procedure was as follows: a subject sat directly in front of a touch screen. He adjusted his distance from the screen to easily click anywhere on the screen without shaking his body, which reduced the interference of unnecessary movements with EEG signals. The tilt angle of the screen was also adjusted to a suitable viewing position so that the head's movement could be minimized during the test, which also served to reduce the noise of the data. The experimenter then put the experimental device on the subject.

Before starting the formal experiment, subjects sat in silence with their eyes closed for about 10 min, in which EEG signals are also recorded as a baseline. After that, the formal experiment began (Fig. 3).

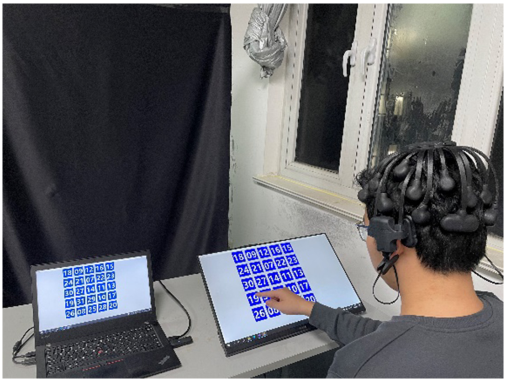


Fig. 3. Experimental environment

This experiment was proposed to mimic the "target searching and action executing" process during the interaction between the human and machine, which can be described as follows: a complete experiment contained six blocks, each of varying difficulty, and in each block, a matrix box appeared on the screen filled with numbers from smallest to largest, but the positions of the numbers were randomized. The subjects were asked to find the target number and clicked it in order from small to large. For example, at the beginning of the block, a 3*3 matrix will appear in the middle of the monitor, and the numbers 0–9 will randomly fill the matrix. More difficult blocks include 16 numbers in the 4*4 matrix or 25 numbers in the 5*5 matrix. Subjects were given short breaks between each block, while subjects were required to be continuous during each block until task completion. Finally, the device will capture EEG signal data from the subject searching for digits to click digits (Fig. 4).

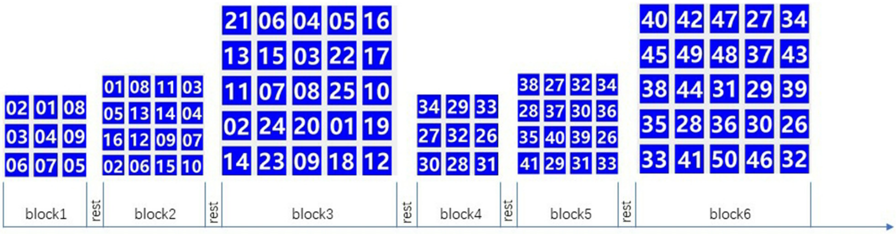


Fig. 4. Time sequence of the experiment

3 Data Preprocess Method

Unlike traditional EEG experiments, where a stimulus is given and the EEG data before and after the presence of the stimulus is recorded, this experiment does not have explicit external visual stimuli. The time point for segmenting the EEG signal is based on when the subject clicks the touch screen. The EEG data between the two moments contained the subject's EEG data for completing the search-click pattern once.

The acquired EEG data were segmented according to the moment of each screen click by the subject. The segmented data were next preprocessed using EEGLab in MATLAB. First, 29 of the electrodes were selected, and a whole-brain average reconstruction reference was used. Secondly, the data were filtered, first using a bandpass filter of 0.5–40 Hz to get the signal band of our interest, and then a bandstop filter of 45–55 Hz to remove the industrial frequency interference. Finally, the Artifact Subspace Reconstruction method of the clean_rawdata plug-in is used for automatic correction to eliminate the following kinds of data segments [14]:

1. Data segments that are too flat within 5 s;
2. Data segments with correlation coefficients less than 0.8 for spatially adjacent channels;
3. Data segments identified by using the ASR algorithm that have Data segments with artifacts.

4 Methodology

Time-series based self-supervised learning mainly uses auxiliary tasks to mine its supervised information from large-scale unsupervised data and trains the network with this constructed supervised information to learn representations that are valuable for downstream tasks. The supervised information in self-supervised learning is not manually labeled [15].

In each trial, subjects repeat such an ordered search-to-click process rather than click-to-search. The corresponding EEG signal is also transitioned from the EEG signal generated by the subject during the search activity to the EEG signal during the click activity. Therefore, the data obtained by normal sampling is defined as positive samples, and the data after manual segmentation to adjust the order is defined as negative samples.

Then, BERT is used to transforming the raw data from each sampling point combined with information from its context into a meaningful feature form. The results are fed into a neural network classifier to complete the supervised learning task using the positive and negative sample labels constructed by ourselves. In downstream task, the k-means method is used to cluster these features derived from the inspiration of the word bag model [16]. If the features extracted by the model can well determine whether the current input sample timing is correct, it can be assumed that the current model can well extract the sequence information implied in the sample data, i.e., it can identify which part of a complete sample belongs to the earlier part and which part belongs to the later part (Fig. 5).

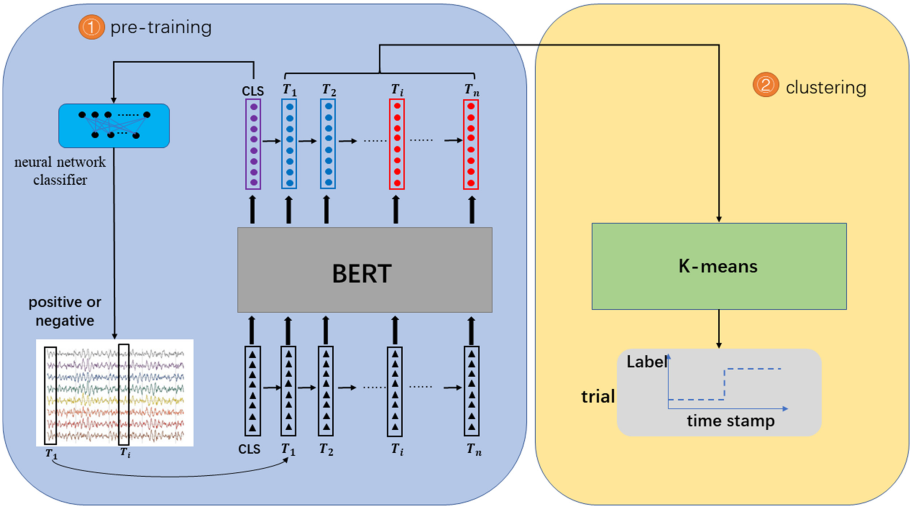


Fig. 5. The structure of the time-series based self-supervised learning task

4.1 Pre-training

Build Self-Supervised Learning Datasets

After counting the time spent by subjects per click, data segments with a single trial duration of less than 1s were removed, which were considered too short and unlikely to contain the complete search-to-click status. For trials longer than 1s, 500 sampling points, i.e., 1s of sampled data, were intercepted backward from the moment the click was completed for subsequent use (Fig. 6).

For each subject's data, a time window of different lengths was taken to intercept the tail of the data according to each subject's performance and then placed at the source data's head (Fig. 7).

Bidirectional Encoder Representations from Transformers

The model uses a self-attention mechanism to encode the raw data into meaningful features. For the current momentary data point, the model selectively extracts information from adjacent momentary data points instead of treating all momentary data points equally, allowing the model to capture the data's internal correlation better.

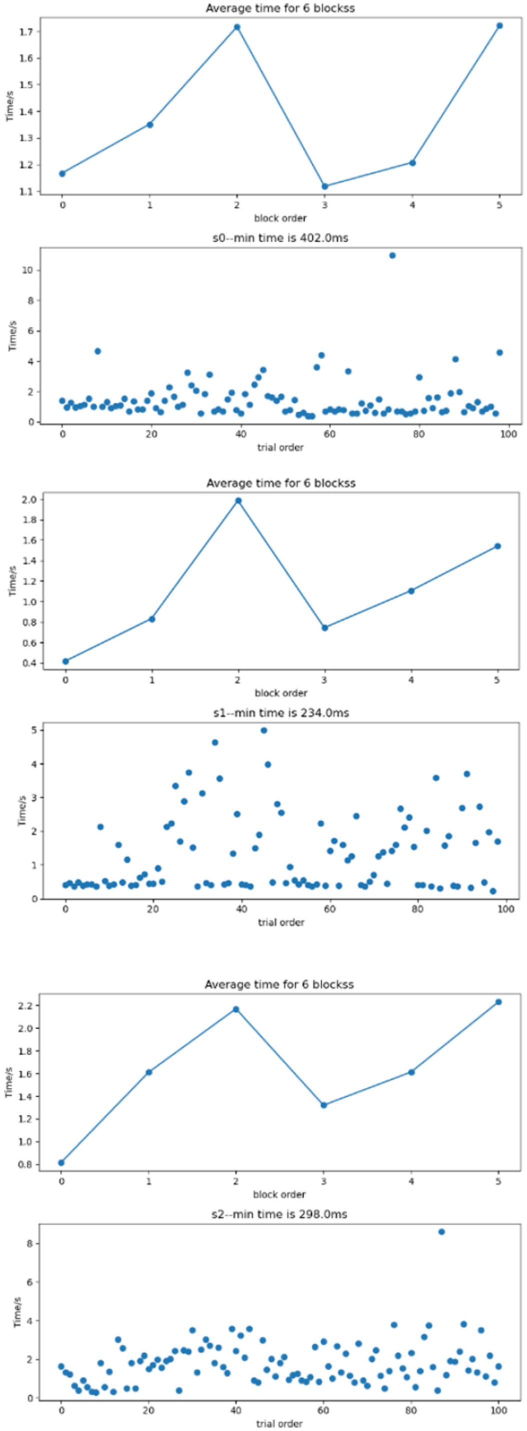


Fig. 6. Subjects s0-s2, (Upper panel) the average time spent for the 6 blocks. (Lower panel) the specific time spent by the subjects for each trial throughout the experiment.

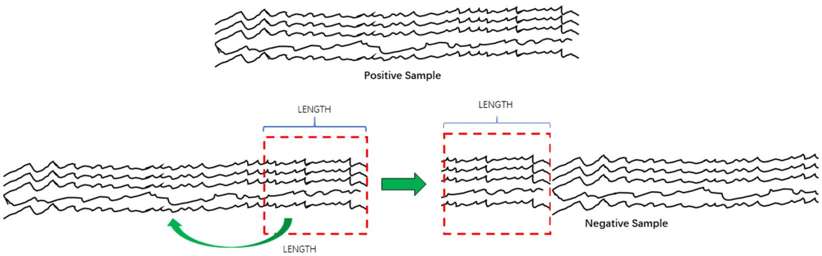


Fig. 7. Construct positive and negative samples by disrupting the temporal order of the pre-processed data

Network Architecture. BERT is a part of Transformer, which abandons the use of circular structure like classical network RNN, solves the problem of parallel computation, and the self-attention mechanism enables the model to take into account the information before and after the data. In the self-attention cell, the input at each moment is transformed into three vectors q, k, v using the three parameter matrices $W_Q,\ W_K,\ W_v$ to be optimized. Then the input is transformed into the following new form, d_k is input vector dimensions:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \tag{2}$$

In order for BERT to capture more relevant information on the time series, multiple self-attention mechanism modules are usually used.

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_n)W^o \tag{3}$$

$$head_i = Attention\left(QW_Q^i, KW_K^i, VW_V^i\right) \tag{4}$$

In this paper, instead of adding position coding, a sliding window-like effect is achieved by adding src_mask, so that points far apart can only be linked by the point between them. The model also incorporates residual networks and layer normalization to reduce the problems associated with the depth of the model. These network structures can be easily built by using the PyTorch framework (Fig. 8).

$$LN(x_i) = \alpha \times \frac{x_i - \mu_i}{\sqrt{\sigma_L^2 + \varepsilon}} + \beta \tag{5}$$

Loss Function

A special vector is added before each sample, and the output of its corresponding position (CLS) is used as the input of the neural network classifier to determine whether the current sample is a positive or negative sample. Since a binary classification task is actually constructed in the pre-training phase, this paper uses a cross-entropy loss function to optimize the model parameters.

$$l = -w\left[y \cdot logx + (1 - y) \cdot \log(1 - x)\right] \tag{6}$$

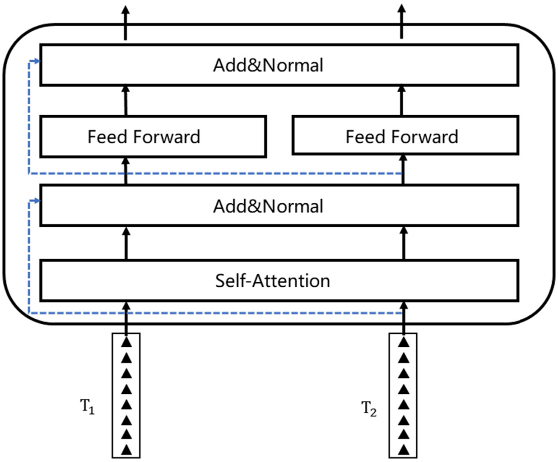


Fig. 8. Bidirectional encoder representations from transformers [17]

4.2 Clustering

The K-means algorithm is a distance-based clustering algorithm that uses distance as the evaluation index of similarity, and considers that the closer the distance between two objects, the greater their similarity. Algorithm flow:

4. K points in the sample are randomly selected as clustering centers;
5. Find the distances of the remaining samples to those K centers and assign those samples to the nearest centers to form the initial clusters;
6. Using the samples from the obtained clusters, the new centers are calculated;
7. Repeat steps 2 and 3 until the cluster center no longer changes.

However, the data to be clustered in this paper are slightly larger, so the K-means++ algorithm is used, i.e., a point is first randomly selected as the first centroid, and the rest of the selected centroids should be farthest from the currently selected centroid. Instead of using all the sample data points, a certain number of samples are drawn from the overall sample to complete both the initialization of the clustering centers and the later iterations to follow up the centers.

For each subject, the data segments generated in all the eligible trials are feature transformed by the BERT model, and the obtained feature vectors are used to generate the cluster centers of K-means (In this paper, two centers are set for the algorithm because we assume that there are two main types of subjects during the experiment.), Finally, based on the obtained clustering centers, the centers to which the feature vectors in a single trial belong can be determined sequentially. Thus, a curve that varies on the time axis can be obtained, representing the variation of the subjects on the two cognitive activities in this trial.

5 Result

The 29-dimensional feature vector of each sampling point is downscaled to 2 dimensions by t-SNE (t-stochastic neighbor embedding) for visualization. This dimensionality reduction method will make samples similar in high dimensional space closer together in low dimensional space, and points that are not similar in high dimensional space are separated in low dimensional space. The 1s data of each trail on the timeline is marked with different shades of blue and red, with dark blue points representing closer to the beginning of the data and dark red points representing closer to the end of the data.

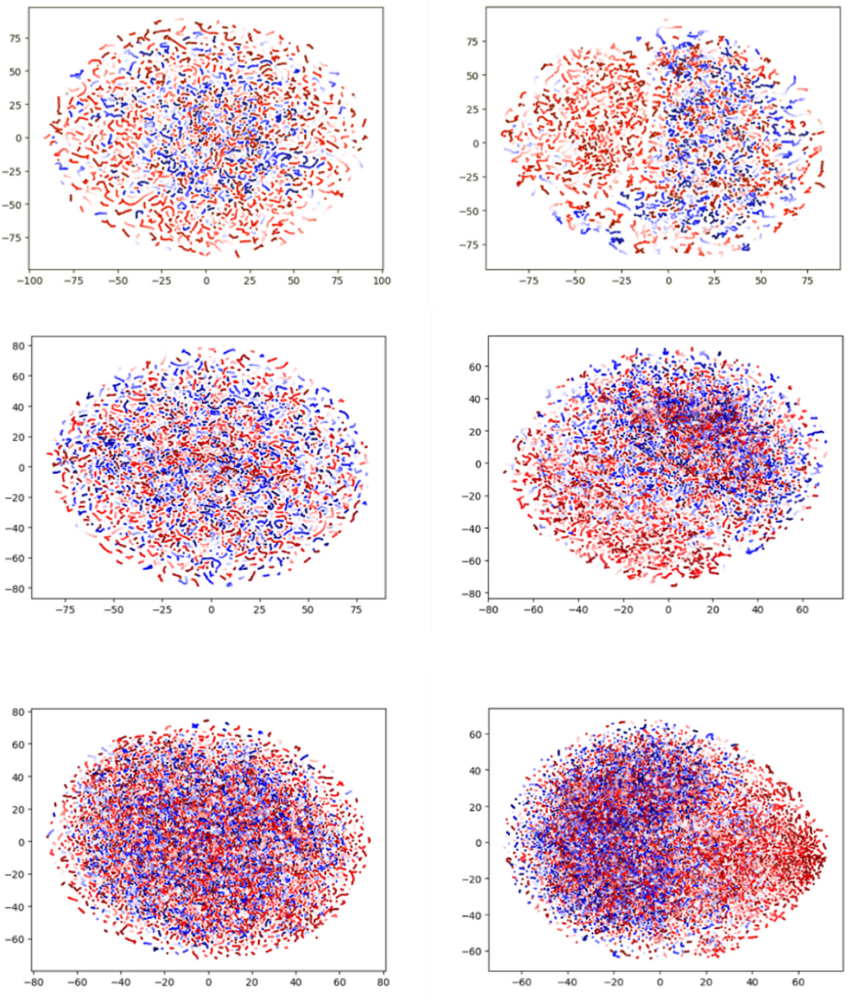


Fig. 9. S0-S2, Raw data(left), extracted feature(right)

As shown in the comparison chart (see in Fig. 9.), there is an obvious change in the distribution of the original data vector and the transformed feature vector in the feature space. The red and blue colors are mixed in the plot of the original data. The red points are still distributed in the feature vector plot at various locations, but the blue points are distributed only at specific locations. This implies that the subjects' cognitive activities during the pre-search phase had a high degree of consistency. However, when subjects drive their hands to click, their cognitive activity may not be purely motor intent, which leads to the red dots in the later stage being similar to the blue dots on the pre-feature space.

In this paper, the search activity is defined as 0, and the click activity is defined as 1. The data points of each trial are converted into the corresponding class according to the class clustering centers learned from the data by the K-means algorithm. As shown in Fig. 10., there is a period of pure 0 at the beginning of the curve, which indicates that the subjects' cognitive activity is homogeneous in the early stages. Over time, the latter part of the curve tends to 1, but it is not pure. This verifies that the blue data points belonging to the earlier period are relatively concentrated. However, the red data points belonging to the later period are also mixed into the blue clusters.

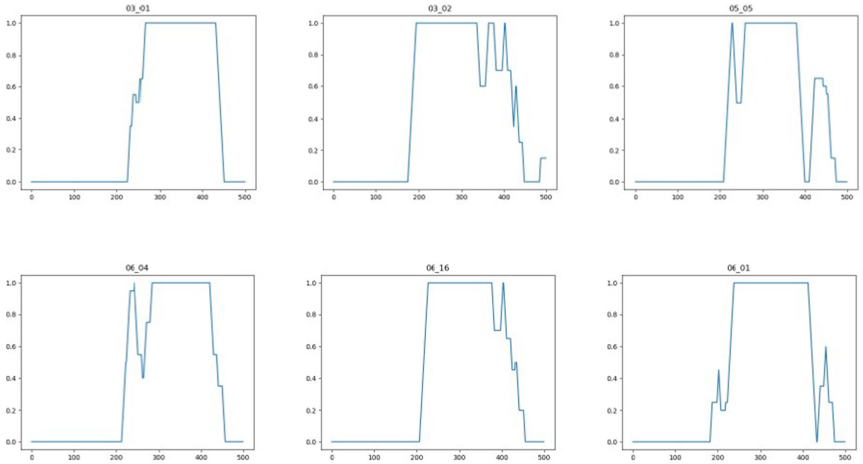


Fig. 10. The segment of data consistent with our expected scenario, i.e., the presence of two changes in cognitive activity from search to click for the subject in each trial. (In order to avoid the interference of a particular outlier, a window of length 10 sampling points is used in this paper and the number of 0 and 1 values within the window is averaged so that the vertical axis of the graph ranges from 0 to 1.)

6 Conclusion

In human-computer interaction, the identification of human cognitive activities is important which can help us design more rational human-computer interfaces. In this paper, in order to study the changes in subjects' cognitive activities in a visual search task based on EEG, a "target search and action execution" task was conducted, where subjects were asked to look for a number and click on it. Because this experiment studies the endogenous cognitive activity of the subjects during the experiment, the experiment is different from previous experimental paradigms, whether studying subjects' emotions or subjects' responses to a particular stimulus. The uncertainty of the time point at which the subjects switched cognitive activities during the task became difficult in this study. Therefore, this paper proposes to use BERT, a time series processing model in the field of NLP, based on the self-supervised learning task, to obtain the encoded form of the sampled EEG signal values at each moment of the experiment. The representational meaning of the same waveform in EEG signals is related to its context, so each moment's encoding requires the integration of information before and after the signal. Raw data will be mapped into the feature space through the model. The results showed that the feature space significantly reflected the clustering of EEG signals near the search point to a particular location compared to the raw data. However, there were also several EEG signals generated later in the task mixed in since the human mind cannot be pure. This may be because the subjects had already started searching for the next target point when they had found the corresponding target point but had not yet completed the click. In the future study, we will further analyze and identify this complex process.

References

1. Lei, W., Le, D.: Risk evaluation of human factors in flight deck system (2010)
2. Wang, Y., Liu, X., Zhang, Y., Zhu, Z., Liu, D., Sun, J.: Driving fatigue detection based on EEG signal, pp. 715–718 (2015)
3. Lin, C.T., Chen, S.A., Ko, L.W., Wang, Y.K.: EEG-based brain dynamics of driving distraction (2011)
4. Hernández-Rojas, L.G., Martínez, E., Antelis, J.M.: Detection of emergency braking intention using driver's electroencephalographic signals. IEEE Lat. Am. Trans. **17**(01), 111–118 (2019)
5. Nguyen, T.-H., Chung, W.-Y.: Detection of driver braking intention using EEG signals during simulated driving. Sensors **19**(13), 2863 (2019)
6. Ditthapron, A., Banluesombatkul, N., Ketrat, S., Chuangsuwanich, E., Wilaiprasitporn, T.: Universal joint feature extraction for P300 EEG classification using multi-task autoencoder. IEEE Access **7**, 68415–68428 (2019)
7. Qian, C., Hou, T., Lu, Y., Fu, S.: Affective recognition using EEG signal in human-robot interaction. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 336–351. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_29
8. Zhang, S., Lu, Y., Fu, S.: Recognition of the cognitive state in the visual search task. In: Ayaz, H. (ed.) AHFE 2019. AISC, vol. 953, pp. 363–372. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-20473-0_35
9. Lee, D.-H., Jeong, J.-H., Kim, K., Yu, B.-W., Lee, S.-W.: Continuous EEG decoding of pilots' mental states using multiple feature block-based convolutional neural network. IEEE Access **8**, 121929–121941 (2020)
10. Wen, D., et al.: Feature classification method of resting-state EEG signals from amnestic mild cognitive impairment with type 2 diabetes mellitus based on multi-view convolutional neural network. IEEE Trans. Neural Syst. Rehabil. Eng. **28**(8), 1702–1709 (2020)
11. Abbasi, M.U., Rashad, A., Basalamah, A., Tariq, M.: Detection of epilepsy seizures in neonatal EEG using LSTM architecture. IEEE Access **7**, 179074–179085 (2019)

12. Sheykhivand, S., Mousavi, Z., Rezaii, T.Y., Farzamnia, A.: Recognizing emotions evoked by music using CNN-LSTM networks on EEG signals. IEEE Access **8**, 139332–139345 (2020)
13. Documents of Quick-30 30-Channel Dry EEG Headset. https://d3ccc04c-e3ae-485a-ab07-ad2d43f82fa1.filesusr.com/ugd/ea87ee_3110a3dc3c1a4e77bdf38b9bd7496759.pdf
14. The Artifact Subspace Reconstruction method of the clean_rawdata plug-in. https://github.com/sccn/clean_rawdata/wiki
15. Zhao, A., Dong, J., Zhou, H.: Self-supervised learning from multi-sensor data for sleep recognition. IEEE Access **8**, 93907–93921 (2020)
16. Shekhar, R., Jawahar, C.V.: Word Image Retrieval Using Bag of Visual Words (2012)
17. Cai, Q.: Research on Chinese naming recognition model based on BERT embedding. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, pp. 1–4 (2019). https://doi.org/10.1109/ICSESS47205.2019.9040736

Human Error and Human Performance

A Qualitative Study on the Workload of High-Speed Railway Dispatchers

Zhenqi Chen[1], Zizheng Guo[1], Guo Feng[2], Lei Shi[1], and Jingyu Zhang[3,4](✉)

[1] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China
[2] Psychological Research and Counseling Center, Southwest Jiaotong University, Chengdu, China
[3] Institute of Psychology, Chinese Academy of Sciences, Beijing, China
zhangjingyu@psych.ac.cn
[4] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

Abstract. In this study, we sought to make an initial exploration to extract the factors that can influence the workload of Chinese high-speed railway (HSR) dispatchers. We also evaluated whether it is applicable to use the Integrated Workload Scale (IWS) scale in this context. By conducting in-depth interviews with five experienced dispatchers, we found that the main factors affecting the workload of high-speed railway dispatchers are the section complexity, the traffic flow, the abnormal situations, and usability of human-computer interfaces. We also found the original IWS scale may not be suitable to assess the workload of Chinese high-speed railway dispatchers due to the differences in work arrangement, equipment, organizational culture, and language. We discussed several ways to reduce the workload of HSR dispatchers and the possible ways to revise this workload measurement in this new context.

Keywords: High-speed railway · Dispatcher · Integrated workload scale · Workload

1 Introduction

High-speed railway (HSR) generally refers to the railway system in which the highest speed of trains can exceed 250 km/h. Since the first high-speed railway was put into operation in Japan (the Shinkansen) in 1964, many developed countries, including Britain, Italy, and Germany, established their own HSR. As a developing country, China has witnessed a remarkable development of HSR in the past 20 years. The length of China HSR has reached 38000 km in 2021 and is still expanding very fast to meet the ever-increasing demand for domestic travel. Although China HSR is the newest system of this kind, the intensive traffic flow, the strict safety requirement, and the implementation of new automated technologies bring new challenges to the dispatching task [1]. Playing a crucial role in the HSR, the dispatchers are responsible for guaranteeing the entire system's safety, punctuality, and efficiency. They have to monitor the system interfaces, organize the traffic flow and resolve potential conflicts. Since their work is essential,

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 251–260, 2021.
https://doi.org/10.1007/978-3-030-77932-0_21

understanding the causes and finding a suitable measurement of their workload is vital for conducting human factor studies in this area. However, current studies may not be enough to answer these two questions satisfactorily [2].

Firstly, previous studies on the analysis of workload sources might not suit the China HSR system. A majority of work in rail human factor studies was conducted on the normal-speed railway system. For example, the U.S. Department of Transportation has conducted a detailed task analysis of American railway dispatchers through observation and structured interviews, but the system is established some 30 years ago [3]. A study of workload on railway traffic dispatchers showed that dispatchers' workload in busy conditions is quite different from the normal condition [4]. An investigation by Reichnach focused on the dispatchers' tasks, such as track-related factors, railway operation-related factors, and dispatcher-related factors [5]. In a study of workload demand on railway dispatchers, the total number of trains and communication frequency was found to play significant roles [6]. In another study, researchers used OWAT, a formula for calculating objective workload, to evaluate and predict the mental workload of dispatchers' monitoring tasks. But the system was still based on a normal-speed railway [7]. To the best of our knowledge, studies about the workload sources of HSR dispatchers were not available.

The second issue is the method of workload evaluation. While objective measures are developing very fast in recent years, self-report questionnaires are still in great need due to the subjective nature of the workload and the easiness of implementation. Whereas general assessment tools such as NASA-TLX, and Subjective Workload Assessment questionnaire (SWAT) are widely used across different working operations, these scales should be used after the mission is finished because it has multiple indicators. However, for tasks such as railway dispatching or air traffic control, an instantaneous measure is needed so the dynamics between traffic flow and workload can be established (for example, the Instant Self-Assessment (ISA) scale [8]). A single item tool that uses the everyday workload description is most suitable to meet such a requirement.

The Integrated Workload Scale (IWS) [9] developed by the Centre for Rail Human Factors, Institute for Occupational Ergonomics, University of Nottingham, was such a tool. Through interviews with railway dispatchers, 47 commonly used descriptors of the workload states were extracted from their everyday experiences. Using carding sorting technique and frequency analysis, researchers classified the descriptors into nine categories. As a result, a unidimensional workload scale with nine-point was formed. The IWS scale has been proved to be reliable and valid to use in the British railway system. However, attributed to the differences in equipment, work content, organizational culture and language, the further verifiable experiment whether this scale is suitable for high-speed railway dispatchers in China.

Therefore, the purpose of this study is to preliminarily investigate the work situation of high-speed railway dispatchers in China and to figure out the source of workload and the applicability of the IWS scale.

2 Method

2.1 Subjects

Five male licensed HSR dispatchers from Chengdu Railway Bureau participated in this study. Four out five dispatchers had bachelor-level degrees, and the average age was 32.6 (SD = 3.90). Their direct HSR dispatching experience was ranging from six months to four years. To note, all of them had worked as signalers and normal-speed railway dispatchers before becoming an HSR dispatcher. Given the HSR system was very new, those with 3 or 4 years of experience are among the most experienced group when they were interviewed (January 2021).

2.2 Interview Protocol

In the interview, the dispatchers were asked to talk about their educational background, working experience, professional training and licensing, the shift arrangement and daily routine. The Behavioral Event Interview (BEI) method was then used to elicit the events that make them feel demanding to find out the key factors affecting workload [10].

2.3 Card Sorting of IWS

We also conducted a card sorting study to investigate whether the IWS scale is usable for China HSR dispatchers. We first translated and back-translated the original nine descriptors of IWS into Chinese, and the mismatches between the versions were resolved by inviting an expert in English-Chinese translation. The nine descriptors were then written on nine independent cards, and we asked the five dispatchers to sort them sequentially according to the workload burdens they described.

2.4 Procedure

The participants were invited to a meeting room of the Human Factors lab at the Southwest Jiaotong University, Chengdu, by the paper's first author. After finishing the informed consent, the interview was conducted online using Tencent Meeting software by a researcher in Beijing (the correspondence author) and another located in Chengdu (the first author). Later, the card sorting was conducted by the first author offline. Afterward, each participant was thanked, paid for 200 yuan and debriefed. The duration was about 120–150 min for each interview, and the whole process was recorded and transcribed for further analyses.

3 Results

3.1 A Brief Description of HSR Dispatchers' Work

The high-speed railway dispatchers worked at the regional dispatching office in which the whole regional railways were separated into several sections. Each section was managed by a dispatching station run by two dispatchers: the main dispatcher and the assistant dispatcher. Generally speaking, the main dispatchers take primary responsibility to monitor and manage the trains. In contrast, the assistant dispatchers manage the signals and equipment (taking the role of signalers and on-site operators in traditional railway systems). The dispatchers work in a rotation manner between the two positions. The working station is a terminal of the CTC system, which connects subsystems and shows all necessary information through the 16 screens. The two dispatchers coordinate and cooperate to complete the routine tasks. Figure 1 shows the workstation and working areas of the two positions.

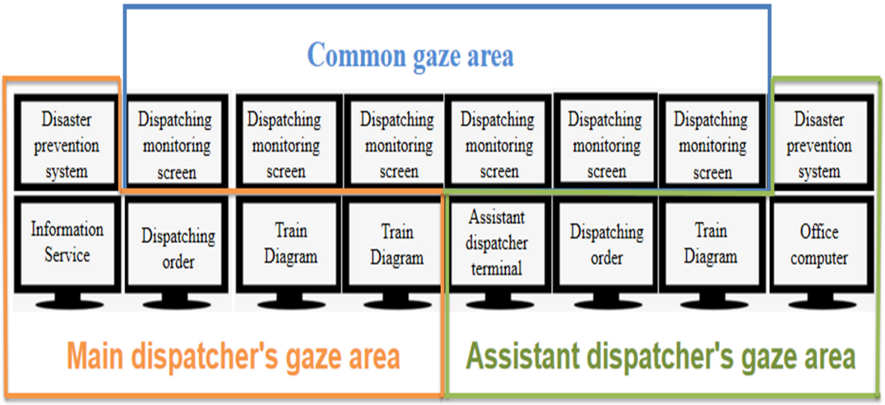

Common gaze area
Disaster prevention system
Dispatching monitoring screen
Dispatching monitoring screen
Dispatching monitoring screen
Dispatching monitoring screen
Dispatching monitoring screen
Dispatching monitoring screen
Disaster prevention system
Information Service
Dispatching order
Train Diagram
Train Diagram
Assistant dispatcher terminal
Dispatching order
Train Diagram
Office computer
Main dispatcher's gaze area
Assistant dispatcher's gaze area

Fig. 1. The overview of workstation and working areas of the two positions.

Through interviews, the dispatchers' tasks can be categorized into five groups: shift handover, emergency handing, operation plan adjustment, construction maintenance and accident prevention. Table 1 shows an overview of some of the dispatchers' primary work.

3.2 The Antecedents of HSR Dispatchers' Workload

Comprehending factors that can influence the dispatching workload is important for future studies. In this study, we extracted the main influencing factors from the interview. These factors can be categorized into four categories.

Section Complexity. The sections varied greatly in their surrounding environment and structure, which may have a considerable impact on the dispatching tasks' difficulty. For example, one dispatcher mentioned that he used to work on one section located in a hilly area with many tunnels. During the rainy seasons, excessive rain can cause water leakage in the tunnels. When this is reported either by the drivers or maintenance staff, the dispatchers should stop the whole line of trains and issue dispatching orders such as turn back, detour and passenger transfer to each train. It reveals that the complex line environment will increase the likelihood of train operation accidents and improve the difficulty of dispatching tasks.

Table 1. Main work activities of the dispatchers

Activity type	Work activities
Shift handover	• Figure out the jurisdiction section's weather conditions and check whether abnormal situations exist in each station • Record the number of train handset through telephone and check the call quality • Fill in the shift record carefully, check Precautions, daily construction plan, determine the time, location, content and impact scale of construction, read the shift plan carefully • After taking over the shift, log in to the dispatching command system, place the I.D. card and dispatching certificate, and place the mouse on the safety island when not needed
Emergency handing	• Pay attention to and analyze dangerous and emergencies on the monitoring screen • Communicate with the station to understand the site situation in detail after finding the dangerous situation • Contact the driver concisely and accurately to report information and dispose of it in strict accordance with the specified procedures • Consider the train parking plan in advance • Inform the relevant departments to carry out an emergency repair and Adjust the operation plan to facilitate the emergency repair departments to enter and leave for maintenance • Organize to restore the order of operation after the repair
Operation plan adjustment	• Fully mastery of the schedule, strictly abide by the running interval and be familiar with the layout of each station; • Make reasonable arrangement according to the train level, grasp the site situation and communicate with the site and adjacent stations in time
Construction maintenance	• Understand the construction and maintenance plan in detail and accurately grasp the influence scope of construction and maintenance; • After mastering the accurate information, draw up the dispatching order, carefully check the dispatching order with the site, and make preparations before and after the construction in strict accordance with the procedure
Accident prevention	• Monitoring train operation and line conditions, monitoring natural disaster alarm system, and monitoring route sequence of each station; • After the alarm prompt box pops up, it is necessary to carefully check the prompt box to ensure that the traffic commander will not miss important information by clicking the "OK" key

In addition to the surrounding environment, the structure of the section may also exert influence. The section may differ in structure due to the number of train stations, the magnitude of stations, the type and location of the crossovers, etc. For in-stance, some stations have a two-way crossover, while others only have a one-way crossover. For dispatchers who need to re-route a train, the tasks can be implemented more efficiently if there is a nearby two-way crossover. Otherwise, he needs to make a more complex detour plan. Therefore, dispatchers must master the relevant information of different sections' infrastructure and stations within their jurisdiction.

Traffic Flow. With the increase of traffic flow, the information and operation dispatchers should deal with will increase correspondingly. Traffic density is a leading cause of workload as the dispatcher said: "the daily work at the junction terminal of dispatcher station is busier, and the requirements for the dispatcher are higher."

Besides, speed mixture will make the dispatching task more complicated. Many high-speed railway lines in China are in the state of mixed operation of Electric Multiple Units (EMU). The maximum operation speed may vary from Dong Che (250 km/h), Gao Tie (350 km/h), to track repairing vehicles (160 km/h). The more trains running at different speeds, the more complex the calculation will be. As a dispatcher said: "If a train running at 350 km/h is after another train slower than it, the former train would be forced to slow down. The train running at 350 km/h will be forced to run. If it wants to run as planned without delay, or it is an important train which could not delay, it is necessary to make some adjustments to detain the slow train in front at the next station, how long to hold the slower train is also a question that needs to consider."

Abnormal Situations. Dispatchers generally mention that it is challenging and stressful to control trains under abnormal situations. According to the "Railway technical management regulations" [11] and "Emergency handling operation manual" [12], the sources of abnormal situations are divided into Fixed Equipment, Mobile Equip-ment, Natural Disasters and Other Types. Fixed equipment failure contains wayside equipment failure, communication signal equipment failure and overhead contact line failure. Mobile equipment failure contains Automatic Train Protection (ATP) equipment fault and EMU fault. Natural disaster contains mud and rock flow, earthquake and snowstorm driving. Other types contains foreign material in track and line.

The handling of the abnormal situation remains a challenging work in train dispatching work for many reasons. First, the occurrence of abnormal situation is al-ways hard to predict. Unpredictability is a major source of stress. Second, once it occurs, the scale of its impact is always very broad. The dispatchers need to coordi-nate with many departments to solve the problem. In addition, although there are many emergency response plans, each situation may have certain unique aspects that the dispatchers should make their specific treatment. Of course, new situations may occur when there is no ready emergency response plan available. Under these circumstances, the dispatcher's working experience is essential. Moreover, when an abnormal situation occurs, the dispatching office supervisor will supervise the dis-patchers' work in a very close manner. Therefore, the occurrence of an abnormal situation will also bring tremendous psychological pressure to dispatchers.

The Usability of Human-Computer Interfaces. At present, the CTC system is used in the dispatching work of China's high-speed railway. The CTC system hugely improves the efficiency of train operation and reduces the workload of dispatchers. In the past, the current two-dispatcher station's work should have taken many more signalers, on-site operators, and so on to deal with. However, the interface of the CTC system can still be improved.

The dispatchers mentioned several problems, including alarm display, monitoring display, decision support information display, and communication apparatus. For example, there is too much information in the alarm prompt box, and the size of the prompt box is too small to read efficiently. The dispatchers also suggest that the alarm level should be modified for some situations, and the false alarm rate should be reduced.

For the monitoring display screen, the train delay information is only provided when the train already passes the station's control point. Although the dispatchers can use the train diagram to estimate its possible delay beforehand, a heavy workload may prevent them from doing so, which will affect the subsequent adjustment of the dispatcher's operation plan.

For the decision support system, dispatchers don't have an assistant tool to help them make decisions, for example, if there is a large train delay area. Which train will be given priority? How to organize the train operation to minimize the total delay time and the influence scope? And how to organize the train to resume on time? CTC system cannot give them potential suggestions; dispatchers rely entirely on their knowledge, skills and work experience to solve it.

For the communication system, several functions can be improved to increase the overall situational awareness. For example, during an emergency, the dispatcher needs to call each train driver one by one after shutting the block with a problem. However, the train drivers can only see the block in front of them closed, so they may keep calling the dispatcher to understand the situation, which will interrupt the dispatcher's work. Functions such as radio broadcasting (for the dispatchers) or situation enhancement visualization (for the train drivers) may provide some improvement.

3.3 The Applicability Analysis of and Modification Suggestions for the IWS Scale

In the interview, we conducted a card sorting experiment to explore whether the translated IWS scale can be used in China's high-speed rail industry. Several findings were summarized below:

1. During the sorting processes, all the five dispatchers expressed doubts about the meaning of the descriptors;
2. Four out of five dispatchers sorted the nine descriptors in a different order as compared with the original scale;
3. In the semi-structured interview, the dispatchers were asked to provide examples and cases for each descriptor. However, they found it hard to do so and suggested that the real workload levels may not that much. All dispatchers suggested that the intensive automation used by the HSR system has changed their work significantly (compared to the normal-speed railway system). In this way, the workload levels might not be that much.
4. We collected the specific terms used by dispatchers to describe their workload. The primary purpose is to ensure that the use of the item is familiar with the dispatcher's

language and can accurately reflect the relevant workload. A total of 28 descriptors describing dispatcher workload were extracted from the interview data. Future studies will collect more data.

To sum up, the initial evidence suggests that the original 9-point IWS scale might not suitable for Chinese HSR dispatchers. Before using the scale to evaluate the workload of dispatchers, it needs to be revised carefully.

4 Discussion

To understand the workload of Chinese HSR dispatchers, the present study conducted an initial interview study. The major purposes were to (1) understand the tasks of the dispatchers, (2) extract factors that may influence their workload, and (3) evaluate the applicability of an instantaneous subjective workload assessment tool (IWS).

The results first revealed the necessary work activities of the dispatchers. Compared to traditional dispatchers, the newly introduced automation has dramatically changed the way of dispatchers. The two-position team can handle the traffic in a broad area, which should have been controlled by many different positions in the old systems, resulting in new research questions for the human factor researchers, such as human-automation interaction.

We also found initial evidence that HSR dispatchers' workload is affected by section complexity, traffic flow, the occurrence of abnormal situations and the usability of the human-computer interface. Whereas the research of airspace complexity in air traffic control is quite many, research in the railway industry is relatively less available [13–15]. Systematic research on this issue may produce effective models, methods and tools to redesign the sectors and traffic plans. The emergency training program is needed to improve the dispatchers' preparedness so that they can deal with the suddenly increased workload in this kind of situation. Several aspects of the human-computer interfaces can be improved. For example, the impact of a natural disaster can be visualized so that the dispatchers can quickly implement corresponding operations. The train's position can be more accurately displayed on the dispatching monitoring screen, which is helpful for the dispatcher to grasp all the information to organize the operation of all trains.

Finally, we found that the original IWS scale may not be applied to Chinese HSR dispatchers' workload assessment. The reasons can be many folds. First, there might be certain language differences. For example, whereas 'challenge' was used as a level 9 descriptor in the original English version scale, the participants in our study considered it to be a moderate level. Therefore, It needs to be revised in Chinese before putting into use.

Certain limitations must be addressed before making a conclusion. First, the number of participants in this interview is small, and all dispatchers are from the same railway bureau. Further studies are needed to collect data from more dispatchers over different areas to resolve potential individual and regional differences in dispatchers' work activities and workload related evaluations. Second, only interviews are used in this study. As this method might be subject to memory bias or social desirability bias, future studies may benefit from using an integrated method, e.g., operation log analysis, observation, accident or near-miss investigations.

5 Conclusion

In conclusion, this study reveals the tasks of the Chinese HSR dispatchers and the influencing factors of workload and finds that the IWS scale needs to be revised to meet the working condition of the China HSR. This study provides initial but valuable information for human factors researchers interested in HSR dispatchers' work. Future work is in great need to further promote our understanding of the nature of dispatchers' workload and the methods to control it.

References

1. Zhou, X., Yao, D., Zhu, M., Zhang, X., Zhang, Z.: Vigilance detection method for high-speed rail using wireless wearable EEG collection technology based on low-rank matrix decomposition. IET Intel. Transp. Syst. **12**(8), 819–825 (2018). https://doi.org/10.1049/iet-its.2017.0239
2. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**(1), 1–17 (2015). https://doi.org/10.1080/00140139.2014.956151
3. Roth, E.M., Malsch, N., Multer, J., Coplen, M.: Understanding how railroad dispatchers manage and control trains: a cognitive task analysis of a distributed team planning task. Hum. Factors Ergon. Soc. Annu. Meet. Proc. **43**(3), 218–222 (1999). https://doi.org/10.1177/154193129904300319
4. McDonald, W.: Train controllers interface design and mental workload. In: Noyes, J., Bransby, M. (Eds.), People in Control, The Institution of Electrical Engineers, London, UK, pp. 239–258 (2001)
5. Reinach, S.: Preliminary development of a railroad dispatcher task load assessment tool: identification of dispatcher tasks and data collection methods. Technical Report No. DOT/FRA/ORD-07–13. U.S. Department of Transportation Federal Railroad Administration Office of Research and Development, Washington, DC, USA (2007)
6. Helbig, L.: A Field Study of Work Demands on Railway Traffic Dispatchers in Centralised Centres (eds.). In: The Proceedings of 2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering and The 3rd International Conference on Maintenance Engineering, vol. III, pp.458–461 (2012)
7. Weeda, C., Zeilstra, M.P., Nadashi, N., Scott, A., Mills, A.: Prediction of mental workload of monitoring tasks. In: Nadashi, N., Scott, A., Wilson, J.R., Mills, A., Rail Human Factors Supporting reliability, safety and cost reduction, Proceedings of the Fourth International Conference on Rail Human Factors, London. pp. 633–640 (2013). https://doi.org/10.1201/b13827-83
8. Brennan, S.D.: An experimental report on rating scale descriptor sets for the instantaneous self assessment (ISA) recorder. DRA Technical Memorandum (CAD5) 92017, DRA Maritime Command and Control Division, Portsmouth (1992)
9. Pickup, L., Wilson, J., Nichols, S., Mitchell, L., Morrisroe, G.: The integrated workload scale (IWS): a new self report tool to assess railway signaller workload. Appl. Ergon. **36**, 681–694 (2006). https://doi.org/10.1016/j.apergo.2005.05.004
10. Bethel, C.L., Murphy, R.R.: Review of Human Studies Methods in HRI and Recommendations. Int. J. Soc. Robot. **2**(4), 347–359 (2010). https://doi.org/10.1007/s12369-010-0064-9
11. China Railway Corporation, Railway technical management regulations. China Railway Publishing House (2014). (in Chinese)

12. Chengdu Railway Bureau, Emergency handling operation manual (2020). (in Chinese)
13. Zhang, J., E, X., Du, F., Yang, J., Loft, S.: The Difficulty to Break a Relational Complexity Network Can Predict Air Traffic Controllers' Mental Workload and Performance in Conflict Resolution. Human Factors, (2019). https://doi.org/10.1177/0018720819880646
14. Majumdar, A., Ochieng, W.Y.: Air traffic control complexity and safety: framework for sector design based on controller interviews of complexity factors. Trans. Res. Rec. Trans. Res. Board **2007**(1), 70–80 (2007). https://doi.org/10.3141/2007-09
15. Loft, S., Sanderson, P., Neal, A., Mooij, M.: Modeling and predicting mental workload in en route air traffic control: critical review and broader implications. Hum. Factors **49**(3), 376–399 (2007). https://doi.org/10.1518/001872007X197017

Computer Aided Search Tasks in a Naturally Occurring Environment

Kyle Harrington[1,2](✉), James Pinchin[2], and Sarah Sharples[2]

[1] NIHR MindTech MedTech Co-operative, The Institute of Mental Health University of Nottingham Innovation Park, Nottingham, UK
Kyle.Harrington@Nottingham.ac.uk

[2] Faculty of Engineering, University of Nottingham, University Park, Nottingham, UK
{James.Pinchin,Sarah.Sharples}@Nottingham.ac.uk

Abstract. A prototype application designed to help coordinate the search activities of small teams of people was developed and tested in a field experiment. Implementing a 2 × 2 experimental design, participants worked in teams in order to find a series of "search targets" in a naturally occurring outdoor environment. The paper demonstrates a novel methodology and reports on the key findings of the study with a focus on the effect of various technological components on search performance, situation awareness and user expectations. Finally, general lessons of the trial are discussed in order to further elucidate the potential benefits of decision-making support tools within missing person incidents and to provide guidance on how such systems could be further improved.

Keywords: Search activity · Search and rescue · Foraging · Missing persons · Visual search · Teamwork

1 Introduction

Informal and familial carers are estimated to have an economic contribution in the UK of £132B a year, which is approximately the total annual spend on the NHS [1]. The main factor in the breakdown of care and subsequent institutionalization of those with care and support needs is carer burden caused by difficult to manage behaviors. Missing incidents are one of the main stressors responsible for both carer burden and subsequent institutionalization [2].

Therefore, it is imperative that guardians and care givers are equipped with the tools and strategies to deal quickly and effectively with these situations as and when they occur; both to minimize the immediate risk of the care recipient and to reduce the long-term likelihood of the breakdown of care.

Missing incidents are activities which usually involve a wide variety of individuals, institutions, and complex socio-technical systems in a time-critical situation. Therefore, a thorough understanding of these incidents depends upon an in-depth analysis not only of individual behavior and psychology but also an investigation of how people work in teams during missing in realistic environments.

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 261–279, 2021.
https://doi.org/10.1007/978-3-030-77932-0_22

Whilst there is a wealth of research into laboratory-based visual search tasks [3, 4], other research into the experience of families of the missing, and some Human Factors research into Search and Rescue operations [5, 6], very little research has attempted to identify the demands and constraints placed upon individuals during the early stages of missing person search, in realistic search tasks within naturally occurring environments.

2 Background

2.1 Visual Search

Perhaps the largest area of research focusing on search behavior is visual search, which has been developed from a perspective influenced heavily by cognitive psychology [7]. Visual search experiments are typically involve participants attempting to find a target item amongst a complex array of distractor items in a controlled environment [8]. In these types of experiments, two methods are commonly used: the present-correct method and the reaction-time method. In regard to the former, participants are briefly presented with an array of items for a set time (known as an inter-stimulus interval) followed by a mask. The participant is then asked whether the target is present or absent from the stimulus array and the target-absent response and percent correct is measured as a function of the inter-stimulus interval. If a task can be performed with high accuracy under these conditions even with very short inter-stimulus intervals, it is taken to mean that these items have been processed in parallel. Conversely, if larger stimulus arrays require higher inter-stimulus intervals, it is assumed that a process with limited capacity is required for the search task [9]. The second type of visual search experiment methodology typically employed uses neither a mask nor an inter-stimulus interval, instead, participants are presented with stimulus array and asked whether the target item is present or absent and their response time is recorded. In this case, if reaction time is independent of the size of the stimulus array, the process is considered to be in parallel [10].

Wolfe [8] describes the how Bottom-Up and Top-Down processes interact in order to direct attention. The ability to perform visual search tasks, therefore, depends upon sufficient difference for bottom-up processes to distinguish a target item from its neighbor, as well as the correct expectation about what the target should look like to direct top-down processes. In the context of missing incidents, this suggests that correct expectations about what the missing person is wearing will help Top-Down processes direct attention to the appropriate aspects of their visual field, whilst Bottom-Up processes will be improved if the missing person is wearing clothes that contrast with their environment.

Findings from visual search experiments shed light on an important aspect of search behavior and demonstrate the relationship between fundamental neurological processes and higher order cognitive expectations. However, whilst profoundly relevant to missing incidents, they highlight only a very small aspect of the overall task. In addition, visual search experiments take place in extremely controlled and artificial environments that lack the context of missing person search and hence, have a low ecological validity.

2.2 Wayfinding, Navigation and Orientation

Navigation is often described as goal-directed movement through the environment which consists of two distinct aspects: Wayfinding and Locomotion [11, 12]. Whilst definitions of Wayfinding and Navigation vary somewhat in the literature, Darken and Peterson [13] describe Wayfinding as the cognitive element of navigation; distinct from movement, it relies upon the development and use of a cognitive map, mental map, or mental model of the environment. Locomotion, or simply, motion, is the real-time aspect of navigation in which an individual traverses through space, maintaining balance, avoiding obstacles and barriers, and moving towards an intended location. Golledge offers a similar description, referring to wayfinding as selecting paths from a network and navigation as both a cognitive and spatial process [14]. Navigation as therefore the aggregation of wayfinding and locomotion, and thus both cognitive and motoric.

A necessary requirement of wayfinding also requires that travelers be aware of their own location relative to their destination, as well as to other salient places or objects en route to their location. This knowledge is called Orientation [12]. A full knowledge of one's location with respect to the Earth's surface (Geographic Orientation) is not required to complete most navigation tasks, it suffices that travelers merely know enough about their current position relative to their target destination. Disorientation may occur frequently during a navigation task, particularly in individuals that have a poor sense of direction [15].

Spatial knowledge is required for wayfinding tasks, though there is still debate about how exactly spatial knowledge is represented. Following on from a seminal paper investigating the navigational strategies employed by rats, Tolman describes the idea of cognitive maps, which are mental representations of spatial information [16].

Perhaps the most supported model of spatial knowledge representation is the Landmark Route Survey (LRS). It not only describes the various levels of representation of spatial knowledge, but also how spatial knowledge develops. Siegel and White theorize that spatial knowledge is often fragmented, incomplete and may at times be inconsistent [17]. The LRS model suggests that spatial representations are often a series of multiple, sometimes contradictory 'mini-spatial representations'. According to this model, landmarks serve as associative cues [21]; stimuli that facilitate the recall of appropriate directional responses, but also function as anchor points for organizing other spatial information [14]. Waller and Lippa [18], also lend support for a variation of this theory, and also note the importance of landmarks as 'beacons'; waypoints which indicate progression towards a navigation goal. Landmarks are more likely to be remembered if they have a dominant physical form, are peculiar in some way, or have associated cultural significance [19]. Therefore, the performance of wayfinding tasks can be improved by providing or enhancing the salience of the visual details of specific locations in the environment [19, 20]. Landmark knowledge is often seen as the appropriate conceptual relationships between certain known stimuli and required action, but these cues are often static, orientation dependent and disconnected [13].

Familiarity with a navigation task can facilitate the acquisition of route knowledge; knowledge of paths between landmarks. At this level of representation, travelers can navigate between two or more landmarks, though this navigation may not be optimal. For example, travelers may know how to get from A to B and from A to C but have no knowledge of how to get from A to C. Route knowledge can facilitate the exchange of spatial information and allows common frames of reference to be established [14].

It is suggested that this type of knowledge is organized in the form of a hierarchy of networks, connecting various locations [14].

At the highest level of representation, survey knowledge (also called configurational knowledge) develops, this involves the ability to estimate relative or absolute distances and directions between two locations, even in unfamiliar routes. Configurations often consist of a hierarchy of locations and routes and the spatial distributions between them. Survey knowledge is often presumed to be the highest level of spatial understanding [21] and is said to constitute much more formal knowledge of spatial relations, including an understanding of Euclidean distances [14]. Taken together, these aspects represent the total spatial knowledge structure, or cognitive map. However, this hierarchy has been challenged by other research, which suggests that even once survey knowledge is acquired, psychological distance in cognitive maps is primarily dependent upon route distance, rather than Euclidian distance [22].

An overall model of navigation is illusive, with some models being too specific to certain environments (hence, not easily generalizable) and others failing to capture the salient aspects of navigational tasks [13]. One model that may be useful in considering how people engage in missing incidents is taken from Jul [23]. This model proposes that navigation tasks consist of a hierarchy decisions and integrates goal formation, strategy selection, cognitive maps, and other navigation elements into a unified theory. To fit this model to behavior during missing incidents, goal formation in this case should be considered which places carers choose to search, rather than what to search for more generally. This model also suggests that there is a reciprocal relationship between wayfinding and locomotion. Goal formation informs strategy selection and together with information from the environment and the travelers cognitive map decisions about locomotion are decided and acted upon. Assessment of the current situation may result in a different strategy being employed, or even a different location being decided upon.

An obvious limitation of applying this navigation model to missing incidents, however, is that it is devoid of the specific context in which missing incidents occur and thus does not adequately capture the salient features which differentiate navigation tasks from missing incidents. In particular, the model does not describe the basis on which these assessments are made in relation to the overall goal, nor does it account for decision-making biases or the high levels of stress that can be encountered if a traveler becomes disoriented or is otherwise experiencing high levels of emotional arousal. Nevertheless, the relevance to caregiver behavior during missing incidents is clear and helps to frame the understanding of missing incidents within a wider, better researched academic context.

2.3 Optimal Foraging Behavior

Related to investigations into Wayfinding and Navigation is the research into animal foraging behavior. According to Kalff, Hills and Wiener [24], foraging describes behaviors as diverse as “Picking bananas from banana trees, searching for nebulas in the night sky and hunting for tuna in the open ocean” (p 168). Missing Incidents are tasks which are very closely analogous to “foraging”, they both involve the allocation of time and resources in order to find a specific object in time and space. The modern scientific understanding to foraging behaviors is largely attributed to two papers; MacArthur and

Pianka [25]and Emlen [26]. The first of these such papers [25], attempts to model “optimum theories” of foraging; that is, the allocation of time and resources in such a way to maximize food gathering whilst minimizing energy and time expenditure. Whilst the second [26], derives a series of predictions about animal behavior from elementary mathematical assumptions about similar optimal strategies.

According to Kalff, Hills and Wiener [24] area-restricted search (ARS), is one of the most well-studied behavioral patterns in foraging. ARS patterns involve high turning angles following resource encounters, but lower turning angles in their absence. ARS is considered to be an optimally adaptive response when resources are clustered (patched environments) and typically localizes animals to where resources are more densely clustered. In addition, this strategy has a minimal informational requirement, depending only that the forager changes their behavior on encountering a resource. The efficiency and simplicity of ARS has led some to propose that this strategy may have provided the biological building blocks for the subsequent evolution of attention [27].

ARS suggests that foragers use temporal cues to determine when to leave designated ‘patches’ of the entire search area. According to predictions made by ARS, foragers will spend more time in resource-rich or dense patches, whilst devoting less time to sparse patches. If a certain temporal threshold passes without encounter, foragers will leave the patch (fewer, shallower turns), whereas encounter with a resource target will increase the total patch residency time incrementally.

Whilst it could be argued that research into the evolutionary foraging behaviors of human and non-human animals has little relevance to missing person search, many of the foraging strategies predicted by MacArthur and Pianka [25] and Emlen [26] have been observed in areas as divergent as information foraging on the internet [28], consumer behavior [29] and even the ways in which humans navigate through their own memories [30]. These findings suggest that the control of attention and the control of movement may share a similar evolutionarily optimal control structure. Which lends support to the idea that there may be common ‘foraging’ strategies, as a vestige of our evolutionary baggage. These strategies (or foraging heuristics) may have been beneficial overall but may in certain specific circumstances manifest themselves as maladaptive behaviors. It may perhaps be the case that people are prone to stochastic movement patterns when looking for a missing person but following more prescriptive methods would be more effective.

2.4 Research in the Wild

Research ‘in the wild’, broadly characterizes a paradigm shift within Human-Computer Interaction (HCI) research in which design, development and evaluation of technological interventions take place in-situ, rather than in an artificial laboratory environment [31]. This marked shift allows for technologies to be evaluated and appraised by potential end-users in a way that much more closely resembles how these technologies would be used in a real-world setting and allows researchers to better understand how technologies are likely to be adopted and perhaps even appropriated. Whilst ‘research in the wild’ is a comparably new research methodology broadly within the discipline of HCI, Hoffman and Woods [32] point to the numerous historical threads with have contributed to the understanding of cognitive and collaborative activity across both human and machine agents. They cite research areas as diverse as knowledge elicitation techniques for expert systems [33], Naturalistic Decision Making [34] and Situated Cognition [35] as various attempts to understand interactions between cognition, culture and complex artefacts in context. Whilst the methodology employed here differs slightly from the definition of “in the wild”, as given by Rogers and Marshall [36], in that it does not “create and evaluate new technologies and experiences in situ” (p.6); it does use a new technology in a naturally occurring physical environment for the purposes of evaluating the likely real-world outcomes and associated behavioral change. This methodological approach also bares resemblance to ‘foraging research’, [24, 27, 30] though to the author’s knowledge, no similar experiments have been conducted investigating human searching and foraging behavior in physical environments.

2.5 Situation Awareness

Situation awareness (SA) describes the knowledge an individual has about their environment, often with respect to a specific task or series of tasks. SA is said to be intrinsically linked to decision-making [37] and may in some cases be linked to performance [38]. SA is considered to be of paramount importance in search and rescue and missing person tasks [5]. However, operationalizing SA, for the purposes of measurement, is not always straight forward and two distinct approaches are available: objective and subjective measures. One example of a subjective measure of SA is the Situation Awareness Rating Technique (SART) [39]. This rating scale was developed for use in Aerospace operations, but is general enough to be used for other domains [5, 39]. SART was developed for a post-trial questionnaire that is relatively quick to administer. The SART questionnaire asks participants to rate the variability of the situation as well as their overall impressions of the quality and quantity of information they are receiving on a Likert scale. Conversely, an “objective” approach to assessing SA was developed by Endsley, the Situation Awareness Global Assessment Technique (SAGAT) [40]. Like the SART, the SAGAT was developed within the domain of Aerospace, as a means of assessing the SA of pilots during a simulation. Endsley suggests that an effective measure of SA involves determining an operator’s knowledge of goal-specific and system critical information. However, whilst there is clearly some value in assessing the SA during the task, the SAGAT is a disruptive technique which involves stopping the simulation or activity in order to determine if an operator is aware of the state of various system elements. Additionally, such an approach would be difficult to implement during an experiment involving physical search activity and would be likely to unduly influence behavior and communication. Objective and subjective measures of SA test important aspects of operator knowledge and self-appraisal of that knowledge. For these reasons, this experiment will take one measure subjective measure of SA (the SART), one objective measure of SA (a post-trial questionnaire) and one measure of performance (number of cones found per participant).

3 Prototype Development

Many existing technologies designed to help with missing incidents depend upon tracking and monitoring technologies which may not necessarily be acceptable to or appropriate for all people [41]. Additionally, compliance, battery life and forgetfulness might be limiting factors in the usefulness of these solutions. Previous research indicates that communication is a crucial element of successful search and rescue activity [5]. Additionally, search and rescue experts have expressed concern that information about the missing person may be inaccurate due to potential sources of communication failure [42]. For this reason, a digital technology which facilitates better communication at the early stages of missing person incidents may improve outcomes for missing people and reduce the stress and emotional burden placed on carers and guardians.

Location and Geo-spatial information are also of paramount importance in Search and Rescue operations, and professional search and rescue teams rely highly on map use in order to maintain high SA of the overall search plan [42]. Furthermore, during missing person searches conducted by professional search and rescue organizations, each individual team has access to a radio in order to communicate with the Incident Commander and other members conducting the search.

Those wishing to develop new technologies to help with search tasks must demonstrate that their solutions do indeed lead to better communication, more overall knowledge of the incident, and ultimately, better search outcomes. For this reason, the search technologies must be designed to allow experimental field trials, without endangering the lives of people who are already at risk. One method which would allow for experimental testing, is the use of surrogate search targets in a naturally occurring environment. The application must be able to detect when participants find these targets in order to determine the performance of teams. Participants must also be given a standardized period of time in which to conduct the search activities so as to make search activities comparable across all trials. For this reason, a timer function should be implemented and controllable by the experimenter. The experimenter must also have the ability to change the available features for the purposes of experimentation and comparison. Finally, the experimenter must be able to have an overall view of the search activity during the experiment for the purposes of safety.

3.1 Functional Requirements of a New Technology

In light of these requirements and considerations, a series of functional requirements are explicitly stated below. The requirements 1–5 represent features a future technology may adopt, whereas the requirements 6–11 are necessary components for testing and evaluating the effectiveness of the application.

Requirements for Application

1. Display Map of current location.
2. Display geo-spatial information of individuals and teams' current locations.
3. Display geo-spatial information of individual's and teams' location history.

4. Give users the ability to refer to geo-spatial locations.
5. Facilitate audio communication.

Requirements for Testing

6. Facilitate the recording of surrogate search targets.
7. Store data of surrogate search targets for later analysis.
8. Record participants location histories for later analysis.
9. Produce an “admin” mode in which displays information and facilitates actions not available to users (such as the location of search targets and starting a timer).
10. Produce a “timer” for the purposes of testing and experimentation, controllable by the admin.
11. Enable the admin to control various features of the application for different experimental conditions.

Using these requirements as a guide, an Android application was developed in such a way to test the effect of various system features and allow for experimental trials with reasonable accuracy. The application recorded participants’ movements on the search task and also recorded when a “search target” with an NFC tag had been scanned. NFC tags were used for the purposes of an experimental trial in order to record when participants had found a designated “search target”. The design of the application went through numerous iterations, before being developed in March 2017 (shown in Fig. 1).

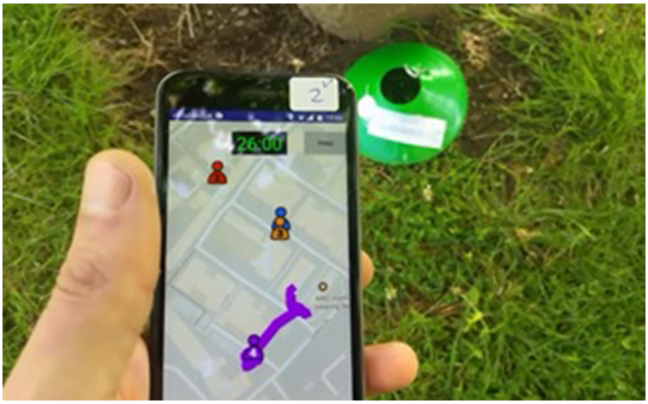


Fig. 1. Application and Search Target

4 Method

Participants arrived at a pre-specified location centrally located on a large UK university campus. They were given an information sheet detailing the study and asked to sign a consent form. Instructional videos were created for the purposes of demonstrating the various system features to participants for each of the conditions. Participants watched these videos after signing the consent form. Following this, they were given an Android mobile phone and a demonstration of the application described above, which was installed. The study implemented a 2×2 experimental design. Participants were given one of two versions of the application: one of these displayed a "shared map", and the other simply displayed the map and participants current location. Half of all participants were connected to a Skype audio conference call and the other half were given the means to make one-to-one phone calls to each other using stored contacts. Each participant was then asked to begin walking in opposite directions until they were no longer visible to their teammates. At this point the researcher began the experiment timer and the participants began the trial.

All participant groups were engaged in searching for a variety of search targets on foot on the grounds of a large university campus whilst using the application described. These targets were sports cones fitted with NFC tags which participants were asked to touch with their mobile phone to confirm they had found the targets. Trials were 30 min long each excluding start times and post-trial interviews, though total participant time was approximately an hour and fifteen minutes. Participant's locations were also monitored at this point to ensure they were not running, and to ensure their safety. Once 30 min had passed, participants were informed that the trial was over and asked to make their way back to the study HQ. Participants were asked at this point not to communicate with each other. Participants then answered post-trial questionnaires described below and a brief semi-structured group interview was conducted.

Conditions. Participants were in one of four conditions, each reflecting different modalities of communication:

Basic Condition (BC). Participants were able to contact each other via standard one-to-one calling features and able to see their current location but had no record of where they or their team members had searched.

Shared Audio Only (AO). Participants were able to communicate to each other via an open radio channel, and able to see their current location, but had no record of where they or their team members had searched.

Shared Visual Only (VO). Participants were able to contact each other via standard one-to-one calling features. In addition, participants were able to see a record of where they and everybody in their team had searched and their current location.

Full Features (FF). Participants were able to communicate to each other via an open radio channel. In addition, participants were able to see a record of where they and everybody in their team had searched, as well as their current location.

4.1 Measurements and Recordings

Participants locations were monitored during the entirety of the experiment, "trails" and cone "scan" activities were all sent automatically by the application to a password protected server, both for use in the trail and later analysis.

Following the trial, participants were issued with a series of questions relating to their performance. Using data generated from the trials, a “map drawing” was produced immediately after the timer had stopped (Fig. 2). This was used as a reference in order to score participants’ objective SA.

Participants also completed a SART questionnaire in order to provide a subjective measure of SA. After the trial was completed and participants had answered the two questionnaires, a semi-structured group interview was conducted and participants were asked how well they thought they communicated, how they thought the technology helped or hindered their task and what improvements could be made if the technology were to be developed further.

The following outcome measures were recorded:

1. Total number of valid search targets found per participant (as a measure of performance).
2. The number of targets which were scanned after they had already been found as a percentage of total cones scanned (taken as a measure of search efficiency).
3. The “Map Score” (as an objective measure of SA).
4. The SART score (as a subjective measure of SA).

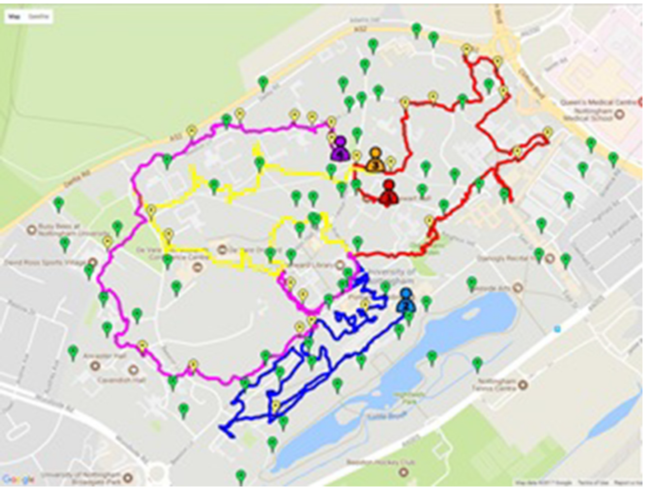


Fig. 2. Map Generated Post-Trial

4.2 Participants and Trial

A total of 52 participants were recruited for this study. Participants could either enlist as complete teams, partial teams, or individuals. Participants who did not enlist as part of a full team were assigned teammates from the participant pool. Of the 52 participants who were recruited, 48 took part in the trial. The 4 participants who were recruited but did not participate in the trial did not attend the trial with a full team and so were unable to take part. Recruitment was conducted largely through an online campaign via Facebook and word of mouth. A Facebook page was created for the purposes of recruitment, and in order to provide a public "scoreboard" for participants' performance during the trial. The advertising campaign took the form of a "Treasure Hunt", and winning teams would win £100 in vouchers as a team prize for all four participants. It was hoped that the prize would act as an incentive for high performance and increase the importance placed on the task by participants. However, for the purposes of fairness a winning team was picked from each of the four conditions in order to ensure that every participant had an equal chance of winning and were not hindered by the experimental condition they were in. Participants were assigned to groups based on an ordering system, so that if the desired number of participant groups were not able to be recruited, there would be a more equal distribution of participants across conditions.

The trial ran from September to November 2017, with three separate blocks of trials, in September, October and November. Between each trial block the researcher performed a visual inspection on the quality, quantity and location of the search targets and ensured the NFC tags were still functional.

4.3 Ethics

Participants were informed prior to the experiment to wear "appropriate" clothing, for the task (e.g., no high heels, weather appropriate clothing). Participants were informed to the purposes and methods employed in the experiment via an information sheet and gave informed consent. Additionally, a tutorial video for each of the conditions was produced and played to participants informing them of specific features and advising caution whilst crossing roads. This study also obtained relevant ethical approval from the University's ethical review board in engineering and was given approval by the department's Health and Safety officer. Participants' data was stored on a password protected server and their data was separated from any potentially identifying information. However, participants were allowed to enter a competition in compensation for their participation, and some participants agreed to sharing pictures of themselves on the study's Facebook page.

5 Results

There was a great deal of variability in the results for all of the dependent variables measured. Despite differences observed between different conditions, few of the results were statistically significant. The results which were significant indicate that knowledge of teammate's locations improved with many-to-many audio communication, and that their own subjective assessment of SA improved with additional visual information about their teammates positions (though, their actual situational awareness or performance did not seem to increase). None of the dependent variables in the results were normally distributed (as indicated by the Sharpo-Wilks), and therefore, non-parametric statistical tests are the most appropriate means of analyzing the data. However, there are no universally agreed upon non-parametric equivalents of the two-way ANOVA, which would be the most appropriate method for analysis. For this reason, the results section will be presented in three parts; a non-parametric analysis of each individual factor, a presentation of the estimated marginal means, and finally, a test for correlation between measures that are assumed to be correlated.

5.1 Non-parametric Analysis of the Effects of Individual Factors

The appropriate non-parametric statistical test for determining if there are differences in the distributions of ratio data between various experimental populations in a between-subjects experimental is the Mann-Whitney U. If each factor (visual information and audio communication) is analyzed individually using this method, two significant results emerge.

The presence of a shared visual map significantly increased participants' subjective SA as measured by the SART questionnaire, from 17.63 (mean of AO & BC) to 20.08 (mean of VO & FF) ($p = 0.04$). However, despite this increase, objective measures of participant's situational awareness actually decreased under these conditions from 6.17 to 9.16, though this difference was not significant.

Conversely, objective SA as measured by the Map questionnaires was positively impacted by the presence of a shared audio channel (conditions AO & FF). The mean value of map scores in conditions with shared audio channels was 6.96/10, as opposed to standard calling features which had an average score of 5.17/10. This was a significant difference with a p-value of 0.004. No other statistically significant differences were observed between conditions.

5.2 Presentation of Means

The highest average performance was obtained by participants who were in the basic condition with no shared visual information and standard calling features (BC; 9.42). The second highest score was achieved by participants who had many-to-many audio communication methods (AO; 8.92), followed by those participants who had many-to-many audio and shared visual information (FF; 7.58), whilst the worst average performance was by participants who had access to shared visual information, but only had access to standard calling features (VO; 6.08). As mentioned earlier, participants who had a many-to-many calling feature typically performed better than those who did not (although the best performers overall did not have this feature), and those participants who were presented with shared visual information performed worse on average than those who did not regardless of audio communication mode, although the difference was not significant.

Subjective ratings of SA increased if either more visual information was present, or if participants had a shared audio channel. The highest average SART scores were obtained by teams who had both additional visual and auditory information (FF; 21.58), scores with either (but not both) of these additional sources of information were comparable (VO; 18.58, AO; 18.33) and the lowest SART scores came from participants who had neither enhancement (BC; 16.92).

However, the effect of various factors on objective SA was not straightforward. Those participants who had access to a many-to-many audio communication tended to perform better than those who did not but additional visual information about teammates locations appeared to benefit only those who had many-to-many audio communication, whilst decreasing the scores of those who did not (FF; 7.03, AO; 6.79, VO; 4.80, BC; 5.50).

Efficiency as measured by repeated scans as a percentage of total scans suggested an interesting relationship. Participants were actually most efficient in the basic condition (BC; 9.1%) but also performed well in the full-features condition (FF; 9.7%), less well in the visual only condition (VO; 13%) and worst of all in the condition with shared audio only (AO; 22.9%).

5.3 Correlations Between Measures

Pearson correlations were conducted on various measures, which could be reasonably assumed to be correlated. However, no such correlations were observed in the data. Firstly, the correlation between SART score (which is a subjective measure of SA) and post-trial Map Score (which is taken to be an objective measure of SA) was tested and no correlation was observed. Similarly, there was no correlation between the number of search targets found and the objective SA. There was a slightly higher correlation between SART and performance, though this was still weak and not statistically significant (0.148).

5.4 Interview Findings

Participants generally indicated that the study design was "well-constructed" (BC, trial 1), and "simple but not easy" (AO, trial 1) and many participants also said that the technology *"has potential helping people with the search"* (FF, trial 2). Participants generally thought the various system features helped their ability to search effectively.

> *"Being in constant contact with each other and being able to see where each other were made us a lot better at what we were doing than if we just had maps and were walking around"* (FF, trial 2).
>
> *"I think the map was particularly important because not only could you see where a person was at that moment, but you could see where that person has been in the past" (FF, trial 1).*

It is also interesting to note that many of those who only received visual information of their own position, but no indication of their teammate's positions (as in BC and AO) expressed that additional visual information would be of benefit to them:

> *"It was good to have that feedback as to where people were, but that feedback would be better on the map. It was silly that we could see ourselves but not other people…It would good to be able to see where everyone else has looked on the map and where they are now"* (AO, trial 2).

Others who also had access to the shared map, (VO and FF) stated that a potential drawback of the highly visual interface was that it may be too distracting.

> *"I think it's maybe not the best idea looking at the screen when you are running around, but the map is really useful. So perhaps I wouldn't look at the screen all the time, but I found myself looking at the screen a lot"* (FF, trial 3).
>
> *"At one point I was looking at the screen and I walked into a building, without looking"* (FF, trial 2).

Overall, participants were generally in favor of additional means of communication to aid the search task, and those who were in conditions without all system features often suggested that such features would be of benefit to them. However, additional suggestions as to how the system could be further improved were made by a few participants and these suggestions typically involved the ability to better plan and coordinate search areas.

> *"I think a bit of planning, the technology seems helpful, but it would be better if it allowed us to plan"*, FF, trial 3.
>
> *"I think the software could help inform some basic strategies, because not everybody will get the idea that you need to spread out. Perhaps some strategies, or whether it's useful to repeat other people's traces and with a missing person they are mobile as well."* FF, trial 2.

6 Discussion

6.1 Performance

There were no statistically significant differences in performance between conditions (where performance was measured cones found per person); neither audio nor visual communication channels appeared to significantly affect total cones found per participant. Neither were there any observed statistically significant differences between groups in the efficiency of search (where efficiency is taken to be an inverse of repeated cones found). However, there was a great deal of variability between individual participants. The standard deviation of cones found was 4.925, whilst the mean score across all conditions was 8.00. Such variation will make statistically significant results less likely. Whilst it is maintained that this experiment represents an important methodological step in evaluating the effectiveness of assistive technologies in a naturally occurring environment, search performance has proved particularly difficult to measure in a way in which controls for a variety of other factors and isolates the effects of various technological interventions. Nevertheless, it is clear that technologies such as these should be properly evaluated in a way which resembles their intended usage. It may be the case that a larger scale study would yield statistically significant differences in search performance between conditions, but future studies may wish to attempt to exert more control over the experimental environment. Good performance on the task is likely down to numerous factors, such as sense of direction, eyesight, and "luck". A repeated measures experiment would be able to control for individual differences, however this would be logistically difficult to implement without replacing the search targets between conditions or changing the environment between conditions.

6.2 Situation Awareness

SA was measured in two ways in this experiment; objectively and subjectively. The objective measures of SA involved asking participants to answer questions about the locations of their teammates once the trial was completed, and the subjective measures involved asking participants to complete a SART, a well-established measure of SA. Interestingly, visual and audio communication channels affected each of these outcomes differently. Shared audio channels were shown to be a significant improvement on the objective measure of SA, whereas, increased visual information only affected the subjective rating of SA. The SART measures amongst other things, the perceived complexity and variability of the situation, but also the quality and quantity of information received. Despite there being no observed differences in the objective measures of SA or performance due to the presence of a shared map, there were significant differences in the SART. Increasing the trust that users' have in the quality of information may be a significant benefit to people in highly stressful missing incident scenarios. Increasing confidence in the quality and quantity of information may have significant real-world benefits, such as reducing the high levels of stress, anxiety and helplessness felt by those conducting missing person searches.

There was no statistically significant correlation between SART and the objective measure of SA. This may indicate that people often incorrectly assess their level of knowledge, or that there is little correlation between "objective" and "subjective" measures of SA. However, perhaps a post-trial questionnaire about various system elements is a measure of recall, not SA. The findings indicate significantly higher recall of teammates locations when many-to-many communication was available, this may be because geo-spatial information presented audibly requires deeper semantic processing than the same information presented visually, and this effort assists recall [43]. Another theory is that participants are more likely to describe landmarks via audio communication and this facilitates survey knowledge [18].

6.3 Extraneous Variables

No trials were run before sunrise or after sunset, but that is not to say that visibility was constant for all trials. Cloud coverage, precipitation levels and time of day may have adversely affected performance in several trials. The weather was not controlled for during the study, and trials were run under numerous environmental conditions, excluding very heavy rain or wind. This is likely not only to affect the visibility of the study, but precipitation may also affect walking speed and crucially, the usability of the touch screen devices participants were issued with. Waldron identified several environmental factors such as temperature and wind speed which affect human behaviors such as travel speed and travel direction [44].

The application was developed on an Android phone, whilst all users in the study confirmed that they were smart phone users, not all were familiar with the Android operating system. An instructional video was presented to all participants explaining various system features and controls. In addition, a brief description of the phone's physical buttons and basic Android operations was given to participants in order mitigate the potentially adverse effects of unfamiliarity with technology. Unfortunately, in a few cases, participants muted their microphones during the trial, whilst this was usually quickly remedied, it did temporary hinder communications between participants.

There were also likely to be other individual differences between participants which could have affected the results, for example navigational abilities [15], familiarity with the campus and inter-team differences.

6.4 The Ecological Validity/Reliability Trade-off

The study took the form of a field experiment and therefore was more ecologically valid than a similar study would have been under laboratory conditions. The price for this higher realism was ultimately less reliability and less control over the environmental factors which may have influenced search task performance. As suggested by Piliavin, Rodin & Piliavin, "the ideal research strategy over the long haul is to move back and forth between the laboratory, with its advantage of greater control, and the field, with its advantage of greater reality" [45]. Whilst it is debatable whether this environment was truly in situ, the use of a naturally occurring environment was crucial in establishing realistic behaviors and limitations of the technology. By conducting the search tasks in the physical world participants were able to comment specifically on how the application was able to meet their requirements and how it affected their physical movement. Participants also mentioned they wanted to ask passers-by if they had seen any search targets; something they would do in a real-missing incident. This experimental set up is far more "in the wild" than controlled laboratory experiments but is not an evaluation of the technology based upon its true intended purpose. Further testing and development would be required before this technology could be tested for use in real missing incidents. Nevertheless, evaluating technologies in environments with high-ecological validity is likely to yield realistic user-behaviors.

6.5 Other Limitations and Areas for Future Work

Whilst the location data for each participant was stored digitally on a remote server for each of the trials, numerous errors with the data storage meant that much of this data was unusable for a more in-depth analysis of pathfinding and search behaviors, and thus made systematic comparisons between conditions impossible.

A further analysis was conducted on the available data in order to determine whether participants during each of the trials were able to detect cones which they passed within 15 m of. The rationale for this analysis was to determine whether participants were less likely to notice search targets which were (theoretically) within their immediate line of sight. Averages for each of the trials which were successfully saved are given in Table1, but shows no clear trends regarding the successfulness of locating targets within a 15-m radius.

Table 1. Percentage of cones found within 15 m of visited locations.

Trial	% found within 15 m	Condition Average
AO1	20	37.28
AO2	59.7	37.28
AO3	32.14	37.28
FF3	37.5	37.5
BC2	68.12	50.725
BC3	33.33	50.725
VO1	61.36	61.36

7 Conclusion

This paper presented a field experiment in which participants, working in teams, were asked to find artificial search targets under strict time conditions in an outdoor area with the use of a digital communication aid. The results show that many-to-many audio communication channels have the ability to increase SA during a search task, whilst additional visual information increases participant's subjective rating of their own SA, despite there being no significant differences in task performance. Whilst there is a clear and pressing need to facilitate effective communication between all concerned parties during missing incidents, more work needs to be undertaken in order to ensure that any potential technological intervention does not unduly distract users from the task.

Acknowledgements. This work was part of an ICase Studentship (13220154) sponsored by the EPSRC and Philips Research, and supported by the RCUK thought the Horizon Digital Economy Research Grants (EP/G065802/1, EP/M000877/1).

References

1. L. Buckner, and S. Yeandle, Valuing Carers: The rising value of Carers Support, Carers, UK (2015)
2. Rowe, M.A., et al.: Ahn, "Persons with dementia missing in the community: is it wandering or something unique?" BMC Geriatr. **11**(1), 1–8 (2011)
3. Ellison, A., Schindler, I., Pattison, L.L., Milner, A.D.: An exploration of the role of the superior temporal gyrus in visual search and spatial perception using TMS. Brain **127**(10), 2307–2315 (2004)
4. Kiss, M., et al.: Efficient attentional selection predicts distractor devaluation: event-related potential evidence for a direct link between attention and emotion. J. Cogn. Neurosci. **19**(8), 1316–1322 (2007)
5. Plant, K.L., Stanton, N.A.: Distributed cognition in search and rescue: loosely coupled tasks and tightly coupled roles. Ergonomics **59**, 1–42 (2016)
6. Adams, J.A., et al.: Cognitive task analysis for developing unmanned aerial vehicle wilderness search support. J. Cogn. Eng. Decis. Making **3**(1), 1–26 (2009)

7. Neisser, U.: Visual search. Sci. Am. **210**(6), 94–103 (1964)
8. Wolfe, J.M.: Guided search 2.0 a revised model of visual search. Psychon. Bull. Rev. **1**(2), 202–238 (1994)
9. Bergen, J.R., Julesz, B.: Parallel versus serial processing in rapid pattern discrimination. Nature **303**(5919), 696–698 (1983)
10. Humphreys, G.W., Quinlan, P.T., Riddoch, M.J.: Grouping processes in visual search: effects with single- and combined-feature targets. J. Exp. Psychol. Gen. **118**(3), 258–279 (1989)
11. Verschure, P.F.M.J., Pennartz, C.M.A., Pezzulo, G.: The why, what, where, when and how of goal-directed choice: neuronal and computational principles. Philos. Trans. R. Soc. B Biol. Sci. **369**(1655), 20130483 (2014)
12. Montello, D.R., Sas, C.: Human Factors of Wayfinding in Navigation. In: CRC Press: Taylor and Francis, Ltd., London, pp. 2003–2008 (2006)
13. Darken, R.P., Peterson, B.: Spatial Orientation, Wayfinding, and Representation, pp. 493–518. In: Stanney, K.M., (ed.) Erlbaum, Mahwah (2001)
14. Golledge, R.G.: Human Cognitive Maps and Wayfinding. In: Golledge, R. R.G., (ed.) Maryland: Johns Hopkins University Press, Baltimore (1999)
15. Kozlowski, L.T., Bryant, K.J.: Sense of direction, spatial orientation, and cognitive maps. J. Exp. Psychol. Hum. Percept. Perform. **3**(4), 590–598 (1977)
16. Tolman, E.C.: Cognitive maps in rats and men. Psychol. Rev. **55**(4), 189–208 (1948)
17. Siegel, A.W., White, S.H.: The Development of Spatial Representations of Large-Scale Environments, pp. 9–55. In: Reese, H.W. (ed.) JAI, Atlanta (1975)
18. Waller, D., Lippa, Y.: Landmarks as beacons and associative cues: their role in route learning. Mem. Cognit. **35**(5), 910–924 (2007). https://doi.org/10.3758/BF03193465
19. Appleyard, D.: Styles and methods of structuring a city. Environ. Behav. **2**(1), 100–117 (1970)
20. Thorndyke, P.W., Goldin, S.E.: Spatial learning and reasoning skill. In: Pick, H.L., Acredolo, L.P. (eds.) Spatial Orientation: Theory, Research, and Application, pp. 195–217. Springer US, Boston, MA (1983). https://doi.org/10.1007/978-1-4615-9325-6_9
21. Hart, R.A., Moore, G.T.: The Development of Spatial Cognition: A Review, pp. 246–288. AldineTransaction, Piscataway (1973)
22. McNamara, T.P., Ratcliff, R., McKoon, G.: The mental representation of knowledge acquired from maps. J. Exp. Psychol. Learn. Mem. Cogn. **10**(4), 723–732 (1984)
23. Jul, S., Furnas, G.W.: Navigation in electronic worlds: a CHI 97 workshop. ACM SIGCHI Bull. **29**(4), 44–49 (1997). https://doi.org/10.1145/270950.270979
24. Kalff, C., Hills, T., Wiener, J.M.: Human foraging behavior: a virtual reality investigation on area restricted search in humans. In: Annual Meeting of the Cognitive Science Society (2010)
25. MacArthur, R.H., Pianka, E.R.: On optimal use of a patchy environment. Am. Nat. **100**(916), 603–609 (1966)
26. Emlen, J.M.: The role of time and energy in food preference. Am. Nat. **100**(916), 611–617 (1966)
27. Hills, T.T., Todd, P.M., Goldstone, R.L.: Search in external and internal spaces: evidence for generalized cognitive search processes. Psychol. Sci. **19**(8), 802–808 (2008)
28. Pirolli, P., Card, S.: Information foraging. Psychol. Rev. **106**(4), 643 (1999)
29. Hantula, D.A., Brockman, D.D., Smith, C.L.: Online shopping as foraging: the effects of increasing delays on purchasing and patch residence. IEEE Trans. Prof. Commun. **51**(2), 147–154 (2008)
30. Hills, T.T.: The evolutionary origins of cognitive control. Top. Cogn. Sci. **3**(2), 231–237 (2011)
31. Chamberlain, A., Crabtree, A., Rodden, T., Jones, M., Rogers, Y.: Research in the wild: understanding 'in the wild' approaches to design and development. In: Designing Interactive Systems Conference, Newcastle Upon Tyne, United Kingdom, pp. 795–796 (2012)

32. Hoffman, R.R., Woods, D.D.: Studying cognitive systems in context: preface to the special section. Hum. Factors **42**(1), 1–7 (2000)
33. Boose, J.H., Gaines, B.R.: Knowledge acquisition for knowledge-based systems: notes on the state-of-the-art. Mach. Learn. **4**(3), 377–394 (1989)
34. Zsambok, C.E., Klein, G.A.: Naturalistic decision making: Lawrence Erlbaum Associates (1997)
35. Hutchins, E.: Cognition in the Wild. MIT press, Cambridge (1995)
36. Rogers, Y., Marshall, P.: Research in the wild. Synth. Lect. Hum. Cent. Inf. **10**(3), i–97 (2017)
37. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Hum. Factors **37**(1), 32–64 (1995)
38. Ma, R., Kaber, D.B.: Situation awareness and driving performance in a simulated navigation task. Ergonomics **50**(8), 1351–1364 (2007)
39. Selcon, S.J., Taylor, R.M.: Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In: AGARD, Situational Awareness in Aerospace Operations 8 p(SEE N 90–28972 23–53) (1990)
40. Endsley, M.R.: Situation awareness global assessment technique (SAGAT). In: Proceedings of the IEEE 1988 National Aerospace and Electronics Conference (NAECON 1988), vol. 3, no. 1988, pp. 789–795 (1988)
41. O'Neill, D.: Should patients with dementia who wander be electronically tagged? No. BMJ Br. Med. J. **346**, f3606 (2013)
42. Harrington, K., Brown, M., Pinchin, J., Sharples, S.: Decision-making within missing person search. Cogn. Technol. Work **20**(4), 665–680 (2018) https://doi.org/10.1007/s10111-018-0515-x
43. Eysenck, M.W., Eysenck, M.C.: Processing depth, elaboration of encoding, memory stores, and expended processing capacity. J. Exp. Psychol. Hum. Learn. Mem. **5**(5), 472–484 (1979)
44. Waldron, J.A.: Human behaviour outdoors and the environmental factors. University of Nottingham (2018)
45. Piliavin, I.M., Rodin, J., Piliavin, J.A.: Good samaritanism: an underground phenomenon? J. Pers. Soc. Psychol. **13**(4), 289 (1969)

Effect of Predictive Next-Letter Highlighting and Its Delays on the Bare-Handed Input in Virtual Reality

Haikun Lin, Shu Ma(✉), Zhen Yang, Chunyan Kang, and Hongting Li(✉)

Department of Psychology, Zhejiang Sci-Tech University, Hangzhou, China
mas@zstu.edu.cn

Abstract. The bare-handed interaction in virtual reality (VR) is gaining popularity because of the high naturalness of the interaction. However, the bare-handed input is disadvantaged by the poor input performance. This study attempted to use a technique - predictive next-letter highlighting (PNLH) to alleviate this problem. In order to investigate the effect of PNLH and the PNLH with different delays on input performance of bare-handed in virtual reality, a total of 30 participants familiar with the keyboard layout were recruited. They were required to complete five input conditions: input without the PNLH, and input with PNLH with a delay of 0, 150, 300, and 450 ms. Results indicated that the PNLH with a delay of 0 and 450 ms had better input performance than that without PNLH. The effect of the delay on input performance varied with duration—a delay of 300 ms will result in the worst input performance.

Keywords: PNLH · Delays · Bare-handed input · VR

1 Introduction

In recent years, collaborative work in virtual reality (VR) has seen a drastic increase in popularity, specifically in the text input, such as giving instructions, entering information, or communicating with people (Löchtefeld et al. 2016; Guttentag 2010; Mujber et al. 2004; Lackey et al. 2016). However, the virtual keyboard could not be shown as approaching the performance of the standard PC keyboard (González et al. 2009). How to improve the input performance in a VR environment is worth exploring.

The PNLH in the soft keyboard was considered an effective technique to improve the input performance (Faraj et al. 2009; Magnien et al. 2004; Bi et al. 2015). PNLH provided visual cues by predicting and highlighting the next possible letter(s) on the virtual keyboard (Sharma et al. 2012). Moreover, PNLH helps users unfamiliar with the keyboard layout not only in finding the target key but also in confirming whether they are pressing the wrong key (Rodrigues et al. 2016). However, virtual keyboards in VR were larger than soft keyboards on smartphones or laptops. The full keyboards in VR were similar to the size of physical keyboards in reality due to Leap Motion capture accuracy issues. Hence, most of the virtual keyboards in VR were out of the participants' central

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 280–291, 2021.
https://doi.org/10.1007/978-3-030-77932-0_23

vision field, but in the peripheral field of view. Thus, whether PNLH can improve the input performance of virtual keyboards in VR requires further verification.

In addition, the PNLH has some limitations which need improvements. PNLH was not able to accurately predict the user's desired letter by highlighting only one key. Several keys must be highlighted simultaneously to increase prediction accuracy. The highlighted non-target stimuli (e.g., a change in key color, key size, or letter color) would capture the users' attention (Rodrigues et al. 2016). Yantis' (1996) proposed that task performance was inferior when participants' attention was distracted by interfering items. We suspected that the performance of PNLH can be improved by reducing the attention capture of non-target. Liu and Bai's (2016) study showed that as the delay increased, participants were able to shift attention away from the non-target position and subsequently inhibit it, and the attention capture effect of non-targets gradually decreased. However, the effect of delays on input performance was in dispute. Previous studies showed that for text input, the faster system response time was, the better input performance would be (Dabrowski and Munson 2011). Furthermore, Dabrowski and Munson's (2011) concluded that the optimal SRT was zero, and delays above 200 ms would be readily perceived and may disrupt the user's work. By contrast, when searching for targets in lines of letters (Schaefer 1990) or 2-D grids of numbers (Thum et al. 1995), optimal response time may produce the fastest performance on the next task. Kohlisch and Kuhmann (1997) also found that participants performed poorly at the lowest delay level and gradually improved with increasing delay time. However, performance deteriorated again beyond a certain delay time. Hence, we suspected that the optimal delay time varies with the different types of tasks; PNLH may have an optimal delay time.

The bare-handed interaction, considered one of the most natural interactions, was a promising application in VR (Prätorius et al. 2014). Therefore, we investigated the effect of the PNLH by the bare-handed interaction in VR (we abbreviate the PNLH by the bare-handed interaction in VR to the PNLH in the following). In summary, this study conducted an experiment to investigate the following questions:

(1) Would the PNLH improve the input performance in VR?
(2) Would increasing the delay time improve the input performance of the PNLH?

2 Method

2.1 Participants

A total of 30 undergraduates and postgraduates (14 males and 16 females, aged 18–25) from a university were recruited for the experiment. All were right-handed daily users of computers with Qwerty keyboards.

2.2 Apparatus

An Asus GL502VS computer (CPU: Intel Core i7 7700HQ; 16 GB, DDR4 2400 MHz memory; 15.6 in. 1920 × 1080 LED display; 8 GB, NVIDIA GeForce GTX 1070 graphics card; and 1 TB hard disk) was used for the experimental program operation and data analysis.

A VR head show (HTC Vive, HTC Inc.), comprising a VR HMD helmet, two laser positioners, and brackets, was used for building and displaying VR environments. The laser positioners were fixed to the diagonal line above the room with brackets to accurately track the position of the VR helmet in the virtual environment. The midpoint of the diagonal line was set as the center to ensure sufficient space for VR.

Furthermore, a Leap Motion controller (Leap Inc. 3.1.3) was used to capture and input the hand movements of users. The controller was fixed on the HMD helmet with the Leap Motion universal VR bracket.

The experimental program was written in C# language and run on Unity Personal (Unity Technologies Inc.). The real-time position of the hand, including the fingers, was captured through the SDK provided by Leap Motion. After a certain algorithm processing, the position was converted into the cursor coordinates projected on the VR space and presented as the shape of the virtual hand. Then, the virtual hand interacted with the virtual keyboard keys to implement the input process of letters. The experiment program recorded the information of the participants and experimental data.

2.3 Experimental Interface

The text input interface in the experiment would be presented perpendicular to the ground, including a short phrase, a virtual keyboard, and an input box. The input box was under the text box and vertically aligned with the text presenting box (Fig. 1).

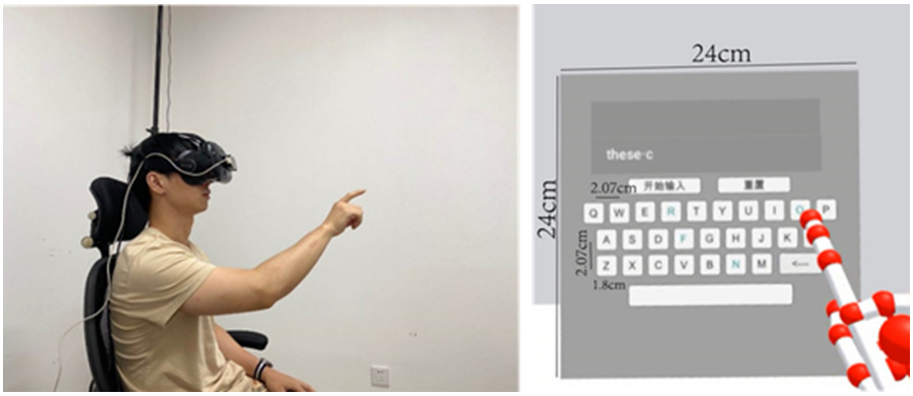


Fig. 1. Experimental scene in reality (left) and experimental interface diagram (right).

The background color of the experiment was the system default, whereas the color of the wall and ground was gray and white, respectively. The background color of the interface was light gray, and the color of the string letters and the letters in the input box was white. Moreover, the colors of the keys of the virtual keyboard and the keyboard letters were white and black, respectively. The virtual key would turn to light blue when the finger successfully touched the key. Then, the virtual keys would change back to their original color after the finger moved away.

The predictive keyboard used in this study was a virtual keyboard that simulates a 100% prediction. After each input of characters, the system predicted the next possible four keys/characters, which were marked with light blue. One of the four keys is the target.

2.4 Input Materials

The material used in the study was from the phrase sets developed by Mackenzie and Soukoreff, which include 500-phrase set with no punctuation symbols and no upper-case letters (Mackenzie and Soukoreff 2003).

In the practice session, five phrases were randomly chosen. In the formal experiment, five groups of phrases with an equal number of characters were presented. Each group consisted of eight phrases. Moreover, the average word length of each group is 4.5 characters (from 23 to 30 characters in each phrase).

2.5 Experiment Design

A within-subject design was used for the experiment. Each participant was required to finish five conditions:

Condition 1: not use the PNLH (NoPNLH)
Condition 2: use the PNLH with a delay of 0 ms (PNLH_0ms)
Condition 3: use the PNLH with a delay of 150 ms (PNLH_150ms)
Condition 4: use the PNLH with a delay of 300 ms (PNLH_300ms)
Condition 5: use the PNLH with a delay of 450 ms (PNLH_450ms)

The dependent variables were the input performance (words per minute, the total error rate, the keystroke reaction time), and the subjective rating on input workload (Workload Profile). They were listed as follows:

1) Words per minute (WPM) is calculated as follows:

$$WPM = \frac{\text{Number of characters} - 1}{\text{Input time}} \times 60 \times \frac{1}{5}. \tag{1}$$

In Eq. (1), the constant 60 is the number of seconds per minute, and the factor of one-fifth accounts for the average length of a word in characters (Arif and Stuerzlinger 2009).

2) The total error rate is calculated as follows:

$$\text{Total error rate} = \frac{INF + IF}{C + INF + IF}. \tag{2}$$

In Eq. (2), the C keystrokes are the correct characters in the transcribed text. The INF keystrokes are the characters that are incorrect and not fixed, whereas the IF keystrokes are the characters that are incorrect and fixed.

3) The Keystroke reaction time (KRT) is calculated as follows:

$$\mathrm{KRT} = \frac{\text{Input time}}{\text{Number of characters} - 1}. \tag{3}$$

The KRT is the average keystroke time for each letter in a phrase after it has been entered. In this experiment, the input time was collected from the input of the first to the last letter (Soukoreff and Mackenzie 2003).

4) Workload Profile

The subjective workload was measured through the Workload Profile (Tsang and Velazquez 1996), based on a multiple-resource model of mental load, including perceptual/central processing, response selection and execution, spatial processing, verbal processing, visual processing, auditory processing, manual output, and speech output. Participants indicated to what extent the recourses were occupied on a 10-point scale ranging from 0 “the resource was not occupied at all” to 10 “the resource was fully occupied.” The test–retest reliability of the scale was 0.92–0.94 (Tsang and Velazquez 1996). Considering that no speech input exists in the input task, we removed the speech input item. The total score of seven items was used in the following analysis.

2.6 Procedure

After arriving in the laboratory, the participants were required to fill in a questionnaire, including gender, age, and experience of using virtual keys. Then, the participants were asked to sit comfortably on a seat that was 60 cm in front of the virtual keyboard, put on the VR helmet, and enter the virtual input environment. Moreover, the experimenter would adjust the height of the seat to ensure that the participants’ eyes were 125–135 cm perpendicular to the ground.

The “memorize” paradigm was adopted to imitate a realistic text input scenario (MacKenzie and Zhang 2001). At the beginning of each trial, a phrase would appear on the text input interface, and the participants were required to memorize it. After memorizing, the participants tapped the “Start input” button, and simultaneously, the phrase would disappear. The participants were asked to input as fast and accurately as possible. If the input was wrong, then the participant may press the backspace key to delete it. If the participant had forgotten the phrase, then they could click the “Reset” button to re-memorize and then input the phrase.

Before the experiment began, the experimenter would introduce the task. Then, five phrases were provided to help the participants get familiar with each condition. During the practice, participants can adjust the posture and distance to the keyboard for comfortable input. In the formal experiment, eight phrases were provided for each condition. The order of the five conditions was counterbalanced. After finishing each condition, the participants were asked to fill in the Workload Profile. When the participants completed all experimental conditions, they were thanked and paid. Figure 2 shows the experimental process.

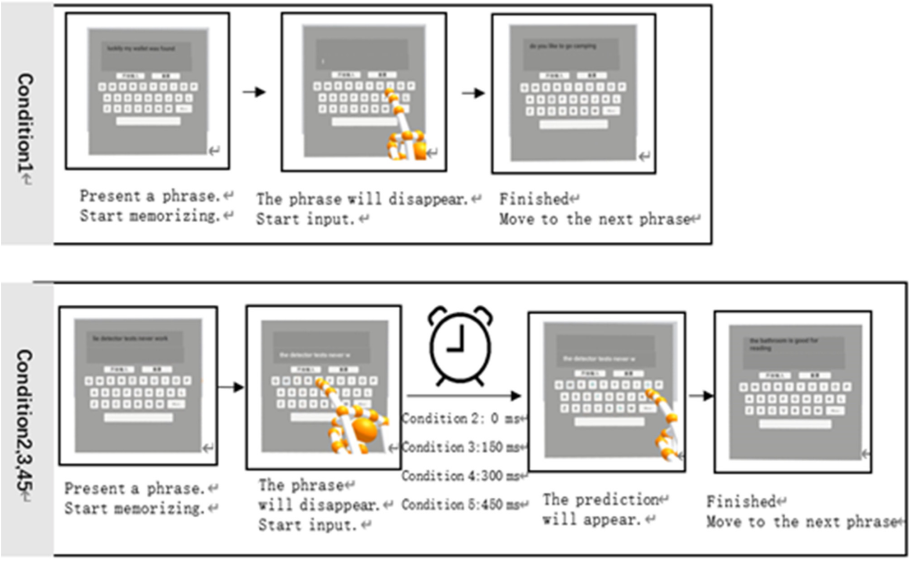


Fig. 2. Input process of each phrase on Condition 1 (up) and Condition 2, 3, 4, and 5 (down).

2.7 Data Analysis

All data were analyzed using SPSS 20.0. The repeated measures ANOVA test was used to analyze WPM, total error rate, KRT, and workload score.

3 Results

3.1 The Effect of Conditions on the WPM

The main effect of conditions on the WPM was significant (F (4, 116) = 2.956, p = 0.023, and partial η2 = 0.093). The post hoc tests showed that the WPM of NoPNLH was marginally significantly lower than the WPM of PNLH_0ms (p = 0.072). The WPM of PNLH_300ms was significantly lower than PNLH_0ms (p = 0.012) and PNLH_450ms (p = 0.008). There was no significant difference between any two of the other conditions (see Fig. 3).

3.2 The Effect of Conditions on the Keystroke Reaction Time (KRT)

The main effect of delays on the KRT was significant (F (4, 116) = 2.669, p = 0.036, and partial η2 = 0.084). The post hoc tests showed that the KRT of PNLH_300ms was significantly higher than NoPNLH (p = 0.031), PNLH_0ms (p = 0.034) and PNLH_450ms (p = 0.003). There was no significant difference between any two of the other conditions (see Fig. 4).

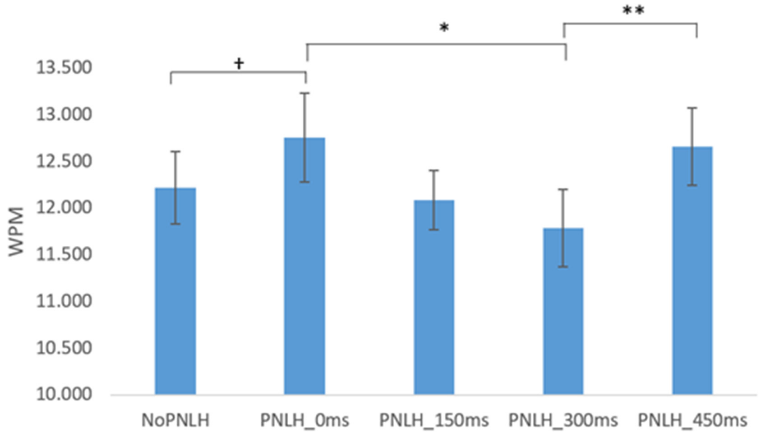


Fig. 3. The Effect of each condition on the WPM. † indicates $0.05 < p < 0.1$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$ (the same below)

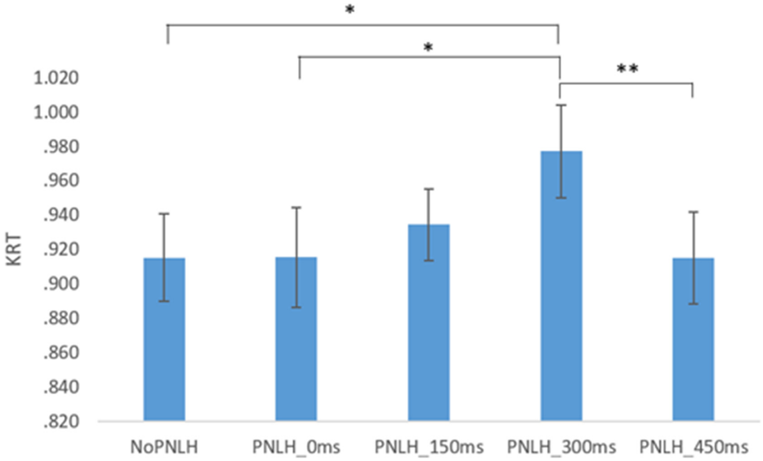


Fig. 4. The Effect of each condition on the KRT.

3.3 The Effect of Conditions on the Total Error Rate

The main effect of conditions on the total error rate was marginally significant (F (4, 116) = 2.412, p = 0.053, and partial η2 = 0.077). The post hoc tests showed that the total error rate of NoPNLH was significantly higher than PNLH_0ms (p = 0.019) and PNLH_450ms (p = 0.050). There was no significant difference between any two of the other conditions (see Fig. 5).

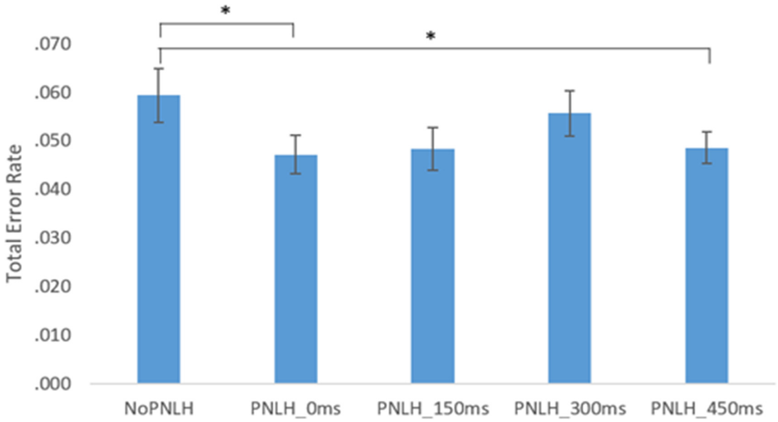


Fig. 5. The Effect of each condition on the total error rate.

3.4 The Effect of Conditions on the Workload Score

The main effect of delays on the workload score was insignificant (F (4, 112) = 0.760, p = 0.554, and partial η2 = 0.026) (see Fig. 6).

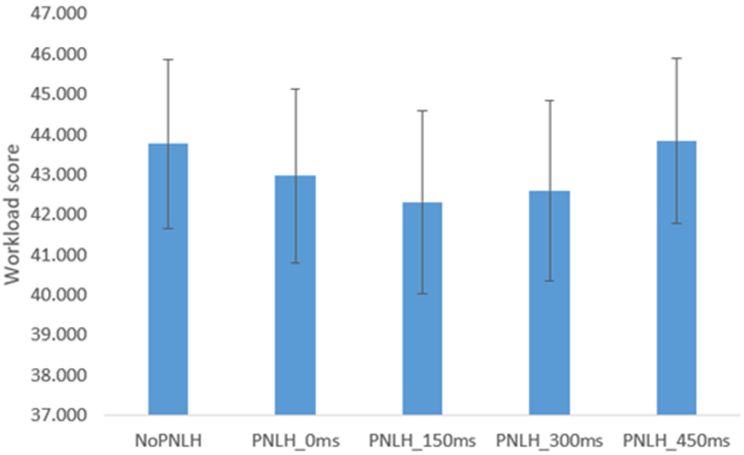


Fig. 6. The Effect of each condition on the workload score.

	NoPNLH	PNLH_0ms	PNLH_150ms	PNLH_300ms	PNLH_450ms	F	p	Partial η2
	(N = 30)	(N = 30)	(N = 30)	(N = 30)	(N = 30)			
WPM	12.221 ± 0.385	12.755 ± 0.476	12.085 ± 0.317	11.788 ± 0.413	12.661 ± 0.414	2.956*	0.023	0.093
Total error rate	0.059 ± 0.006	0.047 ± 0.004	0.048 ± 0.004	0.056 ± 0.005	0.049 ± 0.003	2.412†	0.053	0.077
KRT	0.915 ± 0.026	0.915 ± 0.029	0.934 ± 0.021	0.977 ± 0.027	0.915 ± 0.027	2.669*	0.036	0.084
Workload score	43.759 ± 2.104	42.966 ± 2.163	42.310 ± 2.289	42.586 ± 2.245	43.828 ± 2.060	0.760	0.554	0.026

Data Are Mean ± SD.

4 Discussion

This study explored the following: (1) the effect of PNLH on the bare-handed input performance in a VR environment and (2) the effect of PNLH with different levels of delay (0, 150, 300, and 450 ms) on the performance, and the presence of optimal delay time.

We found that bare-handed input performance on PNLH_0ms was better than NoPNLH. The finding is consistent with previous studies in the soft keyboard (Rodrigues et al. 2016; Bi et al. 2015; Gong et al. 2005).

Furthermore, the effect of the delay on input performance varied with duration, which is worthy of attention and discussion. When the delay was 0 ms, 150 ms, and 300 ms, the input performance tended to decrease along with the increase of the delay. A delay of 300 ms was the worst condition concerning the WPM and KRT. However, when the delay increased to 450 ms, the input performance improved. No more difference in total error rate, WPM, and KRT was observed in PNLH_450ms when compared with PNLH_0ms.

The reason behind the different effects of delay on input performance may be that the information processing during the delay is different. (1) When the delay duration was short, the information processing mainly relayed on the bottom-up processing (Theeuwes et al. 2000). We suspected that the participants relayed on and waited for the cues from the PNLH when the delays were 0, 150, and 300 ms (Dabrowski and Munson 2011). Moreover, during the bottom-up processing, the non-target cues would more likely capture the participants' attention. Hence, the presence of delays reduced the performance: the longer delay duration was, the slower input speed would be. Dabrowski and Munson's study (2011) supported the negative effect of delay duration, where for simple control tasks, the presence of delays decreases user productivity. (2) On the contrary, when the delay duration was longer, the information processing relayed more on the up–bottom processing (Theeuwes et al. 2000). Moreover, the increased delay would inhibit the participants' attention to non-target visual cues in the PNLH. The participants mainly relayed on their memory or experience of the keyboard layout in finishing the input task when the delay was 450 ms. Furthermore, notably, participants could still obtain the benefit from the PNLH with a delay of 450 ms. They have significantly fewer errors compared with NoPNLH.

In summary, every coin has two sides. The delay has negative and positive effects on the input performance of the PNLH. With a delay of 0 ms, attention was captured from the non-target stimulus but without any time delay. With a delay of 150 and 300 ms, the delays were not long enough to reduce the attention capture, but still have negative effects of time delay. With a delay of 450 ms, a positive effect on reducing attention capture from the non-target stimulus is observed, but a negative effect of time delay.

5 Future Work and Conclusion

The study has some limitations and calls for several future works. An analysis of eye movements is necessary to verify the effect of attention capture to better understand the mechanisms of the delay effects on input performance. Ivy et al. (2017) proposed that increasing a constant time delay is a promising strategy for learning and long-term maintenance. Thus, we considered that the delays may also help participants become more familiar with the input material in this study. The findings in our study were suitable for the participants familiar with the keyboard layout. The results may be different for the unfamiliar participants, which needs further investigation. Although the total error rate was significantly decreased by the PNLH, the input speed was only marginally significantly increased, compared with the keyboard without PNLH. Based on Fitts' law, further studies may focus on varying the predicted key's morphology (e.g., magnifying or protruding the predicted keys) to improve the efficiency of input.

In this study, we have proven that the PNLH can improve the performance of bare-handed input in a VR environment for the participants familiar with the keyboard layout. In the meantime, although we failed to further improve the input performance of PNLH by increasing the delays, we found that the effects of the delay on input performance varied with its duration and concluded some guidelines. If delay is inevitable, then a delay of 300 ms was the worst option. Furthermore, if familiarity with the material is needed, then a delay of 450 ms may be a great choice as it is up–bottom information processing.

Declaration of Interest Statement. There are no competing interests in regards to this research by any authors.

References

Faraj, K., Mojahid, M., Vigouroux, N.: Bigkey: a virtual keyboard for mobile devices. In: Jacko, J.A. (ed.) Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction, Part III, pp. 3–10. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02580-8_1

Arif, A.S., Stuerzlinger, W.: Analysis of text entry performance metrics. In: 2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH), pp. 100–105. IEEE, September 2009. https://doi.org/10.1109/TIC-STH.2009.5444533

Bi, J., Liu, H., Ge, L.: Development and verification of the prompt soft keyboards based on Chinese spelling inputs. J. Psychol. Sci. (04), 62–67. CNKI: CDMD: 2.1015.546835 (2015)

Dabrowski, J., Munson, E.V.: 40 years of searching for the best computer system response time. Interact. Comput. **23**(5), 555–564 (2011). https://doi.org/10.1016/j.intcom.2011.05.008

Gong, J., Haggerty, B., Tarasewich, P.: An enhanced multitap text entry method with predictive next-letter highlighting. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1399–1402, April 2005. https://doi.org/10.1145/1056808.1056926

González, G., Molina, J., García, A., Martínez, D., González, P.: Evaluation of text input techniques in immersive virtual environments. In: Macías, J.A., Saltiveri, A.G., Latorre, P.M. (eds.) New Trends on Human-Computer Interaction, pp. 1–10. Springer, London (2009). https://doi.org/10.1007/978-1-84882-352-5_11

Guttentag, D.: Virtual reality: applications and implications for tourism. Tour. Manag. **31**(5), 637–651 (2010). https://doi.org/10.1016/j.tourman.2009.07.003

Ivy, S.E., Guerra, J.A., Hatton, D.D.: Procedural adaptations for use of constant time delay to teach highly motivating words to beginning braille readers. J. Vis. Impairment Blindness **111**(1), 33–48 (2017). https://doi.org/10.1177/0145482X1711100104

Kohlisch, O., Kuhmann, W.: System response time and readiness for task execution the optimum duration of inter-task delays. Ergonomics **40**(3), 265–280 (1997). https://doi.org/10.1080/001401397188143

Lackey, S.J., Salcedo, J.N., Szalma, J.L., Hancock, P.A.: The stress and workload of virtual reality training: the effects of presence, immersion and flow. Ergonomics **59**(8), 1060–1072 (2016). https://doi.org/10.1080/00140139.2015.1122234

Liu, L., Bai, X.: The effects of attentional control setting and types of cues on attentional capture. Acta Psychologica Sinica **48**(9), 1093–1107 (2016). https://doi.org/10.3724/SP.J.1041.2016.01093

Löchtefeld, M., Krüger, A., Gellersen, H.: DeceptiBike: assessing the perception of speed deception in a virtual reality training bike system. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction, pp. 1–10, October 2016. https://doi.org/10.1145/2971485.2971513

MacKenzie, I.S., Soukoreff, R.W.: Phrase sets for evaluating text entry techniques. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 754–755, April 2003. https://doi.org/10.1145/765891.765971

MacKenzie, I.S., Zhang, S.X.: An empirical investigation of the novice experience with soft keyboards. Behav. Inf. Technol. **20**(6), 411–418 (2001). https://doi.org/10.1080/01449290110089561

Magnien, L., Bouraoui, J.L., Vigouroux, N.: Mobile devices: soft keyboard text-entry enhanced by visual cues. In: Proceedings of the 1st French-Speaking Conference on Mobility and Ubiquity Computing, pp. 158–165, June 2004. https://doi.org/10.1145/1050873.1050908

Mujber, T.S., Szecsi, T., Hashmi, M.S.: Virtual reality applications in manufacturing process simulation. J. Mater. Process. Technol. **155**, 1834–1838 (2004). https://doi.org/10.1016/j.jmatprotec.2004.04.401

Prätorius, M., Valkov, D., Burgbacher, U., Hinrichs, K.: DigiTap: an eyes-free VR/AR symbolic input device. In: Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology, pp. 9–18, November 2014. https://doi.org/10.1145/2671015.2671029

Rodrigues, É., Carreira, M., Gonçalves, D.: Enhancing typing performance of older adults on tablets. Univ. Access Inf. Soc. **15**(3), 393–418 (2016). https://doi.org/10.1007/s10209-014-0394-8

Schaefer, F.: The effect of system response times on temporal predictability of work flow in human-computer interaction. Hum. Perform. **3**(3), 173–186 (1990). https://doi.org/10.1207/s15327043hup0303_3

Sharma, M.K., Sarcar, S., Saha, P.K., Samanta, D.: Visual clue: an approach to predict and highlight next character. In: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pp. 1–7. IEEE, December 2012. https://doi.org/10.1109/IHCI.2012.6481820

Soukoreff, R.W., MacKenzie, I.S.: Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 113–120, April 2003. https://doi.org/10.1145/642611.642632

Theeuwes, J., Atchley, P., Kramer, A.F.: On the time course of top-down and bottom-up control of visual attention. Control Cogn. Process. Attention Perform. **XVIII**, 105–124 (2000)

Thum, M., Boucsein, W., Kuhmann, W., Ray, W.J.: Standardized task strain and system response times in human-computer interaction. Ergonomics **38**(7), 1342–1351 (1995). https://doi.org/10.1080/00140139508925192

Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics **39**(3), 358–381 (1996). https://doi.org/10.1080/00140139608964470

Yantis, S.: Attentional capture in vision. In: Kramer, A.F., Coles, M.G.H., Logan, G.D. (eds.) Converging Operations in the Study of Visual Selective Attention, pp. 45–76. American Psychological Association, Washington (1996). https://doi.org/10.1037/10187-002

Methodology to Quantify Accuracy for Procedure Execution Analysis

Sarah O'Meara[1](✉), Kimberly Jenks[1], Christopher Stevens[1], Jennifer Mindock[2], and Stephen Robinson[1]

[1] University of California Davis, Davis, USA
someara@ucdavis.edu
[2] Houston, USA

Abstract. Several industries have sought to identify, mitigate, and predict human errors, as human errors contribute to accidents. We present a new methodology, Procedure Deviation Analysis (PDA), which establishes a human-error taxonomy for repair tasks. Unlike other approaches, PDA defines errors as deviations from the task procedure and considers all operator actions that differ from the procedure to classify and quantify procedure execution accuracy. PDA provides a quantitative measure of accuracy that yields insights for training efficacy and procedure design due to its broader consideration of error. PDA was developed to assess accuracy during repair tasks and to provide comparisons between training techniques. Two main findings of the PDA methodology were the establishment of six deviation modes and the associated application rules. To establish the PDA methodology, eighteen subjects participated in a two-session study and received one of two refresher trainings during the second session. The results indicated that neither refresher training prevented significant performance degradation. The total deviation mode occurrence significantly increased from the first to second session ($p = 0.001$). The average number of steps also increased ($p = 0.036$), due to subjects repeating steps. The percent accuracy significantly decreased from 87.07% to 61.07% ($p < 0.0001$), indicating that knowledge loss from the first to second session that was not adequately addressed in the refresher trainings. Furthermore, PDA highlighted potential areas for improvement in the procedures. PDA provided insights into both training efficacy and procedure quality, two of the key tools used to prevent negative outcomes in high-risk environments.

Keywords: Human error · Error taxonomy · Human performance

1 Introduction

Several industries have sought to identify, mitigate, and predict human errors, as human errors have caused serious accidents in marine shipping [1], aviation [2], nuclear power plants [3], medicine and surgery [4], highway traffic [5], railway traffic [6], mining, construction [7], and maintenance. Human Reliability Analysis (HRA) calculates the probability that human errors occur in predictive applications or post-incident analyses, and several variants have been developed over the years [8–13]. Assessing human error

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 292–307, 2021.
https://doi.org/10.1007/978-3-030-77932-0_24

may also be used as a measure of performance. For example, OCHRA [14] standardizes and systematically detects surgical errors. In this research, we present a new methodology, Procedure Deviation Analysis (PDA), which establishes a human-error taxonomy for repair tasks. In particular, we focus on critical situations where the operator needs to be trained shortly before or during the response and proficient execution of written procedures is paramount. One such use case is during long-duration missions in space, where the nature of the environment makes mistakes costly. Training and procedures are two of the key tools for enabling crewmembers to successfully perform in critical situations, and the PDA methodology provides insights that improve the quality of both.

1.1 Problem Description

Human spaceflight relies heavily on training and procedures to prevent undesirable events in complex systems. PDA was developed to provide quantitative insight into subject execution of task procedures given various training interventions. Unlike other methods, PDA defined errors as deviations from the task procedure and considered all operator actions that differed from the procedure, regardless of the impact on the system state, to classify and quantify procedure execution accuracy. Although the errors captured within PDA may not necessarily prevent the successful completion of the task, they are documented, and are therefore known as "deviations." This was consistent with the assertion that some errors may be characterized as an inevitable part of human adaptation, where the person learned how to optimize their skills through trial-and-error [15]. Rasmussen [15] argued to make these errors observable so systems could be designed to safely allow this experimentation. Therefore, documenting deviations independent of system state may capture this optimization. Furthermore, we did not assign any judgement to the deviation (i.e., good or bad), as we were interested in capturing a wide range of ways in which subjects strayed from procedures. PDA provides quantitative measures of accuracy that yield insights for both training efficacy and procedure design due to its broader consideration of error and detailed analyses. This approach is valuable to the spaceflight community due to the community's existing application of and focus on structured training flows, evaluations, and procedure development.

1.2 Background

Although several industries have used HRA for error analysis, according to a review by Alexander [16] none of the techniques are specifically designed for space operations. Alexander [16] presented the applicability and modification of The Human Factors Analysis and Classification System (HFACS), which identifies and categorizes accident causes [17], to a rocket-powered test flight accident. In the comparison of 15 HRA techniques for space operations, the techniques had different disadvantages like being time consuming ($N = 8$), vague or subjective implementation ($N = 5$), limited scope ($N = 1$), or needed modifications for the relevant application ($N = 1$) [16]. These HRA techniques may be used for complex, large applications that require significant effort, particularly for modeling error. Often, these analyses revealed a series of errors from the individual to organizational level [17]. HFACS identified one level of error as "Unsafe Supervision," that covered failure to provide proper training [17]. Improper training may be highlighted in a post-accident analysis and subsequently corrected; however, it would be beneficial to identify potential training issues and training efficacy in advance. PDA can provide detailed information to find latent errors in procedures used for training, as well as quantitively assess procedure execution accuracy.

To evaluate training, and specifically training to complete procedure-based tasks, a low-level error taxonomy was needed. Many error taxonomies have categories of errors (for examples, see Refs. [18, 19]), such as observation, interpretation, planning, and execution, as is the case in CREAM [13]. For the purposes of improving execution of procedures and the related preparatory training, the execution category was most relevant to our research. Although error taxonomies existed in various domains, the medical domain was chosen as a model for this research due to its critical nature, the use of procedures, and the importance of execution. A review of medical error taxonomies revealed that of 26 taxonomies, 20 addressed external error modes or the behavioral level error classification [20]. Of these 20, a surgical example was chosen as the starting point because of its similarities to the activities subjects performed in this research which incorporated tools, manipulation, and small parts.

2 Methodology

2.1 Approach

To address both training and procedure development, both a person approach and system approach were required. The person approach focused on the individual's actions and assumed the cause for errors originates from within the person (e.g., inattention) [21]. The system approach examined the system and put in places barriers based on the notion of the "Swiss cheese model" that all the holes in these barriers would have to align for an error to occur [21]. These two approaches for error management have different philosophies, but both addressed the use of procedures [21]. By its nature, PDA encompassed both a person and system approach because it captured individual deviations in a generalized taxonomy while maintaining specificity that can be used to both provide user-specific training (person approach) and recommendations for procedure improvements (system approach).

2.2 Purpose

PDA aimed to observe procedure deviations regardless of the impact on the system state. It therefore captured all subject actions without assigning judgement (i.e., good or bad), and subsequently categorized the actions into different modes for further analyses that provided insights into both procedures and training. Accuracy was defined in reference to the procedure, irrespective of the resulting system state, and was a tool to evaluate the procedure as opposed to the system. It would have been inaccurate to describe the deviation modes as errors, which imply a negative consequence. For example, adjacent steps in the procedure may be performed out of order and have no impact on the final outcome but would still be documented as an inaccuracy or procedure deviation. It was also possible for the subject to perform actions outside of the procedure (e.g., tool selection when not specified, writing notes to track parts, temporary part stowage), which may have affected accuracy and/or efficiency, but these actions were not be captured in PDA. While it would be possible to establish deviation modes to address these non-procedural ones, it was not the focus of the present study. The emphasis in this work was to develop the methodology and framework to characterize a range of subject actions and resulting deviations in context of the written procedure.

2.3 Development Process

The PDA methodology development consisted of establishing the relevant procedure deviation modes and the application rules. The application rules determined whether a deviation mode could be assigned to a procedure step and which deviation modes may occur simultaneously. For the procedure used in this study, combinations of two, three, and four modes were possible. Three raters independently analyzed three task execution videos to test the PDA methodology. Any discrepancies in the analyses were discussed to arrive at a consensus and update application rules. The raters iterated with the updated application rules until there were no discrepancies for all selected task execution videos. In this work, the effort focused on developing mature application rules and did not address inter-rater reliability. The overall process is shown in Fig. 1.

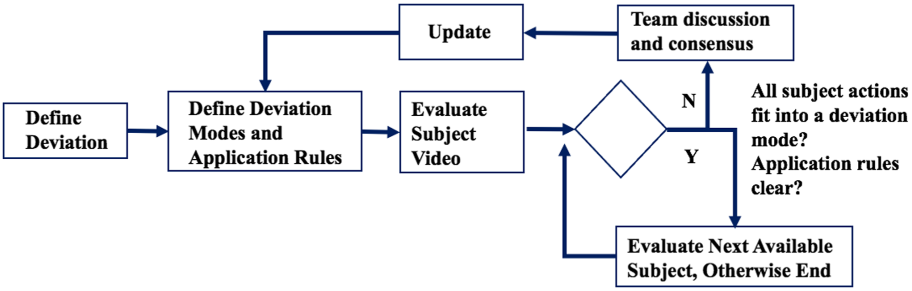

Update
Team discussion and consensus
Define Deviation
Define Deviation Modes and Application Rules
Evaluate Subject Video
N
Y
All subject actions fit into a deviation mode? Application rules clear?
Evaluate Next Available Subject, Otherwise End

Fig. 1. Process to define deviation modes and application rules

The process began by establishing the goals and principles of the analysis, namely that PDA should capture all subject deviations from the procedure. After the definition step, the authors determined a preliminary set of PDA deviation modes based on the literature. From this step forward, the developmental process was iterative and relied on testing the PDA on a subset of task execution videos (i.e., videos of the subjects performing the task). Three raters independently analyzed three task execution videos and compared results. Any discrepancies in the analysis were discussed, the authors arrived at a consensus, and updated the modes or application rules. The updates generally fell into three categories—clarification of wording, new application rule, new deviation mode—with the last being the least common. A typical discrepancy consisted of a misunderstanding of an existing application rule or the rater had encountered a unique deviation and was unsure of how to categorize it. The raters repeated the process until there were no discrepancies across raters for all selected task execution videos.

The procedure deviation modes were based on a list of error modes observed in a laparoscopic cholecystectomy surgical procedure [22] and adapted for the mechanical inspection task specific in this study. Tang et al. [22] separated their "external error modes" into procedural error modes and execution error modes, where the latter address the manipulation of endoscopic instruments. The raters' familiarity with the mechanical inspection task guided them in translating established error modes in a surgical procedure to deviation modes in a mechanical one. As seen in Table 1, several surgical error modes directly mapped to the mechanical procedural deviation modes (Omitted, Partial Omit, Sequential), whereas others were generalized (Extra Action, Execution). As the system and tools used in the mechanical inspection task were not instrumented, nor were the subjects required to use particular tools, quantification of manipulation forces was not possible and surgical error modes 7 and 8 were not used. Conversely, the authors needed to create a "Fragmented" deviation mode. It is important to note that Table 1 presents the final translation of surgical error modes to mechanical procedural deviation modes, which was a product of the raters' task familiarity and an iterative process to establish the application rules.

Table 1. Comparison of surgical error modes [22] to PDA deviation modes

	Surgical Error Modes	**PDA Deviation Modes**
1	Not done	Omitted
2	Partially completed	Partial Omit
3	Repeated	Extra Action
4	In addition	
5	Instead of	
6	Out of sequence	Sequential
7	Too much manipulation	-
8	Too little manipulation	-
9	Wrong manipulation	Execution
10	Wrong object	Execution or Extra Action
-	-	Fragmented

Of the deviation modes, the "Fragmented" and "Extra Action" modes were initially less straightforward to implement. The procedure format used a numbered outline with three levels: section, step, sub-step (formatted as section.step or section.step.sub-step, as needed). Each step or sub-step contained at least one predicate-object pair (a.k.a. "action") in order to complete the step and may include an "Audible Call Out," where the subject had to communicate information about the system. Some procedure steps had multiple actions and it was possible for the subject to complete one or more actions and delay the remaining actions until a later time. This type of deviation had a sequential component, but the appearance of the step multiple times during the task was inherently different than the transposition of an entire procedural step. Thus, a "Fragmented" deviation mode was created to annotate this unique deviation pattern. The "Extra Action" mode was originally a "Repeated" mode, in line with the third surgical error mode (see Table 1). However, the authors realized that 1) it was improbable that a subject would repeat a step in its entirety, and 2) there were several instances when subject deviations did not fall into any mode but needed to be captured. This insight led to the "Repeated" mode transforming into a broader, catch-all mode of "Extra Action." The catch-all nature of the "Extra Action" mode had the benefit of flexibility to encapsulate unpredictable subject deviations without having to revise the existing methodology.

2.4 Methodology Outcomes

The two outputs of the PDA methodology were the establishment of six deviation modes—Omitted, Partial Omit, Fragmented, Sequential, Execution, and Extra Action—and the associated application rules. The deviation modes addressed if a procedure step occurred in the expected order, in its entirety, and as described in the procedure (see Table 2). The deviation modes do not address tool use, or any actions not described in the procedure.

Table 2. Brief description of deviation modes

Deviation mode	Description
Omitted	Does not attempt any part of the step
Partial Omit	Completes part, but not all of the step
Fragmented	Completes the step in multiple parts with other steps in between
Sequential	Performs all or part of the step out of order
Execution	Does not complete the step as described in the procedure
Extra Action	Repeats or attempts to repeat all or part of the step

Application Rules. The most general application rule addressed the procedure level for which a deviation mode can be annotated. The procedure contained six sections and was written to either two (section.step) or three levels (section.step.sub-step), which are considered executable, or containing sufficient instructions for the subject to execute the action(s). PDA deviation modes were only applied to the most detailed level available, and there were 52 of these procedure steps for the selected, experimental task. In general, the numbered procedure step at which a PDA mode can be applied was called the "step" and the PDA deviation mode was always annotated at the step level. Although the PDA modes were annotated at the step level, the actual deviation can occur at the "action" level. A step can be comprised of one or more actions, where an action was defined as a predicate-object pair. Based on the PDA deviation mode's individual definition, it was considered at the step, action, or possibly either level. From the PDA methodology, the deviation modes naturally separated such that two modes fell under each of those three possibilities (see Table 3).

Table 3. Applicable occurrence level by deviation mode

Deviation mode	Step level	Action level
Omitted	Yes	No
Sequential	Yes	No
Execution	Yes	Yes
Extra Action	Yes	Yes
Partial Omit	No	Yes
Fragmented	No	Yes

For a given step, deviation modes could occur singularly, and most could occur simultaneously. For dual mode occurrence, there were the following combinations: Extra Action-Execution, Sequential-Fragmented, Sequential-Partial Omit, Fragmented-Execution, and Partial Omit-Execution. A triple mode could exist for the following: Sequential-Fragmented-Execution and Sequential-Partial Omit-Execution. The Omitted mode was singular and could not occur concurrently with any other deviation modes, as all other deviation modes required that the step occurred. Other deviation modes were mutually exclusive, such as Fragmented-Partial Omit, and Sequential-Extra Action. The diagram in Fig. 2 illustrates how deviation modes may compound. For example, performing a step out of order and incorrectly would result in Sequential-Execution dual mode. A "Correct" mode was shown for completeness in the absence of deviations.

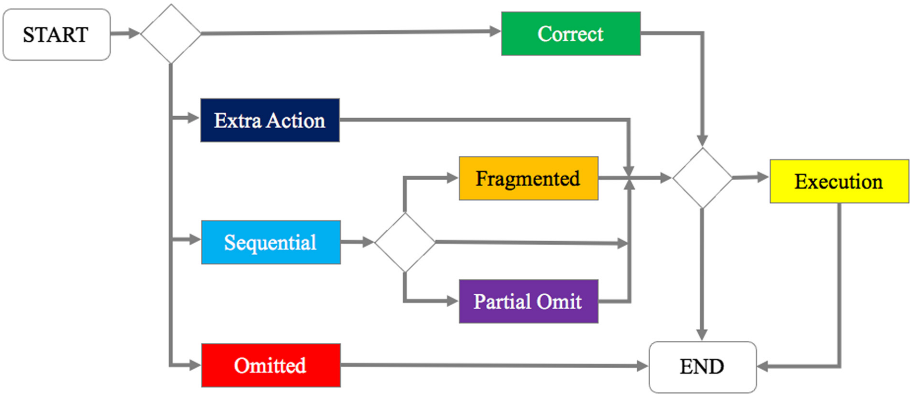


Fig. 2. Diagram illustrating the possible combinations of deviation modes

Almost all of the PDA modes need global knowledge for accurate assignment, or complete knowledge of the task outcome. The Execution mode was the one exception, where only local (or real-time) knowledge was needed. In other words, the rater only needed to review that step in order to assign the Execution mode. For the remaining modes (Omitted, Partial Omit, Sequential, Fragmented, and Extra Action), the rater needed to review the entire task. While the rater assessed the video of the task real-time, continuous updates had to be made to assigned deviation modes based on developing information. In some regards, there was a temporal aspect to the deviation modes where they evolved to different deviation modes. For example, the Omitted mode required that the step was never performed in the entire task. Until the entire task was reviewed, there was a possibility that the step may be performed. If the step was performed at a later time, then the step would be assigned a Sequential mode. Furthermore, the in-progress step had the potential to obtain Fragmented, Partial Omit or Execution modes. A similar analysis was done with the remaining deviation modes, and the results are shown in Fig. 3. From the diagram, the following insights can be attained: 1) the final results were first affected by whether or not a step occurred at all or partially, 2) the modes in the inset could revert to an Extra Action if the step occurred again later in the task, and 3) the Execution mode was independent of temporal dynamics and only depended on real-time information.

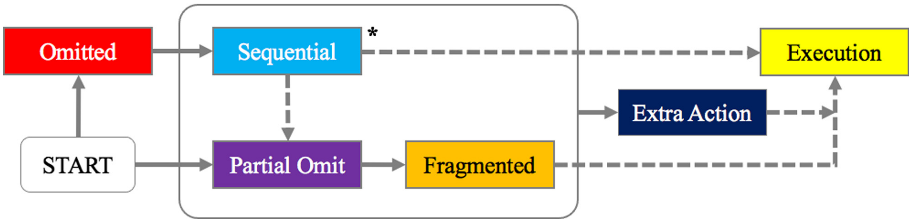


Fig. 3. Potential evolution of deviations modes. Illustrates what deviation modes can be seen real-time, how they may change, and how additional deviation modes are acquired. For example, a step that was not performed at the expected time (Omitted) may become Sequential if the subject preforms the step in its entirety at a later time during task execution. Solid arrows indicate deviation mode updates and dashed lines are for additional deviation mode acquisitions. *The simultaneous occurrence of Sequential-Fragmented would only occur for particular patterns.

In defining the PDA deviation modes and application rules, the raters imposed constraints on which modes could even possibly occur in a procedure step. Any steps that only had one action were not eligible for any action level modes (Partial Omitted and Fragmented). Steps that lacked an evaluation criteria could not be assessed for the Execution mode. In some cases, it would be illogical for the subjects to omit a step because they would not be able to proceed with the remainder of the task. Although, an exceptional circumstance could occur where the subject skipped a large portion of the procedure or exceeded the time limit. Additionally, based on the patterns that had to occur for the Sequential mode, there are some cases when the step was not eligible for this mode (see Sequential subsection). In contrast, the Extra Action mode could always occur because the subject could always repeat part of or the entire step. The task included disassembly to reach the part for inspection followed by assembly to return the task setup to the starting state. Disassembly and assembly steps may be linked and force deviation modes. For example, if the subject partially omitted an action in the disassembly, then there was no physical reason for the subject to perform the assembly action. However, that failure to perform an action was still a deviation from the written procedure and may have an impact on other performance measures, like completion time. Therefore, these linked and forced deviations were annotated.

The application rules determined whether a deviation mode was eligible to apply to a step. The deviation mode definitions, which are detailed in the following subsections, provided the framework to categorize the observed subject performance. Correct steps occurred in the expected order, in their entirety, and as described in the procedure.

Omitted. In the Omitted mode, a subject neglected to perform a step in its entirety and never performed the step at any time during task execution. Since the subject must have neglected to perform the entire step, and therefore any and all underlying actions, this mode was considered at the step level.

Partial Omit. The Partial Omit mode was essentially the Omitted mode considered at the action level. For the Partial Omit mode to have occurred, the subject must have neglected to perform one or more, but not all, actions for the duration of the task execution. Therefore, a step must have contained more than one action to be eligible for the Partial Omit mode.

Execution. The Execution mode compared the subject's execution of a step or action to the criteria listed in procedure, if available and feasible to assess. The procedure listed ten audible call outs without specific wording. The subjects needed to use appropriate, but not exact, verbiage and indicate the appropriate part on the system, as required. There were also two photographs taken of the system and used to check for Execution modes. In contrast, the subjects were permitted to use tools and techniques as desired, and unless specified in the procedure, and could not incur the Execution mode. As a result, the Execution mode only applied to a subset of steps, 18 out of 51. The Execution mode was assessed for the step (including underlying actions) after the subject has proceeded to another step.

Sequential. The Sequential mode occurred when a subject performed a step at an unexpected time in the procedure, either premature or late, and was evaluated at the step level. There were several apparent patterns when sequential modes occurred that can be generally characterized as "transposed", "blocked", or "outlier." A "transposed" pattern occurred when adjacent steps were flipped or transposed (e.g., 1.02 → 1.04 → 1.03 → 1.05). The raters decided to assign the Sequential mode to the higher numbered step (i.e., the step that would have occurred second), to remove the subjectivity of determining if a step was premature (e.g., 1.04) or late (e.g., 1.03). Similarly, in the "blocked" pattern, a group of steps were performed consecutively relative to each other, but the group was out of sequence with the rest of the procedure. In this case, all steps in the group were assigned a Sequential mode. An additional Sequential mode was not assigned to steps within the group that were also out of sequence relative to each other. Therefore, a Sequential mode could only be assigned once per step occurrence. The final pattern, "outlier", referred to a single step that clearly occurred out of order either because it appeared several positions away from where it was expected, or it occurred during another section in the procedure.

Fragmented. The "Fragmented" mode occurred when all the underlying actions of a step were performed, but at different times during task execution. The maximum number of fragments corresponded to the number of underlying actions. Steps with a singular action may never be assigned a Fragmented mode. For the Fragmented mode, there was an inherent non-sequential characteristic of fragmented steps, as actions occurred at an unexpected time in the procedure. An additional Sequential mode was not assigned to these fragments, as it was an inherent property by definition. However, there were three general patterns that occurred where one would necessitate the additional assignment of the Sequential mode to discriminate between them. The first pattern was a simple dispersal, where the fragments occur intermittently between other steps. The second and third pattern occurred when two steps are both fragmented and dispersed by each other. The "ABAB" pattern occurred when the subject began working on the expected step (A), then switched to next step (B), and then alternated between the steps to complete them. The "BABA" pattern distinguished itself because the subject started working on a later step in the procedure first (B), therefore that step needed to be assigned a Sequential mode. It should be noted that the Sequential mode was only assigned to the first "BABA" fragment. This was the only case where the Sequential mode considered part of the step.

Extra Action. The "Extra Action" mode was introduced in the previous section for its unique property that allowed for the flexibility to capture a range of subject procedure deviations. An Extra Action mode was assigned when 1) a step incident occurred multiple times during task execution, 2) the multiple step occurrences was not solely due to a single step fragmented or partially omitted, and 3) the subject attempted any step's action. It is worthwhile to emphasize the third requirement, that the subject made an *attempt*, as it did not require the subject to complete an action or a step. For example, a subject could start manipulating a part associated with the action, and the step with an Extra Action mode would be annotated. The subject was considered as attempting the step if there was visible manipulation of a system part related to a procedure step. Incidental contact, such as moving a part in order to access another part of interest, was not counted. This rule intended to remove rater subjectivity concerning subject intent—was the subject attempting part of the procedure? This rule may have led to an overestimation of subject attempts but provided an objective approach.

Identifying when to assign Extra Action modes proved challenging because a step must occur multiple times during task execution, leading to the question of which step occurrence should be assigned an Extra Action mode. The decision was made to minimize the number of modes assigned because it provided a quantitative, objective measure that could be consistent across raters. Therefore, the raters compared the two step occurrences, identified the more correct step, and assigned the Extra Action mode to the less correct step. In practice, this led to any premature or late step occurrences assigned an Extra Action mode when another step occurrence existed at the expected time in the procedure.

3 Application

3.1 Experimental Context

Eighteen undergraduate subjects participated in the multi-session study approved by the University of California Davis Institutional Review Board to generate data used to develop PDA and assess accuracy during a repair task. In the first session, subjects trained to complete the selected repair task, a float valve inspection for a portable generator, and then performed the task for the evaluation. Subjects followed written procedures to complete the task. A view of the test article is shown in Fig. 4. Subjects in the Treatment group created a self-made refresher video for use in the second session scheduled 6 months later. When subjects from both groups returned for the second session ($N = 16$), they received refresher training prior to the second evaluation: the Control group reviewed the paper procedures and the Treatment group had both their self-made refresher videos and the paper procedures. Videos taken of the task execution during the evaluation sessions were used for PDA methodology development and results. PDA was used to evaluate the hypothesis that self-made refresher videos would result in less performance degradation than the Control refresher training when subjects returned after 6 months.

3.2 Metric Definitions

Three metrics were derived from the PDA analysis: the total deviation mode occurrence, the number of steps, and percent accuracy. For the total deviation mode occurrence, all occurrences across deviation modes were summed for each subject. The number of steps counted the total steps for the entire task execution including all step occurrences for each subject. For example, a step with three fragments would count as three, not one. The percent accuracy was calculated by dividing the number of correct steps by the number of steps. Although not derived from PDA, completion time was also used as another metric in the study to compare performance trends.

3.3 Experimental Data

The raw data used for PDA consisted of videos taken during the evaluations at both sessions and did not include any training runs. There were 18 subjects in the first evaluation and 16 subjects in the second evaluation. The PDA analysis for the second evaluation included a subject that did not finish (the maximum study duration elapsed), but the subject's time was not used in the completion time metric. The raters viewed the videos to annotate the correct steps and deviation modes, which resulted in a coded sequence for each evaluation. The coded sequence was then processed to calculate the metrics of total deviation mode occurrence, number of steps, and percent accuracy.

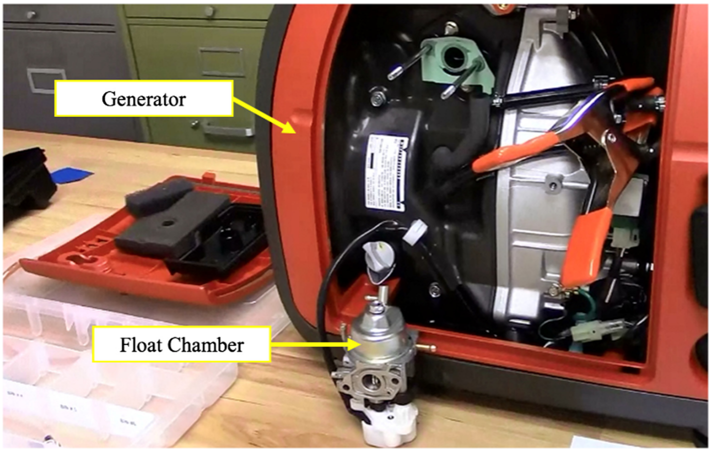


Fig. 4. The task required that the carburetor was removed from the generator. The float valve was located within the float chamber.

3.4 Analysis

Statistical Analysis. The metrics, total deviation mode occurrence, number of steps, and percent accuracy, were used to demonstrate an application—evaluating refresher training efficacy. A linear mixed effects model was used to investigate effects with one between-subjects factor, Group, and one within-subjects, repeated measure, Session. Post-hoc pairwise comparison employed the Tukey Honest Significance Difference test with the Satterthwaite method to calculate adjusted degrees of freedom in R [23]. Effects were considered significant at the $p < 0.05$ level.

For the total deviation mode occurrence, the factor Session was significant (*F(1, 15.84) = 15.84*, $p = 0.001$), but not Group (*F(1, 16.52)* = 0.45, $p = 0.523$) or the interaction of Group and Session (*F(1, 15.84) = 0.37*, $p = 0.551$). The Session factor was also significant for the number of steps (*F(1, 16.16) = 5.25*, $p = 0.036$), but the factor of Group (*F(1, 16.48) = 0.20*, $p = 0.659$) and the interaction of Group and Session (*F(1, 16.16*) = *0.24*, $p = 0.628$) were not significant. The same trend occurred for percent accuracy, where there was a significant main factor of Session (*F(1, 15.16) = 59.32*, $p < 0.001$), but not Group (*F(1, 16.46) = 0.76*, $p = 0.396$) or the interaction between Group and Session (*F(1, 15.16) = 0.44*, $p = 0.519$).

Since the only significant factor was Session and there were not significant interaction effects, the Groups were combined in the following comparisons. The total deviation mode occurrence significantly increased from the first to second evaluation ($p = 0.001$), but not equally across all deviation modes (see Fig. 5). The increase in total deviation mode occurrence from the first session (7.83 ± 7.05, $\mu \pm \sigma$) to the second session (32.75 ± 30.94) followed the trend in increasing completion time. The average completion time increased from 18.48 ± 3.35 min to 25.34 ± 5.82 min. The number of steps also increased ($p = 0.036$), due to subjects repeating steps. The percent accuracy significantly decreased from 87.07% to 61.07% ($p < 0.0001$), indicating that knowledge loss from the first to second evaluation was not adequately addressed by the refresher training in either group.

Therefore, the PDA results did not support the hypothesis that the Treatment group with the self-made refresher videos would experience less performance degradation, as there were no significant differences between groups and performance decreased from the first to second session.

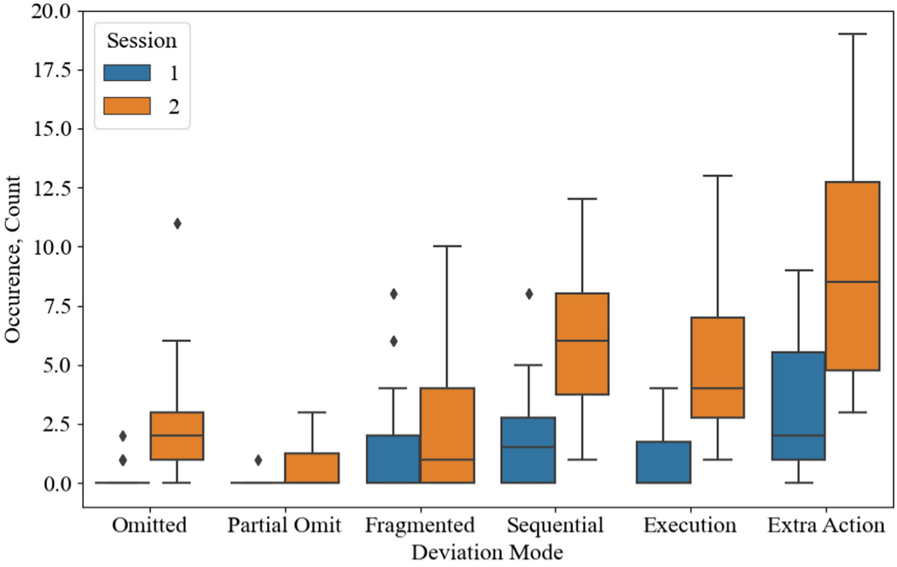


Fig. 5. Comparison of deviation modes occurrences between Sessions for all Groups

Exploratory Analysis and Insights. PDA revealed that some procedure steps tended to accrue deviation modes. These procedure steps may benefit from additional scrutiny to mitigate those deviations. For example, an accrual of Execution modes occurred during procedure steps for layering and replacing tubing. Half the subjects misidentified that the tubing layering was incorrect, which suggested that subjects were not able to identify the layering order. It was suspected that the picture in the procedures of the desired end state did not provide sufficient detail with the lack of labels and text to show or explain the layering order. Furthermore, the tubing removal and replacement also attracted Sequential modes, indicating that subjects either ignored the removal/replacement order or preferred to choose their own order.

Deviations also tended to occur when the procedural step involved small parts. For example, Fragmented and Extra Action modes occurred during the removal and replacement of a gasket. It appeared that the subjects tended to initially neglect the gasket and it may be worthwhile to modify the procedure to call out and highlight this small part. However, there was evidence that subjects self-corrected. It was observed that if a subject recognized an Execution or Partial Omit deviation, Extra Action deviations followed as a correction.

These selected examples demonstrated that PDA provided a technique to produce an annotated sequence of procedure steps that can then be analyzed to provide insight into procedure and training improvement needs. Some deviations appear attributed to the subjects' abilities, like skill-based errors under HFACS "Unsafe Acts," where omissions are memory-based errors and improper execution are technique errors [17]. In other cases, improving the procedures with better figures could prevent deviations and would be similar to a system approach [21].

3.5 Future Improvements and Applications

One of the main limitations of PDA was the implementation time. The time to conduct the PDA varied between the length of the recording for simple assessments and several hours for more complicated ones. The complexity increased with the frequency and number of concurrent deviation modes. Another challenge was compensating for obstructed views, which primarily affected steps involving small parts. For a more practical implementation, the rater could leverage a subset of PDA deviation modes or apply the analysis to selected parts of the procedure, thereby targeting the analysis towards critical aspects of the task.

The PDA methodology has several potential applications that could be derived from this initial work. Now that the methodology has been established, inter-rater reliability could be assessed for applications utilizing multiple raters in the future. In addition to the differences between individual raters, a more in-depth and personalized analysis could be conducted for subjects. Customized procedures are of considerable interest to maximize the training efficacy for individuals, not only on average. More data would need to be collected for individuals in order to customize procedures, and PDA provides a consistent methodology to use across procedures. A larger data collection could also support developing a real-time predictive model for deviation mode occurrence. Furthermore, PDA is not limited to the metrics used in this work and could be used to calculate other measures of interest, such as subject propensity for deviation modes.

For future translation of PDA outside of research, the barriers are lower for the target domain of human spaceflight than might exist in other terrestrial domains. Human spaceflight has a strong organizational culture and infrastructure supporting structured training flows, evaluations, and procedure development that lend well to incorporation of such observation and analysis techniques. Translation to a broader set of real-world applications largely depends on the availability of video data and detailed procedures for application of these analysis techniques, along with the personnel possessing sufficient system knowledge to perform the analysis. In organizations for which detailed level of action is not currently captured for analysis, the establishment of such a capability would no doubt provide helpful insights regardless of the application of PDA or other analysis techniques. Other domains may not have as much infrastructure in place to be leveraged and would likely need increased visibility into the actions taken by operators.

4 Conclusions

PDA was developed to assess accuracy of procedure execution, which can then be used to evaluate training efficacy and improve procedures. The results provided by PDA did not support the hypothesis that self-made refresher videos would lessen performance degradation after a 6-month break. Performance decreased, as reflected in increased total deviation mode occurrence, increased number of steps, decreased accuracy, and increased completion time. These metrics indicated that regardless of the refresher training, performance decreased (i.e., decreased accuracy and efficiency). In addition, PDA highlighted common deviation modes for particular steps. Execution modes suggested that the information in the refresher training and/or procedures was not sufficient. Sequential and Fragmented modes raised the question of whether the step order mattered, and possibly the subjects had found a preferred and more efficient ordering. Omitted and Partial Omit modes drew attention to procedure steps at risk of being overlooked. Finally, the Extra Action mode indicated that subjects sometimes would take self-corrected action. These insights are valuable feedback for procedure design and training, such as areas that needed additional and clearer information or practice. In the study's target domain of spaceflight, in which procedures and training are deeply embedded and relied upon for prevention of undesired outcomes, PDA provides valuable quantitative insights into subject performance and ways to support it.

The dataset generated from PDA contains information about how the subject proceeded through a procedure, which can be used in a variety of ways to assess issues at the person and system level. Transforming observed performance to quantifiable deviations provides metrics to compare groups to better understand individual and group differences. PDA can be used in combination with other quantitative performance metrics to provide a more complete understanding of human performance in terms of accuracy and efficiency. Through the power of the detailed quantification, PDA provides insights into both training efficacy and procedure quality, two of the key tools used to prevent negative outcomes in high-risk environments.

Acknowledgements. This work was funded by the National Space Biomedical Research Institute (project HFP03801).

References

1. Celik, M., Cebi, S.: Analytical HFACS for investigating human errors in shipping accidents. Accid. Anal. Prev. **41**, 66–75 (2009). https://doi.org/10.1016/j.aap.2008.09.004
2. Shappell, S., Wiegman, D.: Human error analysis of commerical aviation accidents: application of the human factors analysis and classification system'. Aviat. Space Environ. Med. **72**, 1006–1016 (2001)
3. Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G., Umbers, I.: Nuclear action reliability assessment (NARA): a data-based HRA tool. In: Spitzer, C., Schmocker, U., Dang, V.N. (eds.) Probabilistic Safety Assessment and Management, pp. 1206–1211. Springer, London (2004). https://doi.org/10.1007/978-0-85729-410-4_195
4. Bonrath, E.M., Zevin, B., Dedy, N.J., Grantcharov, T.P.: Error rating tool to identify and analyse technical errors and events in laparoscopic surgery. BJS Br. J. Surg. **100**, 1080–1088 (2013). https://doi.org/10.1002/bjs.9168
5. Stanton, N.A., Salmon, P.M.: Human error taxonomies applied to driving: a generic driver error taxonomy and its implications for intelligent transport systems. Saf. Sci. **47**, 227–237 (2009). https://doi.org/10.1016/j.ssci.2008.03.006

6. De Felice, F., Petrillo, A.: Methodological approach for performing human reliability and error analysis in railway transportation system. Int. J. Eng. Technol. **3**, 341–353 (2011)
7. Garrett, J.W., Teizer, J.: Human factors analysis classification system relating to human error awareness taxonomy in construction safety. J. Constr. Eng. Manag. **135**, 754–763 (2009). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000034
8. Cooper, S.E., Wreathall, J., Thompson, C., Drouin, M., Bley, D.: Knowledge-base for the new human reliability analysis method "A Technique for Human Error Analysis (ATHEANA)." Presented at the International Topical Meeting on Probabilistic Safety Assessment Moving Toward Risk Based Regulation, Park City, UT (US) (1996)
9. Kim, J.W., Jung, W.: A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. J. Loss Prev. Process Ind. **16**, 479–495 (2003). https://doi.org/10.1016/S0950-4230(03)00075-5
10. Akyuz, E., Celik, M.: Utilisation of cognitive map in modelling human error in marine accident analysis and prevention. Saf. Sci. **70**, 19–28 (2014)
11. Williams, J.C., Bell, J.L.: Consolidation of the error producing conditions used in the human error assessment and reduction technique (HEART). In: Safety and Reliability, pp. 26–76. Taylor & Francis (2015)
12. Marseguerra, M., Zio, E., Librizzi, M.: Quantitative developments in the cognitive reliability and error analysis method (CREAM) for the assessment of human performance. Ann. Nucl. Energy. **33**, 894–910 (2006). https://doi.org/10.1016/j.anucene.2006.05.003
13. Hollnagel, E.: Cognitive Reliability and Error Analysis Method (CREAM). Elsevier (1998)
14. van Rutte, P.W.J., Nienhuijs, S.W., Jakimowicz, J.J., van Montfort, G.: Identification of technical errors and hazard zones in sleeve gastrectomy using OCHRA. Surg. Endosc. **31**(2), 561–566 (2016). https://doi.org/10.1007/s00464-016-4997-4
15. Rasmussen, J.: Human error mechanisms in complex work environments. Reliab. Eng. Syst. Saf. **22**, 155–167 (1988)
16. Alexander, T.M.: A case based human reliability assessment using HFACS for complex space operations. J. Space Saf. Eng. **6**, 53–59 (2019)
17. Shappell, S.A., Wiegmann, D.A.: The human factors analysis and classification system–HFACS (2000)
18. Chang, Y.H.J., Mosleh, A.: Cognitive modeling and dynamic probabilistic simulation of operating crew response to complex system accidents. Part 2: IDAC performance influencing factors model. Reliab. Eng. Syst. Saf. **92**, 1014–1040 (2007). https://doi.org/10.1016/j.ress.2006.05.010
19. Shappell, S.A., Wiegmann, D.A.: HFACS analysis of military and civilian aviation accidents: a North American comparison. In: Proceedings of the Annual Meeting of the International Society of Air Safety Investigators. Gold Coast Australia (2004)
20. Taib, I.A., McIntosh, A.S., Caponecchia, C., Baysari, M.T.: A review of medical error taxonomies: a human factors perspective. Saf. Sci. **49**, 607–615 (2011). https://doi.org/10.1016/j.ssci.2010.12.014
21. Reason, J.: Human error: models and management. BMJ **320**, 768–770 (2000)
22. Tang, B., Hanna, G.B., Joice, P., Cuschieri, A.: Identification and categorization of technical errors by observational clinical human reliability assessment (OCHRA) during laparoscopic cholecystectomy. Arch. Surg. **139**, 1215–1220 (2004)
23. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. arXiv Preprint arXiv:14065823 (2014)

Human Factors Analysis for Aviation Accidents and Incidents in Singapore

H. R. Mohandas(✉) and Tham Kah Weng(✉)

Republic Polytechnic, 9 Woodlands Avenue 9, Singapore 738964, Singapore
{ramanathan_mohandas,tham_kah_weng2}@rp.edu.sg

Abstract. The aviation industry is an important sector in contributing to a nation's economy and its growth and development. The continuous improvement in the aviation industry with the adoption of new technology has developed trust and confidence amongst people all over the world. However even with such improvement and technological advances embedded in the aviation industry, accidents still cannot be avoided. An aviation accident is characterized as an incident that is directly related to an aircraft's service. Aviation accidents are often the result of a number of causes and contributory factors, many of which have a human dimension to them. In some cases, human error is a factor in as high as 70% of aircraft accidents (Feggetter 2007). Hence, the challenge for accident investigators is how best to identify and mitigate the causal sequence of events leading up to an accident. The Human Factors Analysis and Classification System (HFACS) is a general human error framework originally developed and tested within the USA military as a tool for investigating and analysing human causes of aviation accidents (Li et al. 2008). The applicability of HFACS to the analysis of large-scale datasets of incidents and accidents has also been demonstrated in both civil and military aviation environments in Taiwan, India and Australia. Further, this framework has been extended and adapted to analyse the underlying human factors causes in accidents involving remotely piloted aircraft, aviation maintenance and railroad accidents. In Asia, Singapore has a strong and robust economy and operates one of the world's most successful and recognisable airlines. It is therefore surprising that little is known about human error in the aviation context. Thus, one purpose of the present study is to assess the utility of the HFACS framework as an error analysis and classification tool for accidents/incidents in the aviation industry in Singapore. Specifically, HFACS will be applied to commercial aviation accident and incident records maintained by the Transportation Safety Investigation Bureau (TSIB) of the Ministry of Transport (MOT). The TSIB is the air, marine and rail accidents and incidents investigation authority in Singapore. Its mission is to promote transport safety through the conduct of independent investigations into air, marine and rail accidents and incidents. A comprehensive review of 75 accidents and incidents that occurred between October 2000 and December 2019 in Singapore has just been undertaken. This is currently being analysed within a HFACS context and the findings will be reported here. Thus, the utility of the HFACS framework will be appraised, an overview of the 75 accidents will be given, and more specific detail relating to human error will be reported. At a later stage, these findings will be compared with other countries in Asia where similar data is available. Researchers claim that the HFACS framework bridges the gap between theory and practice by

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 308–323, 2021.
https://doi.org/10.1007/978-3-030-77932-0_25

providing safety professionals with a theoretical tool for identifying and classifying human errors in aviation mishaps (Wiegmann and Shappell 2001a, b, c). The system focuses on both latent and active failures and their interrelationships, thereby facilitating the identification of the underlying causes of human error. The findings from this research will allow safety managers, administrators and other aviation professionals working in airlines, airports and aviation regulatory organizations to analyse their safety programmes' strengths and weaknesses. This will also contribute to the build-up of a comprehensive database in Southeast Asia. Finally, this study is unprecedented in Singapore and will offer new insights into the nature and trends of human factors in aircraft accidents and incidents.

Keywords: Human factors · HFACS · Aviation safety · Accidents and incidents · Air traffic

1 Introduction

The growth of the aviation industry has made air transport to be one of the fastest modes of public transport across international boundaries. It has provided profits for many airlines and has contributed towards the improvement of the world economy. An airline's profitability contributes towards the strong economy of a country. The sustainability and growth of the aviation industry involves the prerequisites of safety as an important consideration for the public. Hence, aviation safety has been considered as an important aspect for the healthy development of the aviation industry. The conditions during the second half of the 20th century have changed the perception of the aviation industry and the technical issues turned out to be human error issues that have affected aviation organizations and their safety environment (Zhou et al. 2018). Hence, it has been revealed that human errors have been considered as one of the most frequent errors that lead to aviation accidents (Wiegmann and Shappell 2016). Human error has been mostly found prevalent in almost all the aviation accidents, but still no prevention programs have been designed for reducing the occurrence of such errors.

It is clearly revealed that almost 70–80% of all the aviation accidents have occurred due to human errors (Gong et al. 2014). Human errors resulting in aviation accidents have been considered to be the major concern in aviation industries across the world (Salmon et al. 2012). The International Civil Aviation Organization (ICAO) has stated previously that aviation is arguably the safest mode of mass transportation across international boundaries. The shift of the aviation industry from having a fragile safety record to becoming the first ultra-safe system is due to the incessant investment in safety efforts by the aviation community (Yeun et al. 2014). Even after such advancements in aeronautical technology and weather forecasting, aviation accidents still cannot be eliminated. Aircraft accidents arise due to human errors, loss of control, mechanical failure or bad weather (Li et al. 2015).

The airlines work to provide the best flying experience to their passengers concerning their safety, comfort, punctuality and hospitality. These services onboard are carried out by the cabin crews and pilots who are responsible for looking after their passenger's needs. Hence, appropriate training programmes are provided to the aircrew members for maintaining the safety during the air operations and reducing the impact of any type of errors (Vatankhah 2021).

The Human Factors Analysis Classification System (HFACS) based on the 'Swiss-Cheese' model (Resaon 1990) consists of a human error framework that has been mainly used for analysing the human error accidents in the American military aviation operations and the developers of this frameworks have represented its applicability for commercial and general aviation accident analysis (Wiegmann and Shappell 2001a, b, c; Shappell and Wiegmann 2003). Human factors are considered important for understanding human performance and have been mainly used in the investigation of maritime and railway accidents that occur due to human errors (Reinach and Viale 2006; Baysari et al. 2009; Vairo et al. 2017).

Aviation is a major driver in Singapore's economy. The Singapore aviation sector contributes to about 3% of the country's Gross Domestic Product (GDP) and plays an indispensable role enabling other sectors such as tourism, financial services, retail and manufacturing[1]. Singapore is also an air transport hub of the region with the 9th busiest airport in Asia[2] as well as an established aerospace maintenance, repair and overhaul (MRO) sector.

To support Singapore's airports, airlines and other support facilities, the air transportation sector employs about 119,000[3] people in diverse roles such as flight crew, air traffic controller, ground services personnel, engineers, technicians, etc. The focus of this study will be on the flight crew operating aircraft in Singapore Airspace.

2 Aims and Objectives

This section will elucidate aim and formulate objectives to achieve the identified goals. The study will be backed by the following objectives:

1. To assess the utility of the HFACS framework as an error analysis and classification tool for accidents/incidents in the aviation industry in Singapore.
2. To identify the unsafe acts and human errors in aviation concerning the accidents/incidents in Singapore.
3. To analyze HFACS framework towards implementation of operation safety in aviation in Singapore for preventing the serious consequences of aviation accidents/incidents.

3 Literature Review

3.1 Major Aspects of Aviation Accidents

The complexity of the human factors has been considered as the major cause of occurrence of aviation accidents. Different models have been formed for analysing and investigating the human factors responsible for causing aviation accidents so that the number of

[1] https://www.caas.gov.sg/public-passengers/learn-about-singapore-aviation.

[2] https://edition.cnn.com/travel/article/asia-busiest-airports-2019/index.html.

[3] https://www.iata.org/en/iata-repository/publications/economic-reports/singapore-value-of-aviation/.

such incidents are reduced and prevented (Shappell et al. 2007). Different surveys have been conducted on aviation accidents and incidents for the awareness of air transportation safety (Helmreich and Merritt 2017). Despite progress and awareness concerning these accidents, the cases of air accidents have not reduced significantly.

Aviation accidents are considered as the events occurring due to non-functionality of the operations in the aircraft from the time of boarding of the passengers till the landing of the flights where all the passengers have disembarked (Abeyratne 2012). It also constitutes any operational defect during this time that involves fatal or heavy injury of individuals, damage or loss of the aircraft. Aviation incidents involve the fluctuations in the operational activities of the aircraft that could affect the safety of the passengers. Aviation incidents are mostly represented in the form of faulty actions of the individuals or inappropriate operations related to airborne and ground-based equipment that supports the consequences of the air operations that results in the termination of flight and non-performance of emergency protocols (Clothier and Walker 2015). The major causes of aviation accidents and incidents involve various factors. The major factors involve human errors along with some technical and meteorological factors. The outcomes of poor adherence towards the organization of legislative procedures and the mismanagement of air operations have been the major cause for the occurrence of aviation accidents.

3.2 Human Errors and Unsafe Acts in Aviation Accidents

Human errors and unsafe activities have been considered as the major issues that contribute towards aviation accidents and incidents (Chen and Vincent 2018). The causes of human errors include the errors made by pilots, maintenance engineers, air traffic controllers and related professionals who have a direct impact on flight safety. The failure towards maintaining the operational and organizational activities by the related professionals in aviation contributes towards the occurrence of aviation accidents and incidents by the humans (Reason and Hobbs 2017). Adverse weather conditions contribute towards air accidents in many different ways. The weather conditions are not always found to be appropriate but the information given to the flight crews during the take-off and landing cannot change accordingly (Ji et al. 2011). The decisions of the flight crew are completely dependent upon the information being available to them. No practical need has been made mandatory for the separation of environment from operational factors. The inadequate information provided to the crew members of the flight leads to the lack of critical data collection which in turn leads inappropriate decisions that cause air accidents. In some cases, it has been analysed that the crew members of particular flights have necessary data that are required but often these data are misinterpreted by them (Kelly and Efthymiou 2019). In some cases, it has been analysed that the crew members of particular flights have necessary data that are required but often these data are misinterpreted by them (Oliver et al. 2019). Sometimes even though the flight crew have all the necessary data and are able to interpret the data appropriately, their insufficient skills training and related protocols for taking appropriate decisions to be carried out at particular time is missed upon. This leads to a lack of situational awareness that in turn causes aviation accidents and incidents (Flin and Maran 2015). Increased workload of the crew members has caused deterioration in the operational activities of the flight.

Unsafe acts in the aviation industry have often led towards the occurrence of accidents and incidents (Wiegmann et al. 2005). Unsafe acts have been categorized as errors and violations. Errors are unintentional behaviours and violations are disagreements with rules and regulations (Chen et al. 2017). Skill based errors occur during operational execution that involves practice of the tasks concerning the protocol, training and any alteration in such operational activities results in unsafe situations. Decision errors are another type of unsafe acts which involves the behavioural and actional changes in the operational proceedings which evolves inappropriate decision making that results into unsafe situations (Wagener and Ison 2014). Perceptual errors occur when a person's perception of the situation differs from reality because of faulty information gained from the other sensory units. The unsafe acts in the aviation industry are also brought about by environmental factors, operational conditions and personnel factors (Reason 2008). The physical environment consisting of weather conditions and the technological environment consisting of variations in the designs and automation issues resulting into inappropriate maintenance of the aircraft are also causal factors (Erjavac et al. 2018). The mental state of the flight crew operating the aircraft and their physiological conditions also affects the performance on the flight. Personnel factors involve the management of the crew members of the flight involving their communications, planning, coordination and teamwork issues (Shappell et al. 2007). The inappropriate supervision of the flight crew with the lack of guidance, training, and leadership results in unsafe activities that may cause the air incidents and occurrences (Ancel and Shih 2012). Organizational influences due to the inappropriate organizational climate, lack of operational process and lack of resource management have also contributed to the unsafe acts (Li and Harris 2006).

3.3 Human Factors Analysis and Classification System (HFACS)

The Human Factors Analysis and Classification System (HFACS) has been considered as a systematic framework for analysing accident analysis. It involves the analysis of the various accidents and incidents taking place, mainly due to the human errors. This framework consists of analysing data related to the accidents in the well-structured form. The Human Factors Analysis and Classification System (HFACS) framework established by Wiegmann and Shappell (2001a, b, c) has proven to be a reliable tool in accommodating human causal factors associated with aircraft accidents. Based upon Reason's (1990) model of latent and active failures, HFACS addresses human error at all levels of the aviation system, including the condition of aircrew and organizational factors (Wiegmann and Shappell 2001a, b, c). There are four levels of HFACS. These are: level 1 (unsafe act of pilots or operators), level 2 (preconditions for unsafe acts), level 3 (unsafe supervision) and level 4 (organizational influences). Level 1 (unsafe acts of operators - active failures) is the level at which the majority of accident investigations have been focused in the past. These are the behaviours of the flight crew on the flight deck that contribute directly to the accident. Level 2 (preconditions for unsafe acts - latent/active failures) addresses the psychological precursors to the active failures at level 1. Level 3 (unsafe supervision - latent failures) traces the causal chain of events producing the unsafe acts up to the level of the front-line supervisors. Level 4 (organizational influences - latent failures) describes the contributions of fallible decisions in upper levels of management that directly affect supervisory practices, as well as the conditions and actions of front-line operators (Harris et al. 2007). Hence, the researchers are able to obtain the data related to the accidents from this framework for efficient analysis. HFACS was developed for the US naval aviation for analysing the aspects of human factors in accidents. It is basically known as a generic human error-coding framework. The development of HFACS has been implicated in several papers and books (Weigmann and Shappell 1997, 2001a, b, c, 2016; Shappell and Wiegmann 2001, 2003, 2004). This framework is based and derived from Reason's model of accident causation. Reason's model involves the active and latent failures that combine together along with other factors for coping up with the defences of the system (Reason 1990). The active failures involve the errors related to the performance of operators in the complex system and the latent failures involve the distal errors and system misspecifications, which lie dormant within the system for a long time. The fact was revealed that the complex system types are managed and maintained by human beings and therefore the human decisions and actions need to be implicated in all accidents. The latent failures are spawned in the upper levels of the organization and are related to its management and regulatory structures. Reason's model has been known to be influential in terms of human errors in aviation accidents. This model was not capable of providing remedial solutions towards aviation related accidents. Hence, based on this model, the HFACS framework has been developed in order to fulfil such types of needs (Shappell and Wiegmann 2003). The HFACS framework has shown efficacy in the analysis of accidents in the US commercial aviation, US general aviation and Australian general aviation (Wiegmann and Shappell 2001a, b; Shappell et al. 2007; Shappell and Wiegmann 2003, 2004; Lenné et al. 2008).

HFACS has been considered important for analysing the causes of accidents and is efficient in the identification of the related risk factors of the accidents. Figure 1 represents an overview of the HFACS. This framework has been proved to be an effective tool in identification of human errors in various domains such as railways, mining, etc. (Baysari et al. 2009; Patterson and Shappell 2010; Chauvin et al. 2013). This framework has been found to be efficient in analyzing the General Aviation accident data by the insurance companies (Lemeé 2006). It has also been used in the process of the prospective assessment concerning the effectiveness of aviation safety products that has been developed by NASA aviation safety programmes (Andres et al. 2005; Lechner and Luxhoj 2005; Luxhoj and Hadjimichael 2006).

3.4 Operational Safety by Utilizing HFACS Framework in Aviation

Human errors have been considered as one of the most frequent threats towards safety of lives in aviation (Harris and Li 2010). These errors have been responsible for the creation of failure in the complex systems which have been managed, operated, designed and maintained by the humans (Plant and Stanton 2012). Hence, the decision-making of human beings and their actions at the organizational level are implicated in all types of accidents (Reason 1997). The Human Factors Analysis and Classification System (HFACS) has helped in identifying the human errors for maintaining the operational safety in the aircraft and preventing air accidents and incidents. Many studies have been conducted using the HFACS framework for maintaining the operational safety in aviation. One of the studies involving aircraft registered in Taiwan has been analysed for

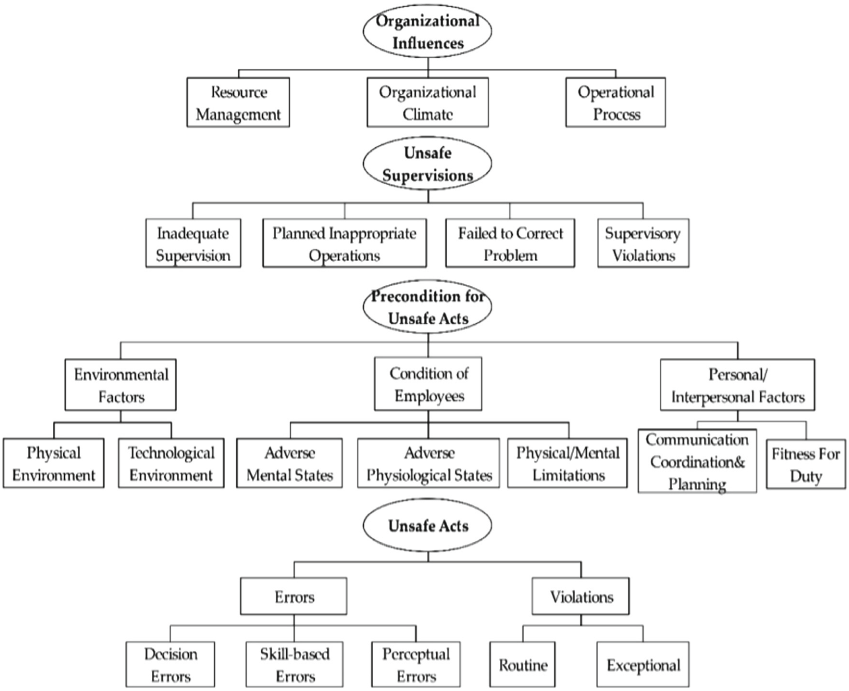


Fig. 1. Overview of human factors analysis and classification system (HFACS)

41 aviation accidents. This study revealed that errors at operational levels are associated with organizational inadequacies. Hence, the findings from this study have evolved the direction for utilizing human error intervention strategy involving remedial safety actions which are aimed at high organizational areas (Li and Harris 2013). Also, the framework has been efficient in analysing the influential factors of violation behaviour that revealed that the attitude of the management may influence the operator's attitude along with their groups, norms, work pressure and behaviour (Fogarty and Shaw 2010).

Safety has been the major priority in the aviation industry and hence safety management systems have been developed by the commercial airlines for reducing the occurrence of hazardous factors in aircraft operations (Liou et al. 2008). Various applications of HFACS in aviation have been studied for identifying the human errors and reducing the aviation accidents. The utilization of HFACS has helped in analysing the human factors caused in accidents involving remotely piloted aircrafts too (Tvaryanas et al. 2006). Maintenances error in the aircraft have been analysed by the adaptation of the HFACS-ME (Krulak 2004). Another study has shown the utilization of the HFACS framework in identifying the human factors in the Asiana Airlines flight 214 accident that occurred on July 6, 2013 (Small 2020). The result of this study has revealed that inappropriate training of the pilots, lack of supervision and deviations from the standard operating procedures (SOPs) were the major issues that contribute towards air accidents. Hence, the HFACS framework helps in identifying errors and implicates the practice of safety measures for reducing the impact of such accidents. The HFACS framework has also shown efficiency in analysing the human errors for Airport surface deviations and runway incursions. For example, the loss of situational awareness has deviated the pilots and air traffic controller's attention leading towards runway incursions (Torres et al. 2011). These situations have been mitigated by training the pilots and the related professionals so that awareness and attentiveness is maintained. Hence, the operational activities need to be strictly maintained for reducing the aviation accidents and maintaining the safety of the aircraft. Tools and techniques have been evolved for mitigating such impact of air accidents so that prior symptoms of such accidents are addressed and immediate solutions are implicated concerning the same.

4 Research Methodology

4.1 Description of the Study Area

The study encompasses civil aircraft flights to and from Singapore's two commercial airports – Singapore Changi Airport and Seletar Airport. As a major international air hub, Changi Airport in 2019 handled about 382,000 commercial aircraft movements comprising both passenger flights and air freights[4]. The flights to and from Changi Airport are operated by a myriad of Full-Service Carriers (FSCs) as well as Low Cost Carriers (LCCs).

Seletar Airport is primarily a general aviation airport although Firefly Airlines, an LCC based in Malaysia, operates its ATR72 aircraft there as well. Other than the Firefly Airlines flights, Seletar Airport also oversees chartered flights, private aircraft and training flights. There are three flight schools in Seletar airport that operate general aviation aircraft such as the DA40 and Cessna 172.

A comprehensive review of all 75 Air Safety Investigation Reports between October 2000 and September 2019 was carried using database records maintained by the Air Accident Investigation Bureau of the Ministry of Transport in Singapore. Of particular interest in this study were those accidents and incidents attributable, at least in part to the aircrew operating the flights. Accidents due solely to catastrophic failure, maintenance errors and unavoidable weather conditions such as turbulence and wind shear were not included. Furthermore, only those accidents and incidents in which the investigation was completed, and the cause of the occurrence determined were included in this analysis. On the whole 48 reports related to aircrew related human factors occurrences met these criteria. The type of aircraft operations involved in this study were commercial passenger aircraft, commercial cargo and general aviation. The total number of accidents analyzed were 22 and the number of incidents were 53.

4.2 Research Design

In accordance with the research query and objectives, the study aims at addressing, a descriptive research design**.** Descriptive research design facilitates explaining the events as they independently occur in nature without any interference or manipulation of data from the researcher. Additionally, it will help in justifying the results of the study in a strategic manner.

[4] https://www.changiairport.com/corporate/our-expertise/air-hub/traffic-statistics.html.

4.3 Research Paradigm

The current study will utilize an interpretivism research paradigm as it aims for assessing the utility of the HFACS framework as an error analysis and classification tool for accidents/incidents in the aviation industry in Singapore. This research paradigm has been implemented when the objectives are directed towards outlining issues and measures of social issues of aviation accidents and incidents which has been found to be prevalent due to human errors.

4.4 Research Approach

The present research study aims to assess the utility of the HFACS framework as an error analysis and classification tool for accidents/incidents in the aviation industry in Singapore. The study adopts a qualitative research approach since the researcher collects data through a comprehensive review of accidents and incidents that occurred in Singapore. The HFACS framework has been applied to commercial aviation accident and incident records maintained by the Transportation Safety Investigation Bureau (TSIB) of the Ministry of Transport (MOT). The TSIB is the air, marine and rail accidents and incidents investigation authority in Singapore which is further used in this study to achieve the aims and objective of the paper. Its mission is to promote transport safety through the conduct of independent investigations into air, marine and rail accidents and incidents. Therefore, the current study utilizes a quantitative research approach. This can be attributed to the fact that interpretivism philosophy is most appropriate to implement with a quantitative approach. Moreover, it has helped in accumulating objective data through a wider sample size.

4.5 Data Analysis

The secondary data involves the comprehensive review of accidents and incidents that occurred in Singapore that needs to be analysed. This collected data is being analysed within the HFACS context. Thus, the utility of the HFACS framework has been appraised for reviewing such accidents and more specific detail relating to human error can be reported accordingly. In a later study these analyses will be compared with data from countries in Asia where similar databases and accident investigation reports are available.

5 Results and Discussion

The HFACS framework bridges the gap between theory and practice by providing safety professionals with a theoretical tool for identifying and classifying human errors in aviation mishaps (Shappell and Wiegmann 2001). The HFACS framework can be used to reliably identify underlying human factors problems associated with the 75 accidents and incidents recorded in Singapore. The HFACS framework has addressed four major types of human errors. Level 1 consists of unsafe acts of operators. Errors and violations are classified at Level 1. Level 2 consists of preconditions for unsafe acts involving latent and active failures. This level has described substandard conditions of operators and their substandard practices. Level 3 consists of unsafe supervision with latent failures traced in the causal chain of events evolving unsafe acts up to the level of the front-line supervisors. Level 4 consists of an organizational influence that describes the decision-making in upper levels of management while supervising the practices.

Although there were 48 flight occurrences associated with aircrew human factors, a total of 76 causal factors were observed and analysed using the HFACS framework. Instances of all but two HFACS categories (i.e. adverse physiological states and personnel readiness) were observed at least once in the database. Refer to Table 1 and Fig. 2 below for a summary of the HFACS factors observed.

Table 1. Summary of observed HFACS factors

HFACS factors	Total numbers	Percentage
Organizational Influences: Resource Management	4	5.3
Organizational Influences: Organizational Climate	2	2.6
Organizational Influences: Organizational Process	6	7.9
Unsafe Supervision: Inadequate Supervision	5	6.6
Unsafe Supervision: Planned Inappropriate Operations	1	1.3
Unsafe Supervision: Failed to Correct Problem	1	1.3
Unsafe Supervision: Supervisory Violation	1	1.3
Precondition for Unsafe Acts: Adverse Mental States	1	1.3
Precondition for Unsafe Acts: Adverse Physiological States	0	0.0
Precondition for Unsafe Acts: Physical/Mental Limitations	1	1.3
Precondition for Unsafe Acts: Crew Resource Mismanagement	7	9.2
Precondition for Unsafe Acts: Personal Readiness	0	0.0
Unsafe Acts: Decision Errors	13	17.1
Unsafe Acts: Skill-based Errors	20	26.3
Unsafe Acts: Perceptual Errors	8	10.5
Unsafe Acts: Routine Violations	1	1.3
Unsafe Acts: Exceptional Violations	5	6.6
Total	**76**	**100**

At the unsafe act level, skill-based errors were associated with the largest percentage of occurrences. Approximately 26% of all aircrew related accidents were associated with at least one skill-based error. The proportion of incidents and accidents has remained relatively unchanged over a 19-year period in the study. Notably the highest proportion of accidents and incidents associated with skill-based errors occurred in the year 2013. Among the remaining categories of unsafe acts, accidents and incidents associated with decision-based errors contributed the next highest proportion (i.e. 17% of the accidents

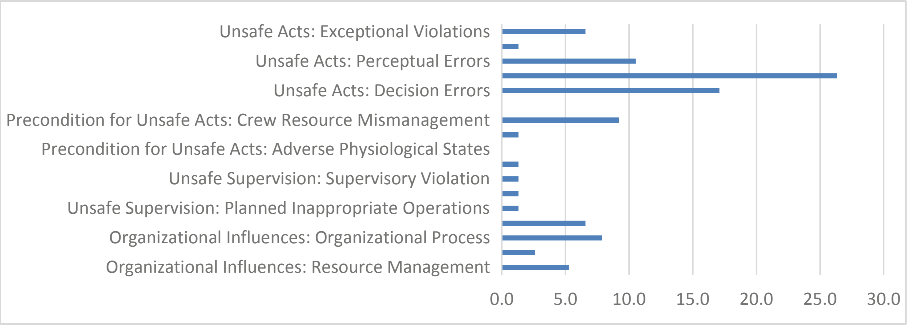


Fig. 2. Comparison of observed HFACS factors in percentage

and incidents examined). The proportion of incidents/accidents associated with decision-based errors remained relatively constant throughout the years of study. Perceptual errors contributed about 10% of the occurrences examined. The number of perceptual errors remained relatively constant throughout the period of study. Occurrences attributable to violations of rules and regulations numbered about 8%.

Within the pre-conditions level, Crew Resource Management (CRM) failures were associated with the largest percentage of occurrences. Approximately 11% of all aircrew related incidents and accidents were associated with at least one CRM failure. However, the percentage of occurrences associated CRM failures remained relatively constant over the 19-year period of the study. There was one incident each of adverse mental state and physical/mental limitation. There were no occurrences associated with personal readiness issues or adverse physiological states.

Compared to the category of unsafe acts, the number of AAIB reports that implicated the aircrew as contributing to an accident or an incident citing some form of supervisor and organizational factor was comparatively smaller. There were 10% of the occurrences that involved some form of supervisory influence and about 16% of occurrences that implicated organizational factors.

A trend analysis was done to compare the number of human factors related accidents and incidents, and the overall aircraft movements in Singapore from 2000 to 2019. Refer to Fig. 3 below. Since 2000, the human factors related occurrences are on an overall rising trend although the numbers have remained fairly constant since 2013. Incidentally, 2013 also registered the peak in the number of occurrences. The overall increase in accidents and incidents can be attributed to the rise in aircraft movements in Singapore, which saw a 120% jump from 2000 to 2019.

In summary, the HFACS framework was found to accommodate all 76 causal factors associated with 48 accidents in which pilots were involved across a 19-year period in the Singapore aviation industry. The error categories that were originally developed for military aviation in the United States are applicable within the commercial aviation space in Singapore. There were two errors factors within the HFACS framework that were never observed in the Singapore AAIB data base. For example, there were no instances in which crew personal readiness and adverse physiological states were causal factors. Unsafe acts in total contributed approximately 62% of occurrences related to

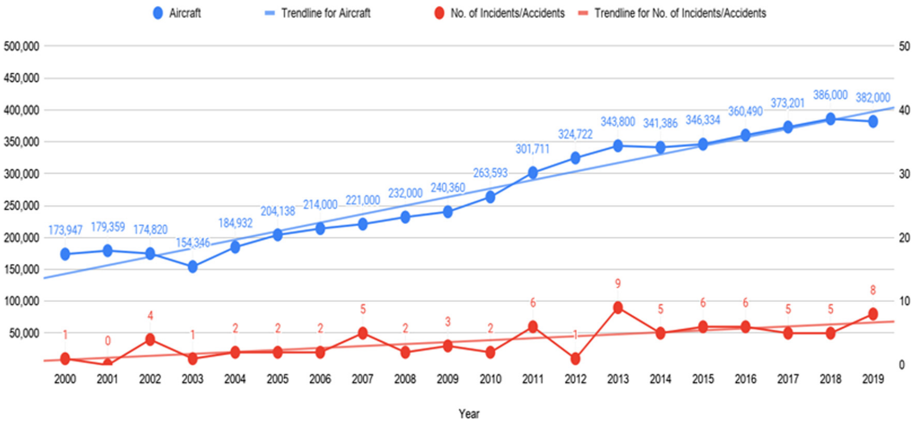


Fig. 3. Comparison of human factors related accidents and incidents, and aircraft movements: 2000 to 2019.

flight operations. The next highest category was organizational influences which was 16%, followed by pre-conditions for unsafe acts at 12%. Finally, the proportion of accidents and incidents associated by unsafe supervisions was 10%.

HFACS was utilized in analysing accident databases done other countries, like India, USA, Africa and Australia (Gaur 2005, Munene 2016, Inglis and McRandle 2007) and it was shown to accommodate all human factors that contributed to aircraft accidents.

6 Significance of the Study

The present study has analysed the identification of human errors which is responsible for causing aviation accidents and incidents in Singapore. The present study has utilized the HFACS framework in identifying the human errors and this could help in prevention of any incidents and accidents. Using this framework aviation organizations will be capable of identifying the weak links which are the causes for the occurrence of aviation accidents. The study has shown efficacy in identifying human performance and reducing the system deficiencies for maintaining safety measures in the aviation industry. The HFACS framework helps in analysing the weak areas that will help in reducing the accident and injury rates. Also, this study has provided a direction towards analysing and reviewing historical accidents and adoption of safety measures for avoiding the possibilities of such accidents. Hence, this framework has been used as an important assessment tool towards the guidance of future investigations and helps in improving the quality and accessibility of human factors that are prone for the occurrence of accidents and incidents. The significance of the study will be increased by comparisons with aircraft accident and incident data in other countries. Comparison of the results of this study against another country's similar study will assist in the interpretation of types of causal factors leading to incidents and accidents in Singapore. It will help clarify strengths and weaknesses in the Singapore aviation system.

7 Conclusion and Recommendation

The present study has aimed at assessing the utility of the HFACS framework as an error analysis and classification tool for accidents/incidents in the aviation industry in Singapore. The analysed accidents and incidents that occurred between October 2000 and December 2019 in Singapore have been undertaken with more specific detail relating to human errors. The different categories of unsafe acts in the HFACS framework have helped in analysing the identification of these errors which can be further used for preventing other possible air accidents and incidents. This framework can be used in other data sets as well as for establishing the relationship of errors with the accidents and incidents in supporting the HFACS framework to be an effective tool in analysing human errors. This study provides an understanding, based on the evidence of how actions and decisions at different levels in the organization facilitates aircraft operations resulting in operational errors and accidents. The present study can also contribute in building up the comprehensive database in Southeast Asia. This study can offer new insights into the nature and trends of human factors in aircraft accidents and incidents.

References

Chang, Y.H., Yeh, C.H.: Human performance interfaces in air traffic control. Appl. Ergon. **41**(1), 123–129 (2010)

Isaac, A.R., Ruitenberg, B.: Air Traffic Control: Human Performance Factors. Routledge, London (2017)

Zhou, T., Zhang, J., Baasansuren, D.: A hybrid HFACS-BN model for analysis of Mongolian aviation professionals' awareness of human factors related to aviation safety. Sustainability **10**(12), 4522 (2018)

Wiegmann, D.A., Shappell, S.A.: A human error approach to aviation accident analysis: the human factors analysis and classification system. Collegiate Aviat. Rev. **34**(2), 102 (2016)

Shorrock, S.T., Kirwan, B.: Development and application of a human error identification tool for air traffic control. Appl. Ergon. **33**(4), 319–336 (2002)

Gong, L., Zhang, S., Tang, P., Lu, Y.: An integrated graphic–taxonomic–associative approach to analyze human factors in aviation accidents. Chin. J. Aeronaut. **27**(2), 226–240 (2014)

Salmon, P.M., Cornelissen, M., Trotter, M.J.: Systems-based accident analysis methods: a comparison of Accimap, HFACS, and STAMP. Saf. Sci. **50**(4), 1158–1170 (2012)

Martins, I.T., Martins, E.T., Soares, M.M., da Silva Augusto, L.G.: Human error in aviation: the behavior of pilots facing the modern technology. In: Marcus, A. (ed.) Design, User Experience, and Usability, Part III, pp. 150–159. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39238-2_17

Chen, W., Huang, S.: Evaluating flight crew performance by a bayesian network model. Entropy **20**(3), 178 (2018)

Wiegmann, D.A., Shappell, S.A.: Human error analysis of commercial aviation accidents using the human factors analysis and classification system (HFACS) (No. DOT/FAA/AM-01/3,). United States. Office of Aviation Medicine (2001a)

Shappell, S.A., Wiegmann, D.A.: Reshaping the way we look at general aviation accidents using the human factors analysis and classification system (2003)

Reinach, S., Viale, A.: Application of a human error framework to conduct train accident/incident investigations. Accid. Anal. Prev. **38**(2), 396–406 (2006)

Baysari, M.T., Caponecchia, C., McIntosh, A.S., Wilson, J.R.: Classification of errors contributing to rail incidents and accidents: a comparison of two human error identification techniques. Saf. Sci. **47**(7), 948–957 (2009)

Vairo, T., Quagliati, M., Del Giudice, T., Barbucci, A., Fabiano, B.: From land-to water-use-planning: a consequence-based case-study related to cruise ship risk. Saf. Sci. **97**, 120–133 (2017)

Shappell, S.A., Wiegmann, D.A.: The human factors analysis and classification system–HFACS (2000)

Weigmann, D.A., Shappell, S.A.: Human factors analysis of postaccident data: applying theoretical taxonomies of human error. Int. J. Aviat. Psychol. **7**(1), 67–81 (1997)

Wiegmann, D.A., Shappell, S.A.: Applying the human factors analysis and classification system (HFACS) to the analysis of commercial aviation accident data (2001b)

Wiegmann, D.A., Shappell, S.A.: Human error perspectives in aviation. Int. J. Aviat. Psychol. **11**(4), 341–357 (2001c)

Shappell, S.A., Wiegmann, D.A.: Applying reason: the human factors analysis and classification system (HFACS). Human Factors and Aerospace Safety (2001)

Shappell, S.A., Wiegmann, D.A.: A human error analysis of general aviation-controlled flight into terrain accidents occurring between 1990–1998. Federal Aviation Administration Oklahoma City Ok Civil Aeromedical Inst (1998)

Shappell, S.A., Wiegmann, D.A.: HFACS analysis of military and civilian aviation accidents: a North American comparison. In: Proceedings of the Annual Meeting of the International Society of Air Safety Investigators, Gold Coast, Australia, November 2004

Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.A.: Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Hum. Factors **49**(2), 227–242 (2007)

Lenné, M.G., Ashby, K., Fitzharris, M.: Analysis of general aviation crashes in Australia using the human factors analysis and classification system. Int. J. Aviat. Psychol. **18**(4), 340–352 (2008)

Patterson, J.M., Shappell, S.A.: Operator error and system deficiencies: analysis of 508 mining incidents and accidents from Queensland, Australia using HFACS. Accid. Anal. Prev. **42**(4), 1379–1385 (2010)

Chauvin, C., Lardjane, S., Morel, G., Clostermann, J.P., Langard, B.: Human and organisational factors in maritime accidents: analysis of collisions at sea using the HFACS. Accid. Anal. Prev. **59**, 26–37 (2013)

Lechner, K.W., Luxhoj, J.T.: Probabilistic causal modelling of risk factors contributing to runway collisions: case studies. Hum. Factors Aerosp. Saf. **5**(3) (2005)

Luxhoj, J.T., Hadjimichael, M.: A hybrid fuzzy-belief network (HFBN) for modelling aviation safety risk factors. Hum. Factors Aerosp. Saf. **6**(3) (2006)

Andres, D.M., Luxhoj, J.T., Coit, D.W.: Modelling of human-system risk and safety: aviation case studies as exemplars. Hum. Factors Aerosp. Saf. **5**(5) (2005)

Harris, D., Li, W.C.: An extension of the human factors analysis and classification system for use in open systems. Theor. Issues Ergon. Sci. **12**(2), 108–128 (2011)

Plant, K.L., Stanton, N.A.: Why did the pilots shut down the wrong engine? Explaining errors in context using schema theory and the perceptual cycle model. Saf. Sci. **50**(2), 300–315 (2012)

Li, W.C., Harris, D.: Identifying training deficiencies in military pilots by applying the human factors analysis and classification system. Int. J. Occup. Saf. Ergon. **19**(1), 3–18 (2013)

Fogarty, G.J., Shaw, A.: Safety climate and the theory of planned behavior: towards the prediction of unsafe behavior. Accid. Anal. Prev. **42**(5), 1455–1459 (2010)

Tvaryanas, A.P., Thompson, W.T., Constable, S.H.: Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. Aviat. Space Environ. Med. **77**(7), 724–732 (2006)

Krulak, D.C.: Human factors in maintenance: impact on aircraft mishap frequency and severity. Aviat. Space Environ. Med. **75**(5), 429–432 (2004)

Small, A.: Human factors analysis and classification system (HFACS): as applied to Asiana airlines flight 214. J. Purdue Undergraduate Res. **10**(1), 18 (2020)

Jupp, J.A.: 21st century challenges for the design of passenger aircraft. In: CD Proceedings: ICAS 2012-28th Congress of the International Council of the Aeronautical Sciences (2012)

Helmreich, R.L., Merritt, A.C.: 11 Safety and error management: the role of crew resource management. In: Aviation Resource Management: Proceedings of the Fourth Australian Aviation Psychology Symposium, vol. 1. Routledge, November 2017

Marais, K.B., Robichaud, M.R.: Analysis of trends in aviation maintenance risk: an empirical approach. Reliab. Eng. Syst. Saf. **106**, 104–118 (2012)

Kharoufah, H., Murray, J., Baxter, G., Wild, G.: A review of human factors causations in commercial air transport accidents and incidents: from to 2000–2016. Prog. Aerosp. Sci. **99**, 1–13 (2018)

Clothier, R.A., Walker, R.A.: The safety risk management of unmanned aircraft systems. In: Handbook of Unmanned Aerial Vehicles, pp. 2229–2275 (2015)

Wiegmann, D.A., Shappell, S.A.: A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System. Routledge (2017)

Chen, J.C., Vincent, F.Y.: Relationship between human error intervention strategies and unsafe acts: the role of strategy implementability. J. Air Transp. Manag. **69**, 112–122 (2018)

Reason, J., Hobbs, A.: Managing maintenance error: a practical guide (2017)

Flin, R., Maran, N.: Basic concepts for crew resource management and non-technical skills. Best Pract. Res. Clin. Anaesthesiol. **29**(1), 27–39 (2015)

Wiegmann, D., Faaborg, T., Boquet, A., Detwiler, C., Holcomb, K., Shappell, S.: Human error and general aviation accidents: a comprehensive, fine-grained analysis using HFACS. Federal Aviation Administration Oklahoma City Ok Civil Aeromedical Inst. (2005)

Yeun, R., Bates, P., Murray, P.: Aviation safety management systems. World Rev. Intermodal Transp. Res. **5**(2), 168–196 (2014)

Li, C.W., Phun, V.K., Suzuki, M., Yai, T.: The effects of aviation accidents on public perception toward an airline. J. Eastern Asia Soc. Transp. Stud. **11**, 2347–2362 (2015)

Vatankhah, S.: Dose safety motivation mediate the effect of psychological contract of safety on flight attendants' safety performance outcomes?: a social exchange perspective. J. Air Transp. Manag. **90**, 101945 (2021)

Abeyratne, R.: Aircraft accident and incident investigation. In: Abeyratne, R. (ed.) Air Navigation Law, pp. 105–111. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-25835-0_5

Ji, M., You, X., Lan, J., Yang, S.: The impact of risk tolerance, risk perception and hazardous attitude on safety operation among airline pilots in China. Saf. Sci. **49**(10), 1412–1420 (2011)

Kelly, D., Efthymiou, M.: An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. J. Saf. Res. **69**, 155–165 (2019)

Oliver, N., Calvard, T., Potočnik, K.: Safe limits, mindful organizing and loss of control in commercial aviation. Saf. Sci. **120**, 772–780 (2019)

Chen, J.C., Lin, S.C., Vincent, F.Y.: Structuring an effective human error intervention strategy selection model for commercial aviation. J. Air Transp. Manag. **60**, 65–75 (2017)

Wagener, F., Ison, D.C.: Crew resource management application in commercial aviation. J. Aviat. Technol. Eng. **3**(2), 2 (2014)

Reason, J.T.: The Human Contribution: Unsafe Acts, Accidents and Heroic Recoveries. Ashgate Publishing Ltd. (2008)

Erjavac, A.J., Iammartino, R., Fossaceca, J.M.: Evaluation of preconditions affecting symptomatic human error in general aviation and air carrier aviation accidents. Reliab. Eng. Syst. Saf. **178**, 156–163 (2018)

Ancel, E., Shih, A.: The analysis of the contribution of human factors to the in-flight loss of control accidents. In: 12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, p. 5548, September 2012

Li, W.C., Harris, D.: Pilot error and its relationship with higher organizational levels: HFACS analysis of 523 accidents. Aviat. Space Environ. Med. **77**(10), 1056–1061 (2006)

Liou, J.J., Yen, L., Tzeng, G.H.: Building an effective safety management system for airlines. J. Air Transp. Manag. **14**(1), 20–26 (2008)

Gaur, D.: Human factors analysis and classification system applied to civil aircraft accidents in India. Aviat. Space Environ. Med. **76**, 501–505 (2005)

Inglis, M.S.J., McRandle, B.: Human factors analysis of Australian aviation accidents and comparison with the United States. Australian Transport Safety Bureau (2007). https://www.atsb.gov.au/media/29953/b20040321.pdf

Munene, I.: An application of the HFACS method to aviation accidents in Africa. Aviat. Psychol. Appl. Hum. Factors **6**(1), 33–38 (2016). https://doi.org/10.1027/2192-0923/a000093

Preliminary Analysis of Human Error Prediction Model by Using Biological Information

Yuto Saito, Muhammad Nur Adilin Mohd Anuardi(✉), Ryota Matsubara, and Midori Sugaya

Shibaura Institute of Technology, 3-7-5 Toyosu, Koto, Tokyo 135-8548, Japan
{al17048,nb17106,doly}@shibaura-it.ac.jp,
ryota@sic.shibaura-it.ac.jp

Abstract. Increasing in aging population forced the society to act more than their limit. For instance, an action such as driving, where we need our mental concentration at most, could lead to serious accident from a simple mistake because of overwork. Therefore, it is crucial to prevent the accident. Many researchers focus on biological information to predict the error because human error always related to a person's cognitive condition such as stress and discomfort. However, existing studies on the human error prediction model have not conducted a detailed analysis, and also have not considered individual differences. Therefore, the purpose of this study is to analyze the biological information immediately before and after the occurrence of human error in order to construct a prediction model for human error considering individual differences. In this study, we developed the Stroop task to be used as the mental workload and measured the subjects' biological information. As a result, we proposed 10 [s] as the time intervals for before and after the consecutive of the occurrence of the human errors for better analysis. Besides, the biological information measured from all subjects suggested that pNN10 can be considered as the predictive indicator for human error occurrence. However, other biological information also expressed vary results where our next step needs to consider the individual differences by increasing the sample size. In addition, the logistic regression will be considered for machine learning to be used for the human error prediction model construction.

Keywords: Human error · Electroencephalography · Heart rate variability

1 Introduction

Aging population especially in Japan has accounted for 28.8% of the nation's total population today [1]. Several problems surfaced as a result of the increase in the aging population such as the decline in the labor force. Due to this, the retirement age has been considered to be increased even though the human's ability is declining with the increase in age. Thus, it can be said that increasing in aging population forced the society to act more than their limit. To work over the limit is the same as taking a risk where it

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 324–335, 2021.
https://doi.org/10.1007/978-3-030-77932-0_26

is possible to expect complication especially in elderly people. For instance, an action such as driving, where we need our concentration at most, could lead to serious accident from a simple mistake. Therefore, it is necessary to prevent accidents due to human error. Human error is a human action that produces unintended consequences, which is an important factor of the accidents.

Many studies focus on the detection of human error in order to avoid failure. However, past research on the human error prediction model has not performed thoroughly. A proper error analysis such as engaging with biological information could reduce the potential for accidents and increase safety. This is because human error always related to a person's cognitive condition such as stress and discomfort.

In recent years, some methods using biological information have been proposed for the analysis of the mental state of human error [2, 3]. Since biological information can acquire mental states that occur unconsciously, it is considered to be suitable for making predictions to prevent serious accidents in situations where highly developed and complicated judgments are required. Mark et al. [2] asked the subjects to perform a military simulator task (a task to discover a suspicious airplane on a radar) while acquiring and observing the changes of 20 types of biological information including electroencephalography (EEG), heart rate variability (HRV), and electrodermal activity (EDA). It was shown that the error occurrence affected the changes in the biological information acquired. However, too many evaluation items were affected by the overall evaluation of the experiment.

Nishikawa et al. and Kishimoto et al. also conducted a human error analysis using biological information such as brain waves and pulse [3, 4]. They proposed on the fact that the incidence of error increases when the work environment is unpleasant. Here, NASA-TLX, which is a subjective workload assessment, was used as an evaluation of the work environment [5]. A pattern to analyze the error occurrence effect on the biological information was created based on the change of before and after the error occurrence. Here, a significant difference was confirmed only in a part of the measured pulse.

Yoshida, on the other hand, reviewed the research by Kishimoto et al. and proposed the Stroop task as a workload for the subjects [6]. Yoshida managed to obtain the prediction model and correlation values for human error that are significant for EEG and pulse rates. However, Yoshida only considered the analysis before and after task, which not contemplate the human error in details. Since the calculated values may differ depending on the process of human error occurrence, it is necessary to consider the evaluation before and after the error occurrence in order to analyze human error and build a robust prediction model.

In addition to Yoshida analysis, it did not consider individual differences. Since biological information differs among individuals, and it depends on the environments, it is necessary to construct a human error prediction model that considers individual differences. Therefore, the purpose of this study is to analyze the biological information immediately before and after the occurrence of human error in order to construct a prediction model for human error considering individual differences.

2 Proposed Method

2.1 Overview

In this study, we consider multiple factors related to human error, which we will measure and analyze the biological information such as brain wave and pulse rate. With the measured and analyzed biological information, we will construct and propose a real-time evaluation system to be used as a human error prediction model that consider individual differences. In order to realize these, the following issues were verified: (a) We verified the time interval for the error occurrence analysis such as before, during, and after the occurrence. This verification will come up with the correlation of the human error rate. (b) We verified the biological information indicators by comparing the integration of EEG and HRV measurement that correlates with the error.

2.2 Evaluation Method

In verifying (a) and (b) described in Sect. 2.1, the following two methods were used. 1) Determine the time intervals to be considered as before and after for the human error occurrence analysis. Here we will find the correlation between the average value of biological information with the rate of human error occurrence. 2) Examine the indicators that will affect the mental workload by considering the items used in the pre-questionnaire before the experiment related to human errors. In this study, following the researches by Yoshida and Kishimoto et al., we used the integration of brain wave and pulse rate measurement. This is because the brain wave and pulse rate vary among individuals and sensitive to environments. Thus, the integration measurement can be considered for detail analysis of the mental state at the time of the human error. Table 1 shows the EEG indicators and psychological state which will be investigated in the study.

In addition, the pNNx of HRV indicator was used in the study which refers to a study by Ikeda et al. [7]. pNNx can measure the fluctuation of the parasympathetic nerve to determine the relaxed state. The pNNx was calculated from the ratio of the absolute values of the adjacent heartbeat intervals which exceeding x [ms].

Table 1. EEG indicators.

Indicator	Frequency band (Hz)	Psychological state [8, 9]
Low α	8–9	Relax, peaceful, conscious
High α	10–12	Relax, good concentration
Low β	13–17	Thinking, accident and environment awareness
High β	18–30	Alertness, upset
Low γ	31–40	Memorization, high mental activity
Mid γ	41–50	Visual information processing

2.3 Stroop Task

Since it is necessary to generate a task that leads to human error, we decided to use the Stroop task in this study, as it was judged that it is appropriate to use it for mental workload by previous studies [10]. Figure 1(a) shows an example of the Stroop task. The subjects will have two versions of the Stroop task: (1) Color-based and (2) Meaning-based. The subjects need to answer based on each version where the two information were interfering with each other. As in Fig. 1(a), the character is “red” in color interferes with the character meaning which is “green”. When answering the color-based Stroop task, the correct answer is “red”, while for the meaning-based Stroop task, the correct answer is “green”. As for the experiment system, the Stroop task is performed by the subjects while having biological information measurements at the same time. JavaScript was used as the development language to build the experiment system.

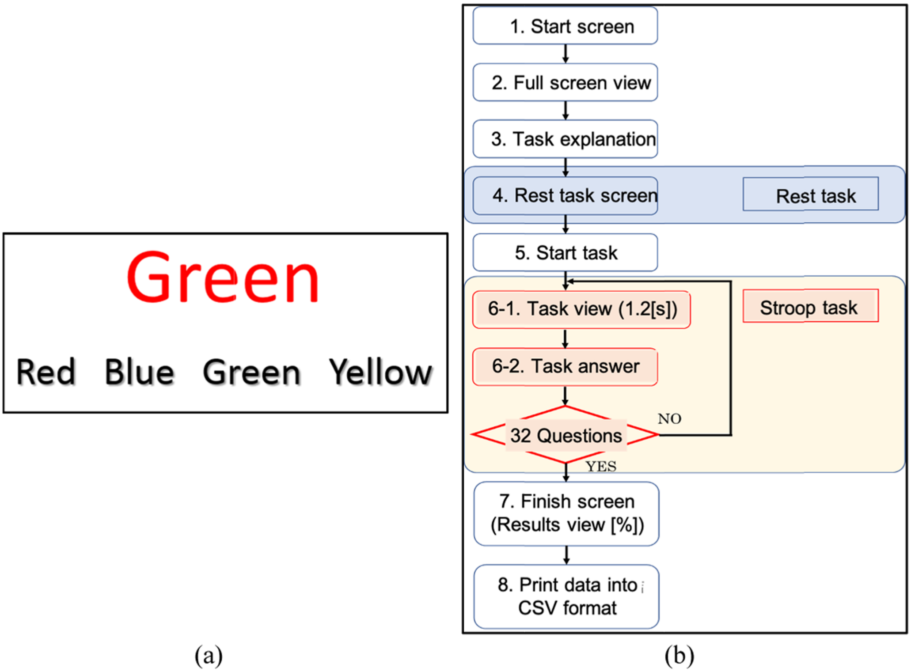


Fig. 1. (a) User Interface (UI) of Stroop task used for experiment. (b) Measurement process. (Color figure online)

2.4 Measurement Process

In this study, the Stroop task used for the mental workload is performed at the same time with the biological information data measurement using the browser. The measurement process was designed as in Fig. 1(b). Figure 1(b) shows the flow for one-time measurement. However, in the experiment, the flow of step 3 to 7 was repeated twice for color-based and meaning-based of the Stroop task. During the measurement, the following data were recorded and saved as comma-separated values (CSV) file: 1) Time when to move to the next step, 2) Time when answering the question, 3) Time taken to answer the question, 4) Answer to the question, 5) Percentage of the correct answer, 6) Average response time for questions with the Stroop effect, 7) Average response time for questions without the Stroop effect.

The Stroop task screen was displayed as in Fig. 1 and the subjects need to choose the answer from four options. After subjects chose the answer within the time given which 1.2 [s], the screen will proceed to the next question. However, if the subjects could not answer within the time given, the screen will automatically proceed to the next question right after 1.2 [s]. In addition, the Stroop task was randomly displayed in order to cancel out the effect of an unobserved factor. The system was developed using the jsPsych library, which is a tool for conducting psychological experiments.

3 Experiment

3.1 Overview

In the preliminary experiment, we measured the subjects 'brain wave and pulse rate during the Stroop task. Four subjects (2 males, 2 females in their 20s) participated in the experiment. There are 32 questions from each Stroop task that need to answer by the subjects. The unanswered and incorrect answers were treated as human errors. The experiment was carried out according to the following procedure.

1. The subjects answer the pre-questionnaire
2. EEG and pulse sensor were set on subjects.
3. Rest for 60 [s] (Rest 1)
4. Meaning-based Stroop task (Task 1)
5. Rest for 60 [s] (Rest 2)
6. Color-based Stroop task (Task 2).

3.2 Results

Pre-questionnaire

Table 2 shows the results of the pre-questionnaire and the rate of human error. A correlation was found between the 'Psychological stress response' and the rate of human error from the pre-questionnaire of the four subjects. In addition, a strong inverse correlation was found in the 'Physical stress response' with the rate of human error. Higher value in Table 2 shows a higher stress level for each item.

Biological Information Measurements. Figure 2, 3, 4, and 5 show the time transition of biological information measured and the human error occurrence of the four subjects. However, due to device malfunction, the pNN10 measurement of the subject A had a defect, which will not include in Fig. 2. In Fig. 2, the biological information measured for subject A before and after the human error occurrence fluctuates greatly especially on β indicator of Task 2.

Table 2. Pre-questionnaire results and human error rate.

Stress item	Subject A (Male)	Subject B (Female)	Subject C (Female)	Subject D (Male)
Work burden	5	6	5	4
Work control	0	0	1	1
Interpersonal relationships at work	0	0	1	1
Work suitability	0	1	1	1
Psychological stress response	17	4	0	3
Physical stress response	2	1	7	6
Workplace support	0	1	6	1
Total stress score	24	13	23	17
Physical condition (1: Good, 5: Bad)	5	2	1	3
Weather (1: Good, 5: Bad)	1	1	1	3
Human error rate	0.5	0.44	0.31	0.31

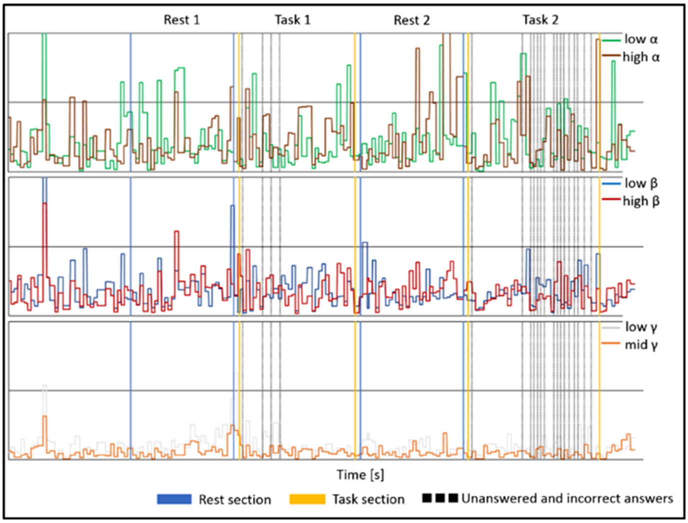


Fig. 2. Biological information measured and human error occurrence of Subject A.

In Fig. 3, it can be seen that pNN10 was lower during task section compared to rest section. As a result of the analysis of variance (ANOVA), a significant difference was found between each rest and task for the pNN10. In addition, it can be seen that pNN10 has a trend where it rises immediately after the human errors occur and falls immediately during the human error occurrence. For the EEG indicators, the α bandwidth, which indicates good concentration, was lower during the task section than the rest section. Meanwhile, the β and γ bandwidth, which indicate thinking and high mental activity respectively, were higher during the task section compared to rest section. As a result of ANOVA, a significant difference was found only between Rest 2 and Task 2 for β and γ bandwidth.

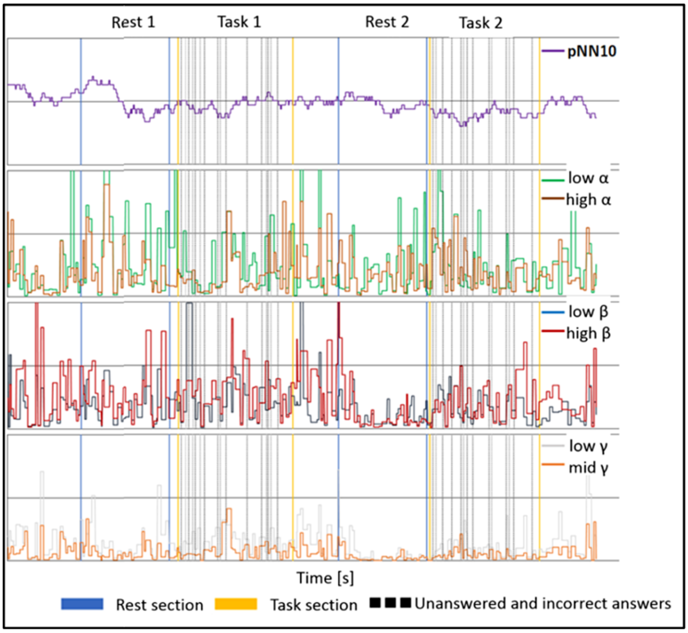


Fig. 3. Biological information measured and human error occurrence of Subject B.

In Fig. 4, it can be seen that pNN10 was lower during task section compared to rest section. As a result of the analysis of variance (ANOVA), a significant difference was found between each rest and task for the pNN10. For the EEG indicators, the β and γ bandwidth, which indicate thinking and high mental activity respectively, were higher during task section compared to rest section. As a result of ANOVA, a significant difference was found only between Rest 2 and Task 2. In Fig. 5, it can be seen that pNN10 was lower during the task section compared to rest section. As a result of the analysis of variance (ANOVA), a significant difference was found between each rest and task for the pNN10. In addition, it can be seen that the fluctuation of EEG is large during Task 1.

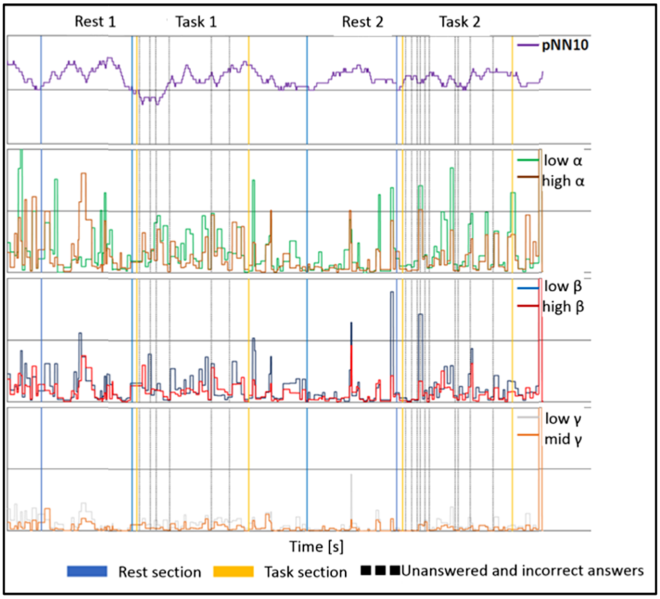


Fig. 4. Biological information measured and human error occurrence of Subject C.

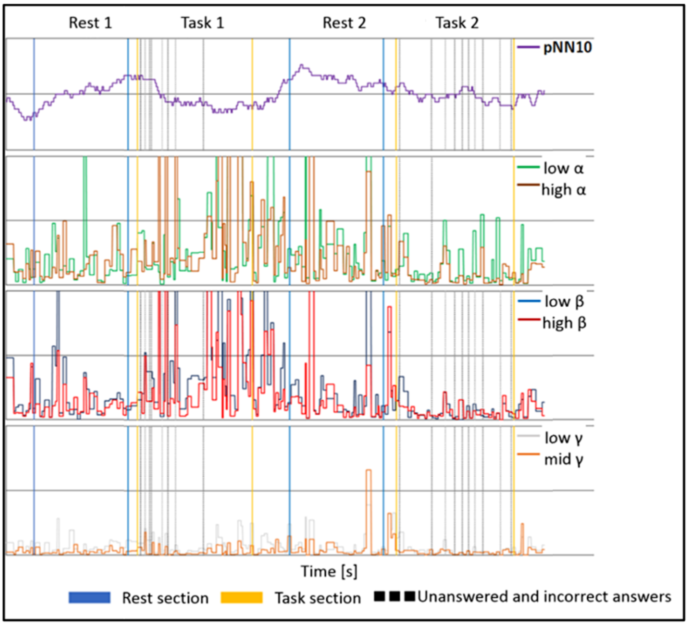


Fig. 5. Biological information measured and human error occurrence of Subject D.

Here, Fig. 6 shows the definition of the time interval before and after human error. 10 [s] were defined as the time for before and after the human error. However, if the error occurs continuously (within 4 [s]), the following two conditions were investigated in order to determine the time for before and after the human error: 1) Assume that human errors occur continuously (the time interval will be 10 [s] before and after of each error), 2) Summarize the continuous part as one human error group (the time interval will be 10 [s] before and after the human error group). These two conditions were compared using the biological information measured from subject B.

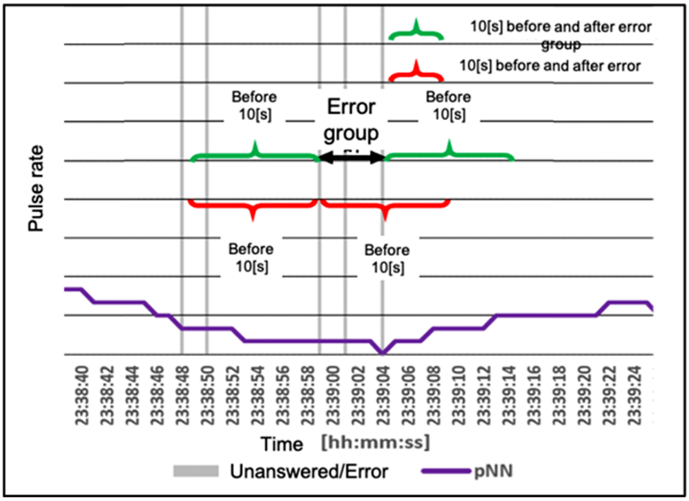


Fig. 6. Definition of time intervals before and after human error.

For the comparison, we calculate the rate of significant difference ($p < 0.05$) from the ANOVA of each biological information indicators between the human error time interval of two conditions with the rest section. Figure 7 shows the comparison of ANOVA significant difference rate of the two conditions. From Fig. 7, it can be seen that the significance rate is higher for most of the biological information in 10 [s] before and after the error group condition. In addition, significant differences were observed between the two conditions for high β and low γ.

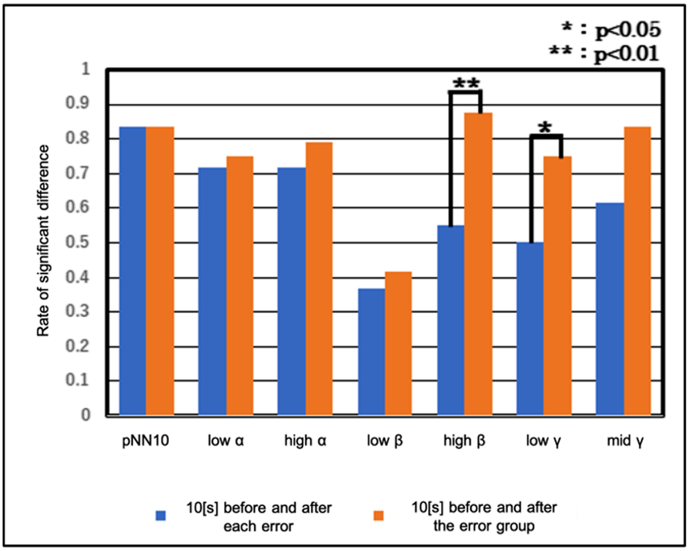


Fig. 7. Comparison of the two conditions shown in Fig. 6 from the biological information of subject B.

4 Discussion

From the pre-questionnaire, 'Psychological stress' and 'Physical stress' items can be examined as pre-indicators that are prone to human error. In addition, it can be considered that the endurance of physical stress reduced human error as a reason that showed an inverse correlation between physical stress and the occurrence of human error. On the other hand, the biological information measured express lower pNN10 value during the task section than the rest section for all subjects. Here, we suggested that pNN10 can be used as a predictive indicator for human error.

As for the EEG indicators, many significant differences were observed between the Rest 2 and Task 2. It was suggested that the subjects were trying to concentrate even on the Rest 1 when comparing with Task 1. Thus, subjects may felt fatigued due to the repeating tasks and decided to take proper rest during Rest 2. Besides, there were many fluctuations in the indicators when human error occurs. However, there were some parts where the biological information does not change during the error occurrence. Here, we suggested that the subjects did not notice that they committed an error. Thus, it is necessary to compare the biological information during the human error occurrence and absence.

Next, we decided to re-examine the biological information measured by considering the time interval before and after the human error occurrence. Here, we compared the biological information measured in two conditions of time interval: 1) 10 [s] before and after each error condition and 2) 10 [s] before and after the error group condition. As a result, a higher significant difference rate was observed in 10 [s] before and after the error group condition. Therefore, we deduced that the human error did not occur instantaneously, but continuously. However, since the sample size is small, it is necessary to increase the number of subjects. In fact, the Stroop task performed in this study was based on the number of question. Therefore, it is necessary to investigate and compare the human error from the Stroop task that is based on the time limit instead of the number of question.

5 Conclusion

In this study, we analyzed the mental state before and after the human error occurrence and constructing a prediction model for human error considering individual differences. Here, we found that it is better to define consecutive human errors as one human error group where we decided to analyze the measured data 10 [s] before and after the human error group. Moreover, the 'Physiological stress' and 'Physical stress' items from the pre-questionnaire can be used as the predictive indicators since the results of the questionnaire showed correlation with the human error occurred during the task. Besides, in terms of biological information measured, pNN10 expressed fluctuations for all subjects during the task which can be considered as the predictive indicator. However, only two subjects showed fluctuations from β and γ indicators from EEG measurement during the task.

In the future, it is necessary to increase the number of subjects, since biological information vary among individuals. In addition, it is important to repeat the task for each subject in order to obtain more robust data. By improving these points, we aim to construct a prediction model for human error by considering individual differences as our next step. The machine learning then will be performed and the accuracy will be verified where the indicator that correlated with the error as the explanatory variable and the error occurrence as the objective variable. For our future step here, the logistic regression will be considered for machine learning to be used for the model construction.

Acknowledgments. The research was supported by Hitachi, Ltd. We also would like to thank the Research and Development Group, Naoko Takada, Tomoaki Hikita, and Yukinori Katagiri for their useful advice.

References

1. Statistics Bureau of Japan, Ministry of Internal Affairs and Communications, Statistics Bureau Home Page/Population Estimates Monthly Report (2021). https://www.stat.go.jp/english/data/jinsui/tsuki/index.html
2. St. John, M., Kobus, D.A., Morrison, J.G., Schmorrow, D.: Overview of the DARPA augmented cognition technical integration experiment. Int. J. Hum.-Comput. Interact. **17**(2), 131–149 (2004)
3. Kishimoto, T., Yoshida, R., Tobe, Y., Sugaya, M.: Examination of evaluation method on human error during work by bioinstrumentation. In: IEEE 27th International Requirements Engineering Conference Workshops (REW), Jeju Island, Korea (South), pp. 92–95 (2019)
4. Nishikawa, M., Nishihara, N., Tanabe, S.: The effect of moderately hot environment on performance and fatigue evaluated by subjective experiment of long time exposure. J. Environ. Eng. (Trans. AIJ) **74**(638), 525–530 (2009)
5. National Aeronautics and Space Administration: The NASA TLX Tool: Task Load Index, TLX @ NASA Ames (2020). https://humansystems.arc.nasa.gov/groups/tlx/

6. Yoshida, R.: Examination of human error prediction method by adaptive index selection (Thesis). Shibaura Institute of Technology (2020). (in Japanese)
7. Ikeda, Y., Horie, R., Sugaya, M.: Estimate emotion with biological information for robot interaction. In: 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2017), Marseille, France, pp. 6–8 (2017)
8. Lim, C.A., Chia, W.C., Chin, S.W.: A mobile driver safety system: analysis of single-channel EEG on drowsiness detection. In: International Conference on Computational Science and Technology (ICCST), Kota Kinabalu, pp. 1–5 (2014)
9. Hirai, F., Yoshida, K., Miyaji, I.: Comparison analysis of the thought and memory at the learning time by the simple electroencephalograph. In: Multimedia, Distributed, Cooperative and Mobile Symposium Proceedings, pp. 1441–1446 (2013). (in Japanese)
10. Shigemori, M.: Stroop task as a human error measure. In: The Proceedings of the 70th Annual Convention of the Japanese Psychological Association, vol. 70, p. 2PM106 (2006). (in Japanese)

Using Eye Tracking to Analyze the Effects of Spatial Contiguity in MOOC Video Subtitles

Xinyong Zhang(✉)

Renmin University of China, No. 59 Zhongguancun Street, Beijing, China
x.y.zhang@ruc.edu.cn

Abstract. In MOOC videos, the display positions of the subtitles are usually fixed at the bottom of the screen, which leads to the separation of the learning contents and the interpretation text (i.e. the subtitles) and thus inevitably affects the learning performance. Especially when students choose a foreign course with a non-native language as the teaching language from a global MOOC platform, this influence could be more significant. This work carried out a knowledge retention test as well as a knowledge transfer test to investigate the effects of the video subtitles with the implementation of the principle of spatial contiguity on learning performance, and further used the results of eye tracking to explain the underline mechanisms that why this principle could effectively improve the MOOC student's performance. This paper verified the necessity of applying the principle of spatial contiguity in the subtitles of MOOC instructional videos.

Keywords: Spatial contiguity · MOOC · Eye tracking · Learning performance

1 Introduction

In recent years, online education has increasingly become popular. MOOCs (Massive Open Online Courses), as a kind of typical online education resources, provide great opportunities for millions of students all over the world to access high-quality instructional resources via the internet. The websites of MOOCs overcome the geospatial limitations, and those students even could obtain certifications from the most famous universities for the courses they have taken and successfully finished. More and more educational institutions have been well aware of the importance of MOOCs, especially in the age of the COVID-19 pandemic, and paid more attention to the developments of MOOC resources, and there have been a number of influential MOOC websites. The contents in MOOC websites should be effective and efficient for users to learn with good performance. Researchers in different fields had extensively studied this issue from different aspects, especially from the perspective of cognitive psychology to deeply study and understand the behavior features of MOOC users so as to

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 336–354, 2021.
https://doi.org/10.1007/978-3-030-77932-0_27

lay the theoretical foundations and propose the guidelines for the designs and developments of MOOC resources.

In general, MOOCs have different static and/or dynamic instructional materials such as text, graphics, images, animations, audio and videos, and thus they are a kind of typical multimedia learning environments. In these environments, however, it would be inevitable to decrease the user's learning performance if there are additional cognitive loads due to the unsuitable designs of multimedia resources. Therefore, a kind of typical researches focused on how to properly integrate and present different instructional resources together, i.e. design and develop online resources for learning efficiently, based on cognitive load theory (CLT) and cognitive theory of multimedia learning (CTML). Researchers summarized and proposed a series of design guidelines after a number of experimental studies. For examples, the well-known principles of multiple representation, spatio-temporal contiguity, split attention, individual differences and coherence, which were proposed by Mayer and Moreno [18], had been widely accepted by the academia and extensively applied in practice. This work is related to the principle of spatio-temporal contiguity, which means that the different materials that art logically related to each other should be clearly organized together in time-line or visual space.

Using the three assumptions of CTML proposed by Mayer [20], the principle of spatio-temporal contiguity could be explained from different perspectives. The first is the assumption of dual channels. It assumes that human beings are able to acquire and process external information mainly via the auditory-verbal and visual-pictorial channels. According to this assumption, it is able to effectively facilitate the integration of verbal and visual representations in the cognitive activities of the brain, reduce the cognitive load of learners, and thus improve the learning performance when harmoniously integrating verbal and visual materials together. The second is the assumption of limited capacities in each of the channels. It is believed that the acquisition and processing of information in human cognitive activities have to consume a certain cognitive resources, which are limited either in verbal or visual channel. Thus, the information capacity that can be processed in each of the channels is also limited. According to this assumption, the larger the learner's intrinsic cognitive load, the smaller the learner's extraneous cognitive load. Therefore, the activities irrelevant to learning, such as visual searching, should be reduced as many as possible. For example, it is able to effectively reduce unnecessary visual searches, and benefit to the increase of learning performance under the limited cognitive capacity condition when placing explanatory text adjacently with their related images. The third is the assumption of active processing. It is taken for granted that learning involves a series of cognitive processing, and the learner will take the initiative to choose learning materials, consciously organize and integrate them with the existing knowledge. If the interrelated text and pictures are also close neighbors to each other in space, the cognitive load of integrating different materials could be effectively reduced.

There had been a number of studies that investigated and verified the positive effects of spatial contiguity on learning performance. However, there was still no study to investigate its effect in MOOC video subtitles, especially for non-native speakers of English. Regarding research methods, besides the techniques of EEG, ECG and traditional questionnaire, the method of eye tracking had also been increasingly applied in the studies of multimedia learning. According to the assumption about the coherence between cognition process and eye movements, the measures of eye movements can be used as the direct evidences for the involvement of cognition during the user's learning processes. With the increasing maturity of eye tracking technology, it had been widely applied in different fields of researches, especially including the field of learning and education since 2000. Employing this technology, it is able to more deeply understand the information processing pattern in the learning process, the effectiveness of multimedia learning resource designs, and the validity of existing theories. This work used a eye tracker (Tobii TX300) and carried out an eye tracking experiment to examine the effects of spatial contiguity principle when applied in MOOC video subtitles for Chinese learners to reveal their cognitive involvement when they were learning courses in English and provide useful design guidelines for the instructional resources in international MOOC platforms.

2 Related Work

This section provides a review about the related work from the perspectives of the significance of spatial contiguity, the effects of video subtitles and the application of eye tracking on/in multimedia learning.

2.1 Significance of Spatial Contiguity for Multimedia Learning

Online learning is very different from the traditional way of learning. Volery and Lord pointed out that one of the key success factors of online learning is the usability of the resources that needs to be maintained and increasingly improved in technique [29]. This is also the intrinsic motivation that the researchers, such as Mayer [18], carried out intensive researches on the design and optimization of online resources, and proposed a number of principles for practical applications, of which the common one is spatial contiguity.

Actually, Tarmizi and Sweller firstly mentioned the principle of spatial contiguity [28]. In their work on how to guide students more effectively to solve mathematical problems by examples, they pointed out that the pictures in the learning materials and the corresponding explanatory text should be placed as close as possible in space. Otherwise, students will switch their attention between different information sources, leading to a significant increase in cognitive load. Mayer preliminarily verified the practical effects of spatial contiguity [16]. He designed two instructional materials to demonstrate the working principle of tire pump. One put the pictures and the corresponding explanatory text closely on the same page, and the other put them on two different pages. The experimental results indicated that the former, which reflected the spatial contiguity principle, improved the students' transfer test scores by 75% and was able to incorporate the students' knowledge more better [21]. After that, Mayer and other colleagues further refined the experiment by designing different layouts of text and illustrations with or without annotations as the stimuli. The results showed that when the illustration with annotations were arranged close to the corresponding text, the effect was better than that of the controlled layout in which the illustration, without annotations, was separated from the corresponding text, and the students' performance in the knowledge transfer test could be improved by nearly 50%, especially more obvious improvements for those students without relevant knowledge background.

Moreno and Mayer further changed the scenes of multimedia learning from static to dynamic, and studied the effects of animation instructional materials, which employed different information integration methods to properly reflect the effectiveness of spatial contiguity, on learning performance [22]. Their experiment revealed that the animation course in which the explanatory text was close to the instructional picture could result in better learning performance. This result confirmed that the effectiveness of spatial contiguity to promote learning performance was valid not only in static instructional materials (such as books, newspapers and magazines) but also in dynamic materials (such as teaching video, film). In other words, the principle of spatial contiguity could also be properly applied in the design of dynamic instructional materials to support multimedia learning more better [19]. Based on the existing theories and his own experimental results, Mayer systematically incorporated this principle into CTML, and believed that the principle of spatial contiguity could theoretically reduce the cognitive load of learners, make them more focused, and thus effectively improve the learning performance [17]. Subsequently, using an eye tracker, Jonson and Mayer further explained the internal mechanism of the effectiveness of this principle from the perspective of learners' eye movement features, i.e. the layout pattern that arranged the relevant text and pictures closely to each other could make the heat map of learners' gaze points more concentrated, the viewpoint switched more frequently, the switching time more short, and fixation on pictures more longer [9].

In China, there were also a certain amount of research on the related field of multimedia learning. One of the main research topics was the design and development of multimedia instructional resources [40]. These research focused more on the application of multimedia learning theory in order to optimize the presentation of instructional information [7,8,10]. The corresponding research on the patterns of information presentation was not only from the perspective of multimedia combinations, but also some of them specifically focused on the principle of spatial contiguity or its applications. Wang et al. further added color cues to the classic animation materials about "how lightning works", and employed eye tracking technology to verify whether or not color cues had positive supplementary effects to the layout pattern of text and pictures with spatial contiguity

[34]. The results showed that although color cues did not independently and significantly improve the learning performance, the eye tracking data showed that color cues could effectively promote the cognitive processing during fixating on text and pictures, and further enhance the effects of spatial contiguity. The validity of spatial contiguity had been extensively verified for the learning of different knowledge domains [38], Wang et al. further considered several adjustment factors, such as learning materials, learning environments and subjects' education levels, to address the drawbacks in previous studies, and provided a meta-analysis about the reliability of spatial contiguity from the perspectives of retention test, transfer test and subjective cognitive load. They reported 1) that the effect of spatial contiguity was more prominent to facilitate the deep understanding of learning contents in the environments with higher interactivity and shorter duration of learning; 2) that it was more prominent to promote knowledge acquisitions for primary and middle school students; and 3) that the effects in the learning media of books or paper was more prominent than that in electronic virtual environments.

With respect to the application studies of spatial contiguity, researchers in China had addressed the topics of special education, middle school curriculum, foreign language teaching and multimedia courseware. Regarding to primary school students with dyslexia, Li pointed out that the integration of text and pictures based on the principle of spatial contiguity was more effective than the combination of pictures and speech [13]. The investigation of Wan and Bi indicated that the teachers in middle schools should also pay attention to the application of spatial contiguity when they used illustrations in classroom teaching [30]. Regarding to foreign language teaching, this principle should be implemented no matter it was the information presentation in preschool English courses [12] or multimedia information presentation in college English [41].

2.2 Effects of Video Subtitles on Multimedia Learning

Although the role of subtitles in some scenarios was doubtful, it was believed that the use of subtitles would not significantly improve learning performance but lead to "redundancy effect" [23]. With the widespread use of online instructional videos, however, the value of the subtitles was approved in more scenarios, and researchers carried out a number of studies addressing the issues about the necessity, attributes, information capacity and presentation style of subtitles.

The early studies mainly focused on the visual factors that could directly affect the readability, such as font shapes, sizes, colors, brightness and backgrounds used for displaying subtitles [44]. Further research showed that these factors generally had significant effects on reading performance, and that each of the factors had its own suitable levels to apply [39], or needed to be treated differently in different scenarios [27]. The dynamic features when displaying subtitles, such as frequency and speed, had also been proved to be able to significantly affect learning performances [6, 26]. As for the effect of the information capacity of subtitles on learning performance, researchers had carried out comparative studies in different situations, showing that the instructional videos with subtitles could bring better learning performance and even could improve students' subjective satisfactions [1,3,35]. Using a questionnaire, Wang specifically investigated the effects of the subtitles that were synchronized with the speaker voice in the video, and confirmed that the subjective satisfaction of students with the use of subtitles was positive [31]. After that, he further explored the effects of the subtitles with different information capacities, such as full subtitles, selected keyword subtitles and no subtitle, on students' learning performance. His work confirmed the necessity of captioning videos, especially the superiority of only presenting the selected keyword subtitles, and also indicated that the "redundancy effect" did not obviously appear [32].

The usefulness of subtitles had been highly concerned especially in the field of foreign language learning [42]. Original movies and televisions in foreign languages are often used as the materials to train students' listening comprehension, and the original context can also enhance students' interest in learning. Liu studied and compared the influences of different display styles on English learners' listening comprehension [15]. The display styles specifically designed in Liu's work included 1) entirely displaying a subtitle sentence in advance but highlighting the corresponding words in sync with the speech, similar to that used in karaoke music videos; 2) displaying the corresponding words one by one in sync with the speech; 3) displaying subtitle sentences with fade-in and fade-out animations; and 4) the traditional method to display subtitles, used as a baseline for comparison. The experimental results indicated that the first display style was able to promote students' listening comprehension, on which the third style also had a significant effect, while the word-by-word display style had a negative effect. Dong et al. further analyzed the effect of subtitle information capacity on students' listening comprehension, and they argued that the students, with different listening levels, could distribute their attention according to the information capacities, and that the keyword subtitles were more reasonable choices so that students could balance the processing of verbal and visual information, e.g. text and images [5]. From the perspective of subtitle languages, Wang studied and analyzed the influence of English full subtitles, Chinese full subtitles and English keyword subtitles on students at different grades [33]. The main results showed that the keyword subtitles were more suitable for students with higher language proficiency and were conducive to learners' understanding of vocabularies; and that full English subtitles were more helpful to lower grade students than keyword subtitles. Li also carried out a research from the same perspective, but the differences were that the subjects participating that research were undergraduates major in English language, and that the subtitles included a bilingual display mode. The results indicated that bilingual subtitles were more valuable than monolingual subtitles [14].

Furthermore, some researchers had studied the designs of subtitles for specific devices, specific video formats, and even special learner groups. With the increasing popularity of mobile devices, Zhao et al., using EEG, studied the effect of subtitle information capacities on improving students' academic performances [43]. They reported that the effect of keyword subtitles was comparable to that of full subtitles in mobile terminals. For the videos converted from PPT documents, Pang et al. studied the influence of subtitles with different presentation modes on learning performance, and the results indicated that the mode to present all the subtitles as a catalog was better than the mode to sequentially present each of the subtitles [24]. With respect to hearing impaired users, Cao et al. employed an adaptive engine technology to automatically match the style, speed, navigation and automatic word segmentation of subtitles so as to achieve a barrier-free presentation for hearing impaired users [2].

Some studies also considered the effect of subtitle position on learning performance. Shui and Wang investigated the effect of different screen positions on college students' identification of the emotions in Chinese idioms. The results revealed that the error rates in the positions of top right and bottom left were significantly lower than those in other positions, showing a significant location effect [27]. However, the experimental task was relatively too simple to represent the cognitive load level in multimedia learning environments. Liu investigated the effect of different positions of subtitles on students' listening comprehension of foreign language, but did not find significant differences [15].

2.3 Eye Tracking Devices Used in the Studies of Multimedia Learning

In the research field of multimedia learning, it is a common method to use advanced technologies to collect data, such as eye tracking and EEG as mentioned above. Especially, the technology of eye tracking was the most representative one, and since the beginning of this century, it has been widely used in different research [11]. Eye tracking can effectively provide a number of quantitative measures about the temporal and spatial distributions of eye movements, such as fixation time, saccade path and heat map of gaze points, which provide direct or indirect evidences for interpreting the cognitive processing of subjects.

With regard to the principle of spatial contiguity, Schmidt-Weigand et al. used eye tracking technology to analyze the differences in the distributions of users' attention between pictures and text in different layouts, and found that students would pay more attention to pictures in the case that the relevant pictures and text are closely placed to each other, and that there was a positive correlation between fixation time and learning performance [25]. Using eye tracking, Diao et al. analyzed the effects of different learning materials, such as plain text, videos with or without subtitles, on students' learning behaviors in online learning environments. Their results uncovered that the cognitive load of video materials with subtitles was the lowest, while that of plain text in online courses was the highest [4]. Wang summarized the selection of eye tracking measures and matters needing attention for the research on multimedia learning [36], and then she analyzed the effects of subtitle information capacities on learners under two conditions of declarative and procedural knowledge types [37]. The result showed that the learning performance had positive correlation with the fixation durations in subtitle areas, further analysis implied that it was more reasonable that the videos with different knowledge types should configure different types of subtitles, e.g. full vs. keyword subtitles.

3 Experiment

As mentioned above, the factor of subtitle position had been taken into account in some studies, but it can be found that there were obvious deficiencies in their experimental designs. First, although different presentation positions had been considered, essentially the subtitles were still displayed in a fixed position. The significant differences observed in this case reflected the screen position effect rather than the effect of spatial contiguity between the subtitles and their corresponding video contents. Second, the experimental tasks were too simple or aimed at learning foreign languages more better, and the evaluation measures were mainly based on the scores of traditional single and/or multiple choice questions, which did not properly reflect the influence on the capabilities of knowledge retention and transfer of the subjects in learning process.

The principle of spatial contiguity had been widely used in static teaching materials, but not in dynamic teaching videos, in which the captions are usually fixed at the bottom of the screen. In this study, the key point of setting the position of subtitle is that it is not presented in a fixed position on the screen, but dynamically presented in the adjacent position according to the actual area of the learning content in the video.

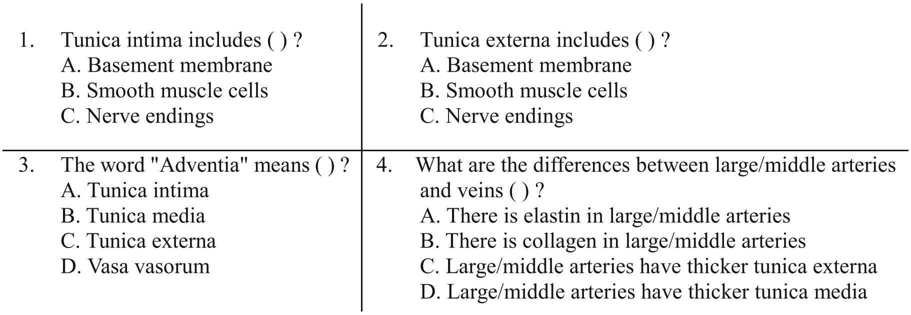
3.1 Experimental Materials

Video Learning Material. In this study, there were three main principles for selecting video learning materials. First, the focus in the video material should be changed, and the change should not be too fast. Second, the subjects were non-native speakers of the language used in the video so that it was able to enhance the dependence of the subjects on subtitles. One of the independent variables in this experiment was subtitle position, and the video with non-native language of the subjects could make the experimental effects more prominent. Finally, the difficulty of the video learning material should be moderate for the subjects, without a specific requirement for any professional knowledge background. According to these principles, we picked out and clipped a 7-min teaching video from the chapter of "Blood Vessels" in the biology popular science course in the public MOOC platform of Khan Academy.

There were two groups of subjects to learn the video material, with two different subtitle displaying modes respectively. The subtitles used for group A were always at the bottom of the screen, while those for group B dynamically changed according to the learning content in the video clip. For example, as in the screen-shots in Fig. 5, when the teacher's mouse cursor was at the top right of the screen, the subtitle for group A was displayed at the bottom of the screen, and that for group B was adjacent with the teaching content.

Learning Test Materials. The design principles of the test materials also had three main points. First of all, test scores should not only reflect the learner's ability to remember in the learning process, but also reflect the learner's ability to understand the learning material. Secondly, the questions used in the learning test should be closely related to the changes of subtitle positions. For example, when the teacher was pointing the mouse cursor at the content she/he was talking about, the corresponding subtitle should be displayed near to the cursor, and the knowledge related to that content could be used in the learning tests. Finally, the questions of the learning tests should evenly cover the whole contents of the video material. According to these principles, the test materials were divided into two parts.

- *Retention Test Materials.* Retention tests can measure how well learners remember what they have learned. There were 3 single choice questions and 1 multiple choice question used to examine the subjects' capability of memorizing the learning material.
- *Transfer Test Materials.* Transfer tests can detect how well learners understand what they have learned. There were 3 short-answer questions and 2 multiple choice questions used to examine the subjects' understanding of the learning material.

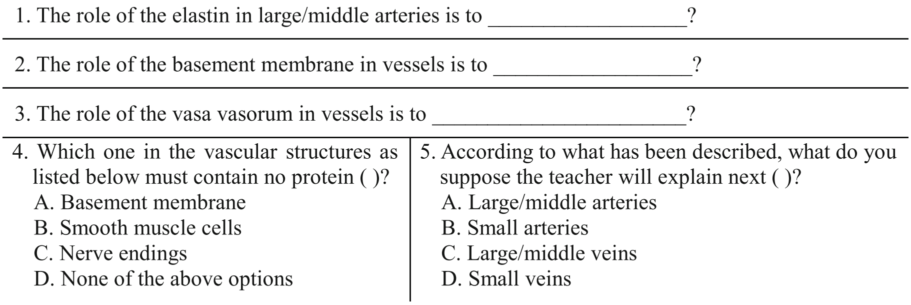


1. Tunica intima includes ( ) ? A. Basement membrane B. Smooth muscle cells C. Nerve endings	2. Tunica externa includes ( ) ? A. Basement membrane B. Smooth muscle cells C. Nerve endings
3. The word "Adventia" means ( ) ? A. Tunica intima B. Tunica media C. Tunica externa D. Vasa vasorum	4. What are the differences between large/middle arteries and veins ( ) ? A. There is elastin in large/middle arteries B. There is collagen in large/middle arteries C. Large/middle arteries have thicker tunica externa D. Large/middle arteries have thicker tunica media

Fig. 1. The questions used in the retention test.

3.2 Subjects and Apparatus

A total of 28 subjects with normal or correct-to-normal vision were recruited in the campus of Renmin University of China. After a pre-recruitment investigation, they had no specific knowledge background on the teaching video content (related to science and general biology). The 28 subjects were randomly divided into Group A and Group B, which happened to include 6 boys and 8 girls. The subtitles used for Group A were presented in traditional form, while the subtitles



1. The role of the elastin in large/middle arteries is to ________________?

2. The role of the basement membrane in vessels is to ________________?

3. The role of the vasa vasorum in vessels is to _____________________?

4. Which one in the vascular structures as listed below must contain no protein ( )?
A. Basement membrane
B. Smooth muscle cells
C. Nerve endings
D. None of the above options

5. According to what has been described, what do you suppose the teacher will explain next ( )?
A. Large/middle arteries
B. Small arteries
C. Large/middle veins
D. Small veins

Fig. 2. The questions used in the transfer test.

for Group B were designed based on the principle of spatial contiguity. Before the experiment, the subjects needed to sign the informed consent form.

An eye tracker, Tobii TX300, was used to track the subjects' eye movements during the learning process. Its working sampling frequency was set 250 Hz. The device and the corresponding data analysis software system were installed on a HP DC7700 desktop computer, with an Intel dual-core 2.0G CPU, 8 GB RAM, and a 23-in LCD display at the resolution of 1920 × 1080 pixels.

3.3 Experiment Procedure

The experiment consists of five main steps, which were as follows:

- (1) *Introduction of the experiment and device calibration.*
Before the formal start of the experiment, the experimenter first introduced the basic content, process and matters needing attention, and carefully answered the questions raised by the subject. The experimenter then had the subject sit about 70 cm in front of the computer screen and made sure that when the subject was looking straight in front of them, her/his gaze was roughly in the center of the screen. Before the experiment begins, the experimenter needed to calibrate the eye tracker to an acceptable accuracy, otherwise the experiment could not proceed to the next step.
- (2) *Background knowledge preview.*
The subjects skimmed the test questions for a short period of time to prepare for later video learning. Learners are easily nervous when suddenly exposed to unfamiliar knowledge. This unstable emotion may be reflected in the eye tracking data, which will cause certain interference to the experiment. Therefore, it was necessary for the subjects to have a general understanding of the learning content and the corresponding test questions in advance. This step also allowed the subjects to relax before the formal experiment.
- (3) *Video learning.*
The subject was going to begin the video learning at this stage. The subjects in Group A watched the instructional video with subtitles at the bottom of the screen. When the subjects in Group B watched the video, the subtitles

would change their position dynamically. Considering that the subtitles in most of the videos are displayed at the bottom of the picture in general, the experimenter reminded how the subtitles would be rendered on the screen.

- (4) *Retention test.*
  In this stage, the subject was given 5 min to answer the retention questions as shown in Fig. 1.
- (5) *Transfer test.*
  In this stage, the subject was given 5 min to answer the transfer questions as shown in Fig. 2.

4 Results

4.1 Learning Performance Analysis

Table 1 shows the test scores of the two groups. As can be seen, the average scores in both retention test and transfer test as well as the total score of Group A were lower than those of Group B. Among them, the change ranges of the transfer test score and the total score were relatively larger, while the change range of retention test score was small.

A one-way between subjects ANOVA showed that the factor of subtitle display mode had no significant main effect on retention test score ($F_{1,26} = 0.310, p = .582$), but with significant main effects on transfer test score ($F_{1,26} = 6.520, p < .005$) and total score ($F_{1,26} = 4.393, p < .005$).

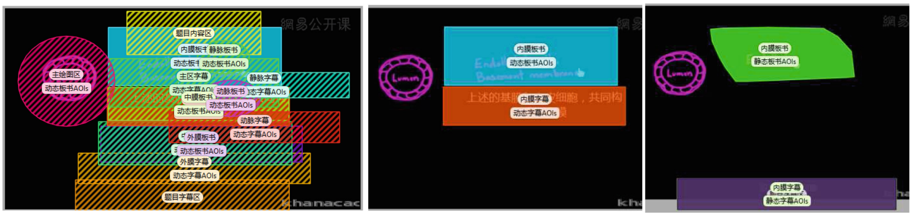
Table 1. Average scores of the two subject groups ($M \pm std$).

Group	Score in retention test	Score in transfer test	Total Score
Group A, watching the MOOC video with the subtitles constantly displayed at the bottom of the screen.	3.64 ± 1.08	2.64 ± 1.22	6.29 ± 1.77
Group B, watching the MOOC video with the subtitles dynamically displayed according to the principle of spatial contiguity.	3.86 ± 0.95	3.93 ± 1.44	7.79 ± 2.01

4.2 Eye Tracking Data Analysis

According to the specific content of the selected video material, there were seven pairs of areas of interesting (AOI) defined to cover different video contents and their corresponding subtitles. For example, when the video was explaining the tunica intima of blood vessels by blackboard-writing, the writing region and its corresponding subtitle region were defined as a pair of AOIs to collect eye movement data. Thus, all the AOIs could be divided into two categories. One was the AOIs responsible for content presentation, the other was responsible for subtitle display. In general, according to the time line of the video, only a pair of AOIs was activated, while the rest were inactive. Furthermore, there should be no overlap between any pair of AOIs.

As shown below, we collected and analyzed four eye tracking measures according to the defined AOIs. The subjects whose effective eye tracking data had relatively low sampling ratio, e.g. less than 85%, would be excluded from the analysis. Thus, there were one subject in Group A and two subjects in Group B were excluded. In addition, the spatial distribution characteristics of the scanning paths and the heat maps of gaze points were also analyzed (Fig. 3).



(a) the distribution of all the AOIs for Group B. (b) a pair of activated AOIs for Group B, they were adjacent to each other. (c) a pair of activated AOIs for Group A, they were separated.

Fig. 3. The AOIs defined in the instructional video.

Gaze Visit Times and Fixation Times. The numbers of gaze visits and fixations in the AOIs were counted. Using a multivariate ANOVA for the differences between the two groups, it could be found that the effect of subtitle display mode was significant on visit times in subtitle AOIs ($F_{1,23} = 5.891, p < .05$), but not on visit times in content AOIs ($F_{1,23} = .000, p = .997$). Similarly, there was a significant effect on fixation times in subtitle AOIs ($F_{1,23} = 10.435, p < .005$), but not on fixation times in content AOIs ($F_{1,23} = 2.173, p = .154$).

Gaze Visit Duration and Fixation Duration. A multivariate ANOVA indicated that the factor of subtitle display mode had significant effects on the measure of visit duration in both subtitle AOIs ($F_{1,23} = 8.838, p < .01$) and content AOIs ($F_{1,23} = 6.166, p < .05$). Similarly, it had significant effects on fixation duration in both subtitle AOIs ($F_{1,23} = 8.329, p < .01$) and content AOIs ($F_{1,23} = 6.527, p < .05$).

Figure 4 shows the comparisons of the total numbers and the total duration of gaze visits as well as those of fixations in content AOIs and subtitle AOIs. As can be seen, in content AOIs, the total visit and fixation times of Group B were both smaller than those of Group A; while in subtitle AOIs, the total visit and fixation times of Group B were both higher than those of Group A. The results of ANOVA indicated that the factor of subtitle display style resulted

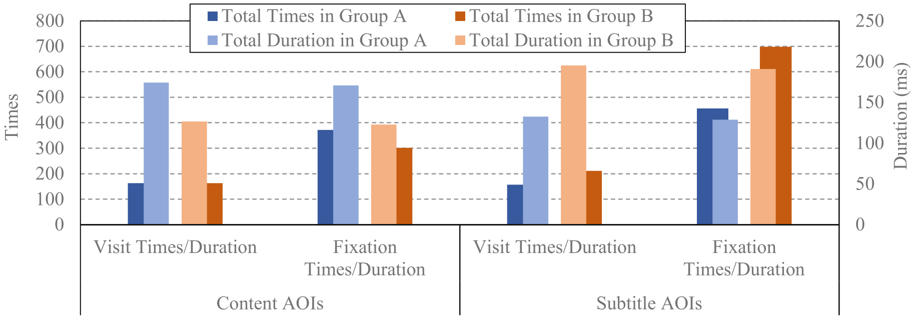
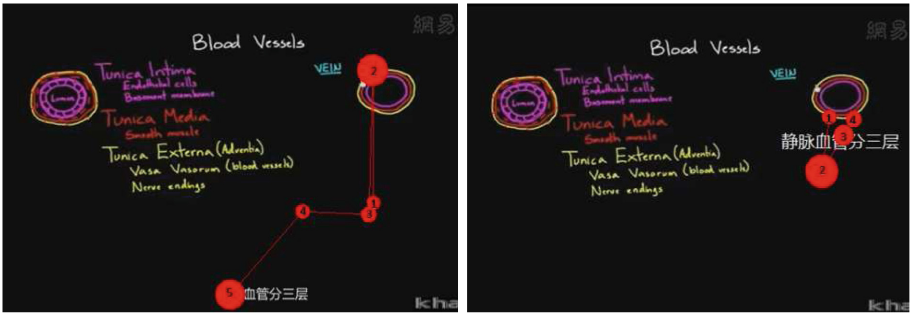


Fig. 4. Total gaze visit times and duration as well as total fixation times and duration in different AOIs.

in significant differences of visit times and fixation times only in subtitle AOIs. Similarly, in content AOIs, the total visit and fixation durations of Group B were both shorter than those of Group A; while in subtitle AOIs, the total visit and fixation durations of Group B were both longer than those of Group A. However, the effects of subtitle display mode on visit and fixation durations in both subtitle and content AOIs were all significant.

Eye Scan Path. Eye scan path can directly display the process of searching and scanning when the learner is acquiring external information. As shown in Fig. 5, the red circle represents the fixations of the learner, and the straight line between two circles indicates the learner's saccade between two fixations. A series of fixation-saccade-fixation, which is arranged in chronological order, directly reflects the learner's cognitive processing of learning materials. The circle diameter represents the duration of fixation. In general, the longer the duration, the higher the cognitive load. At the same time, The length of the straight line between two fixations represents the search path of the learner. Generally, the longer the search path, the lower the search efficiency.

After a careful observation on Fig. 5, it was found that the eye scan paths in the two groups had a consistent feature, i.e. the learner firstly looked at the picture area where the teacher was explaining and demonstrating the teaching content, and then searched for the corresponding subtitle. According to the sizes of the circles, it can be seen that the learners in the two different groups had similar fixation durations in the picture area (content AOI) and the subtitle area (subtitle AOI). According to the numbers of fixations on the scan paths, it can be seen that there were more meaningless fixations generated when the learner's fixation switched from the content AOI to the subtitle AOI in Group A, such as the third and fourth fixations. It implied that the learner had relatively higher cost of acquiring useful information when the subtitle was separated from the corresponding teaching content.

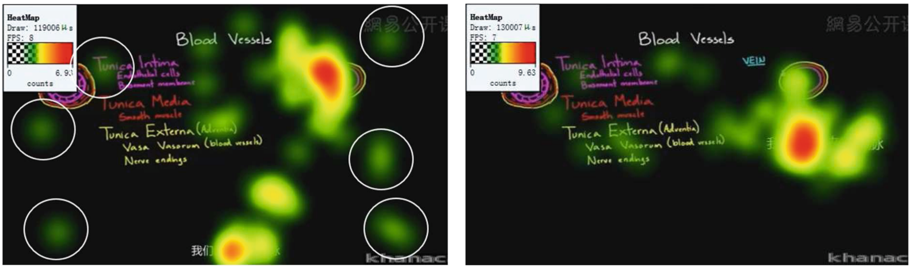


(a) a scan path sample in Group A

(b) a scan path sample in Group B

Fig. 5. Two scan paths for a same learning content in both groups. Note that real positions of the fixations in Goup B should be moved upward a certain distance according to the context of the learning content. (Color figure online)

Heat Map of Gaze Points. Heat map can intuitively show the distribution of the learner's fixations in the learning process and directly reflect the areas of learner's concentration of interests. In general, the more concentrated the distribution of gaze points, the more concentrated the learner's interested ares, and the lower the cognitive load in the learning process.



(a) a heat map in Group A

(b) a heat map in Group B

Fig. 6. Two heat maps for a same learning content in both groups

It was found that the heat maps had distinct characteristics under different subtitle conditions. Figure 6 shows two heat maps that were clipped from Group A and B respectively for the same content. As can be seen, there were two areas with the most prominent heats in the map of Group A, but only one in the map of Group B.

5 Discussion

From the perspective of the influence of subtitle display mode on learning performance, overall it had reached at a significant level. Although no significant effect was observed in the retention test, the subtitle displaying positions reflecting the principle of spatial contiguity in Group B still improved the score of retention. This point was consistent with the change trend of the score in the transfer test as well as the total score. Therefore, it was possible that the questions used in the retention test was not full enough to properly reflect the effect of subtitle display mode. In other words, this work is not enough to draw a negative conclusion that the position of subtitle presentation has no significant effect on retention performance.

According to the eye tracking measures depicted in Fig. 4, the significant effect on the performance in the transfer test was most probably due to the significant increases of the visit and fixation times as well as the significant increases of the visit and fixation durations because the four measurements had a consistent change trend only in subtitle AOIs, i.e. these measures in Group B were all significantly higher than those in Group A. A further comparison of the measures in different AOIs revealed that the learner could allocate more attention to content AOIs when the learning contents and the corresponding subtitles were separated, but at the same time, the learner could also allocate enough attention to subtitle AOIs, thus resulting in two prominent heat areas, in other words, both areas could generate obvious cognitive loads respectively; and that the learner could allocate more attention only to subtitle AOIs when the learning contents were very close to their corresponding subtitles so that there was not a prominent heat area generated in the learning content area as that in the subtitle area, in other words, the cognitive loads in the two areas were obviously different. Combining the results of the two aspects, it appeared that the learner could acquire more information from pictures when they were separated from their corresponding subtitles, but she/he could not had a deep enough understanding of the learning content, while the learner could acquire more information from text so as to understand the learning content more fully when the subtitles and the pictures were adjacent to each other. If there were two prominent heat areas, it implied the inevitable competition of cognitive resources between the content AOIs and the subtitle AOIs; and if only one prominent heat area, it implied the natural integration of cognitive processing. Furthermore, the information from text could better enable the learner to acquire comprehensive cognition and promote the transfer of knowledge.

Additionally, there were more findings that could be revealed from the eye tracking results. With respect to the heat map of Group A, there were a number of areas with relatively lower heats outside the regions where the learning contents and the subtitles were displayed, such as the spots marked with white circles in Fig. 6a. At the same time, there was an obvious secondary heat area between the two prominent heat areas that were related to the picture and the corresponding subtitle, respectively. As for the results of Group B, the heat areas corresponding to the adjacent picture and subtitle areas were almost integrated together, and the number of lower heat areas was less than that in the heat map of Group A. The heat areas were more concentrated as a whole in the map of Group B. These findings implied that the learner could have more cost of information search and recognition when the learning content and the subtitle were separated, and there was more than one saccades taking place to switch the fixation from the picture area to the subtitle area. The scattered areas, with lower heats, outside the regions of the two prominent heat areas indicated that the learner still might adopt the strategy of random search even if the subtitles were constantly displayed in the lower part of the screen, implying the additional increasing of extraneous cognitive load.

6 Conclusions

Employing an eye tracking device and an online learning video selected from a public MOOC platform, this work investigated how the subtitles that were displayed according to the principle of spatial contiguity could affect the learning performance in video materials, in which the instructional contents were presented and explained using a foreign language but the subtitles were displayed using the native language of the learner (e.g. English vs. Chinese). The results of the experiment indicated that compared with traditional subtitles, the subtitles with the implementation of spatial contiguity could significantly improve the learners' performance in the test of knowledge transfer; and that the total performance still had a significant improvements while there was no significant improvement in the test of knowledge retention.

A further analysis of eye tracking data revealed that when the learning content AOIs and the subtitle AOIs were separated from each other, the learner could allocate more attention to the former, but she/he could also allocate comparable attention to the later, resulting in two prominent heat areas of gaze points; and that there were a secondary heat area between the two prominent heat areas as well as a number of scattered areas with lower heats outside the regions of displaying the learning content and its subtitle. When the two kinds of AOIs were close to each other, the learner could allocate more attention to the subtitle AOIs but relatively less to the content AOIs so that there was no prominent heat area generated in the content AOIs but in the subtitle AOIs.

According to the results of eye tracking, we can properly explain the cognitive mechanism of the learner's performance improvement in the test of knowledge transfer. When there are two separated prominent heat areas, the learner's limited cognitive resources inevitably need to be dispersed to deal with different information sources. The scattered areas with less heats and the secondary heat area in the heat map of gaze points imply that the learner probably could perform a random visual search on the display, thus it is inevitable to increase the extraneous load in the learning process. When there is only one prominent heat area, the learner can make full use of cognitive resources and make full use of germane load to achieve a more comprehensive understanding of the learning materials, so as to achieve higher performance in knowledge transfer.

References

1. Bernard, R.M.: Using extended captions to improve learning from instructional illustrations. Br. J. Edu. Technol. **21**(3), 215–225 (1990)
2. Cao, Y., Wu, P., Lin, X., Jia, T.: The design of video barrier-free subtitles for deaf people based on adaptive learning needs. Mod. Educ. Technol. **29**(7), 99–105 (2019)
3. Chung, J.M.: The effects of using video texts supported with advance organizers and captions on Chinese college students' listening comprehension: an empirical study. Foreign Lang. Ann. **32**(3), 295–308 (1999)
4. Diao, Y., Liu, M., Yang, H.: Study on online video open courses learning behavior in the eye tracker test. Mod. Educ. Technol. **24**(11), 45–51 (2014)
5. Dong, J., Zhou, Y., Liu, G.: Perceptual load and selective attention in multimedia English videos watching: The effect of keyword captions on video comprehension. Technol. Enhanced Foreign Lang. Educ. **2013**(153), 57–64 (2013)
6. Gao, D., Zhong, Y., Zeng, L.: A research on how frequency influences the recognition speed of common Chinese characters. J. Psychol. Sci. **18**(4), 225–229 (1995). https://doi.org/10.16719/j.cnki.1671-6981.1995.04.008
7. Huang, S., Shi, G., Li, M.: The effect of resource presentation on the learning effects of different learning styles. J. Zhejiang Norm. Univ. (Nat. Sci.) **41**(4), 410–415 (2018). https://doi.org/10.16218/j.issn.1001-5051.2018.04.008
8. Jiang, W., Bao, H., Yang, Z., Wang, D., Li, H.: The influence of emotional design and presentation on multimedia learning: behavior and eye movement studies. Educ. Modern. **6**(45), 273–278 (2019). https://doi.org/10.16541/j.cnki.2095-8420.2019.45.091
9. Johnson, C.I., Mayer, R.E.: An eye movement analysis of the spatial contiguity effect in multimedia learning. J. Exp. Psychol. Appl. **18**(2), 178–191 (2012)
10. Kang, C., Zhou, A.: The effect of the representational mode of information and learners' cognitive style on learning in the multimedia environment. J. Psychol. Sci. **33**(6), 1397–1400 (2010). https://doi.org/10.16719/j.cnki.1671-6981.2010.06.019
11. Lai, M.L., et al.: A review of using eye-tracking technology in exploring learning from 2000 to 2012. Educ. Res. Rev. **10**(10), 90–115 (2013)
12. Li, H., Jin, M., Zhang, X.: A brief analysis of information presentation and representation of children's English castle. Educ. Teach. Forum **2013**(27), 138–139 (2013)
13. Li, K.: The effect of the multimedia-based presentation mode on dyslexic children's learning achievements. Chin. J. Spec. Educ. **2013**(9), 46–49 (2013)
14. Li, M.: Impact of bilingual subtitles on students' vocabulary acquisition based on RHM. Foreign Lang. Lit. **33**(2), 140–147 (2017)
15. Liu, G.: The influence of subtitle presentation on listening comprehension of Chinese English learners. Master's thesis, Jiang Nan University (2012)
16. Mayer, R.E.: Systematic thinking fostered by illustrations in scientific text. J. Educ. Psychol. **81**(2), 240–246 (1989)
17. Mayer, R.E.: Multimedia Learning, 2nd edn. Cambridge University Press, Cambridge (2009)
18. Mayer, R.E., Moreno, R.: A cognitive theory of multimedia learning: implications for design principles. In: Annual Meeting of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 1–11 (1998)

19. Mayer, R.E., Moreno, R.: Animation as an aid to multimedia learning. Educ. Psychol. Rev. **14**(1), 87–99 (2002). https://doi.org/10.1023/A:1013184611077
20. Mayer, R.E., Moreno, R.: Nine ways to reduce cognitive load in multimedia learning. Educ. Psychol. **38**(1), 43–52 (2003)
21. Mayer, R.E., Steinhoff, K., Bower, G., Mars, R.: A generative theory of textbook design: using annotated illustrations to foster meaningful learning of science text. Educ. Tech. Res. Dev. **43**(1), 31–43 (1995). https://doi.org/10.1007/BF02300480
22. Moreno, R., Mayer, R.E.: Cognitive principles of multimedia learning: the role of modality and contiguity. J. Educ. Psychol. **91**(2), 358–368 (1999)
23. Moreno, R., Mayer, R.E.: Verbal redundancy in multimedia learning: when reading helps listening. J. Educ. Psychol. **94**(1), 156–163 (2002)
24. Pang, L., Xiang, H., Zuo, J., Huang, W.: A study on the effect of subtitles on learning performance in instructional video. Comput. Knowl. Technol. **15**(36), 205–206, 211 (2019). https://doi.org/10.14004/j.cnki.ckt.2019.4347
25. Schmidt-Weigand, F., Kohnert, A., Glowalla, U.: Explaining the modality and contiguity effects: new insights from investigating students' viewing behaviour. Appl. Cogn. Psychol. **24**(2), 226–237 (2010)
26. Shui, R., Fu, D., Li, Z., Shen, M.: The effects of speed, step and window size on comprehension of Chinese text presented in leading format. J. Psychol. Sci. **24**(2), 141–144, 233–253 (2001). https://doi.org/10.16719/j.cnki.1671-6981.2001.02.005
27. Shui, R., Wang, L.: The effect of font size and position on the Chinese character comprehension in the multimedia courseware. Chin. J. Appl. Psychol. **14**(2), 187–192 (2008)
28. Tarmizi, R.A., Sweller, J.: Guidance during mathematical problem solving. J. Educ. Psychol. **80**(4), 424–436 (1988)
29. Volery, T., Lord, D.: Critical success factors in online education. Int. J. Educ. Manag. **14**(5), 216–223 (2000)
30. Wan, Y., Bi, H.: A study on the use of textbook illustrations in classroom teaching. Educ. Sci. Res. **2012**(11), 65–69 (2012)
31. Wang, C.: An investigation of narrative subtitles in online instructional video clips. Dist. Educ. China **2015**(12), 61–66 (2015)
32. Wang, C.: Are subtitles necessary in online learning video? The modification of redundant effect. e-Educ. Res. **37**(3), 59–65 (2016)
33. Wang, D.: The influence of different presentation styles of subtitles on foreign language learners' video comprehension. e-Educ. Res. **33**(5), 85–90 (2012). https://doi.org/10.13811/j.cnki.eer.2012.05.001
34. Wang, F., Duan, Z., Zhou, Z., Chen, J.: The spatial contiguity effect in multimedia learning: the role of cueing. Acta Psychol. Sin. **47**(2), 224–233 (2015). https://doi.org/10.3724/SP.J.1041.2015.00224
35. Wang, J., Hao, Y., Lu, J.: An experimental study on the effect of teaching video presentation on autonomous learning. e-Educ. Res. **35**(3), 93–99 (2014). https://doi.org/10.13811/j.cnki.eer.2014.03.016
36. Wang, X.: Research of eye-tracking experiment in multimedia learning. Res. Explor. Lab. **34**(3), 190–193, 201 (2015)
37. Wang, X., Wang, Z., Hou, A.: The eye movement study on the design of subtitles in network teaching videos. Mod. Educ. Technol. **26**(2), 45–51 (2016)
38. Wang, Y., Xie, H., Wang, F., An, J., Hao, Y.: Text-picture integration in multimedia learning: a meta-analysis of the spatial contiguity effect. Psychol. Dev. Educ. **32**(5), 565–578 (2016)
39. Wang, Z.: Experimental study on multimedia subtitle display technology. China Educ. Technol. **2003**(7), 83–85 (2003)

40. Wang, Z., Feng, X.: The research focus and development trend of multimedia learning research in China in recent years. Mod. Educ. Manag. **2017**(9), 57–63 (2017). https://doi.org/10.16697/j.cnki.xdjygl.2017.09.011
41. Yang, Y., Mi, T., Tang, C.: A study of information presentation approach in multimedia foreign language teaching based on the cognitive theory of multimedia learning. Technol. Enhanced Foreign Lang. Educ. **130**, 42–46 (2009)
42. Yao, B.: A review of studies on the effects of captions on second language acquisition. J. Lanzhou Jiaotong Univ. **32**(2), 87–90 (2013)
43. Zhao, X., Yang, X., Li, X.: The brainwave experimental study of the effect of subtitle's presentation modes of m-learning courseware on the attention. Mod. Dist. Educ. Res. **2017**(1), 95–104 (2017)
44. Zhou, X., Zhang, W.: Caption design in the tv textbook. China Educ. Technol. Equip. **2006**(3), 27–29 (2006)

Cognition and Design

System Performance and Empathetic Design Enhance User Experience for Fault Diagnosis Expert System

Li Lin[1,2], Xiaoyue Pang[3], Jingyu Zhang[1,2], Xianghong Sun[1,2], and Liang Zhang[1,2](✉)

[1] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
zhangl@psych.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Siemens Ltd. China, Beijing 100102, China

Abstract. The fault diagnosis task in factories is technically demanding and time-consuming, especially for the green hands. In order to assist mechanicians to detect the fault parts, the fault diagnosis expert system is introduced to support decision-making and problem-solving processes. In this vision, how to enhance mechanicians' willingness to use such a system becomes a key issue to the industry. The present study aims to find out the elements that affect the user experience of the fault diagnosis expert system. We carried out an experiment to investigate how system performance and empathetic design promotes user experience, and we also explored how those factors apply to different users with high and low expertise. Fourteen participants who majored in automatic control with a basic understanding of industrial automation devices took part in the study. During the experiment, the participants were instructed to repair a UV control cabinet with the help of the fault diagnosis expert system. The system has four levels of system performance (high, medium, low, and all-errors) * two levels of empathetic feature (standard design vs. empathetic design). After each task, the participants filled out a questionnaire to evaluate their experience. The results showed a beneficial effect of both system performance and empathetic design. Also, the interactive effect indicates that for the system with qualified performance, the empathetic design enhanced user experience; while for the system with all-error performance, the empathetic design no longer helped. In addition, the empathetic design played a better role among inexperienced users, while experienced users tended to be more sensitive to changes in system performance.

Keywords: System performance · Empathetic design · User experience · User trust · Fault diagnosis · Expert system

1 Introduction

In factories, there are a series of complex technical tasks. For example, when certain machine equipment breaks down, the mechanicians have to diagnose which specific

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 357–367, 2021.
https://doi.org/10.1007/978-3-030-77932-0_28

part causes the failure. This fault diagnosis task is technically demanding and time-consuming, especially for the green hands. The introduction of the fault diagnosis expert system is aimed to assist mechanicians to detect the fault parts and solve them more efficiently [1, 2].

The main idea underlying a fault diagnosis expert system is that the system would learn the knowledge from the experienced mechanicians and support the decision-making and problem-solving process for the newcomers. Therefore, the fault diagnosis system grows faster when more mechanicians are using it and providing feedback. One of the key problems in developing such a system is how to enhance mechanicians' willingness to use such a system.

One essential factor in user experience is the performance or accuracy of the system. System performance can affect users' acceptance and trust. For example, an early study has found that the trust of operators in machines is affected by system performance. Specifically, the decline in system performance reduces the operator's trust in the machine. Even if the system performance is restored immediately, it takes longer time for the operators to restore the trust in the machine [3]. In addition, recent research has also found that the accuracy of the recommendation results will affect the user experience of the system, such as perceived usefulness and convenience [4].

Another possible solution is to take the design into consideration. In psychological studies, empathy is an important component of altruistic behavior [5, 6]. In addition, when an inanimate object resembles a human to some extent, humans will also empathize with it. In other words, by including human characteristics, a system also stands a chance to arouse empathy among users. And accumulated evidence from previous studies has shown that anthropomorphism provides effective means for inducing empathy [7, 8]. For instance, a recent study has shown that the adoption of virtual animated character stimulated the empathy of users and enhanced their willingness to help a food recognition application [6]. However, as far as we know, it still remains unknown whether empathy manipulation also benefits user experience in industrial scenarios.

Finally, the individual indifference in users themselves should never be ignored when investigating user experience. For inexperienced users, the information provided by the system may be relatively more valuable; while for individuals with rich maintenance experience, it is easier to carry out the diagnosis fault task by themselves, so the use of the system may increase their unnecessary workload [9]. Therefore, users with maintenance experience may be less affected by system differences in design.

The present study aims to find out the elements that affect the user experience of the fault diagnosis expert system, so as to provide users with a better experience and to earn trust from the users. We carried out an experiment to investigate how empathetic design benefits user experience under various conditions of system performance, and we also explored how those factors apply to different users with high and low expertise. Based on the results of previous studies, we hypothesized that both system performance and empathetic design would affect the user experience of the fault diagnosis system. Also, their effect would be modulated by the users' expertise in maintenance.

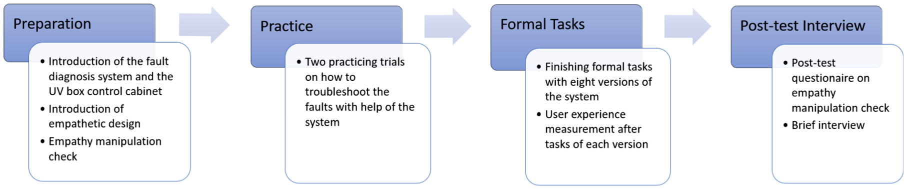
2 Methods

2.1 Participants

Fourteen participants (two females), aged 20–32 (Mean = 24), were recruited online in this study. All of them were undergraduate or graduate students majoring in automatic control. They had a basic understanding of relays, PLCs, and automation devices, and were familiar with industrial circuits. Half of the participants had prior experience in mechanical or electrical maintenance. The experiment lasted approximately three hours. All participants gave their informed consent and got compensation of 200 RMB (1 dollar ≈ 7 RMB).

2.2 Procedure

The procedure of the experiment is shown in Fig. 1. During preparation, the participants were introduced to the interfaces of the fault diagnosis system (see Sect. 2.3 Experimental Platform) and the mechanics of the equipment needs troubleshooting, namely, the UV box control cabinet. Also, the participants were shown the empathetic design versus the normal system, and the participants rated their empathetic feeling toward those two system designs. In the practice phase, the participants completed two practicing trials on how to troubleshoot the faults of the UV box control cabinet with help of the fault diagnosis system. Then the participants finished the formal troubleshooting tasks with eight versions of the system (see Sect. 2.4 Experiment design). After experiencing each version of the system, the participants filled out questionnaires to measure the user experience (see Sect. 2.5 Measurements). Finally, the participants completed a post-test empathy manipulation check and received a brief interview from the experimenters. The interview questions included how they valued the system performance and empathetic design of the fault diagnosis system, and what other approaches they think could improve their willingness to use the system in a real scenario.


Preparation

- Introduction of the fault diagnosis system and the UV box control cabinet
- Introduction of empathetic design
- Empathy manipulation check

Practice

- Two practicing trials on how to troubleshoot the faults with help of the system

Formal Tasks

- Finishing formal tasks with eight versions of the system
- User experience measurement after tasks of each version

Post-test Interview

- Post-test questionaire on empathy manipulation check
- Brief interview

Fig. 1. The experiment procedure.

2.3 Experimental Platform

In this study, we developed an interactive prototype of the fault diagnosis expert system in different versions in aspects of empathetic features and system performance. The system consists of five main user interfaces as follows.

a) *Welcome and login interface*: The user is welcome to log in, and the user can enter the system after being identified on this interface.
b) *Fault selection interface*: After entering the system, the user can select the fault machine and the fault phenomenon. After choosing the fault phenomenon, the users click "Next" to the next step.
c) *Troubleshooting interface*: The system provides possible reasons and suggestions based on the fault phenomenon selected by the user, and the number of "stars" is used as the recommendation level. After the user clicks on the item that has been checked, this item will enter the checked problem interface (on the right side of the troubleshooting interface). At the same time, the system will update the troubleshooting suggestions based on the user's click interaction.
d) *Checked problem interface*: This interface will show all the problems that have been checked. After inspection to determine the cause of the failure, users can select the real cause of the failure and click "Generate Report" to complete the troubleshooting. If all suggested items are not the real cause, the users can also click "Add" to add other items inspected.
e) *Report interface*: In the report interface, the user can review the maintenance- related information, including the maintenance report and the maintenance process.

During the experiment, the participants were instructed to use the fault diagnosis system to repair a UV box control cabinet. The UV box control cabinet showed one failure in each trial. These failures involved different parts that need to be checked, including the power adapter of the light, PLC, touch screen, PLC output cable, intermediate relay, relay cable, PT100 sensor, irradiance meter, fan power supply, and the light bulb. All the troubleshooting tasks were generated by the expert in charge of the UV box control cabinet to keep the difficulty and complexity similar among tasks. Due to the covid-19, the experiment was conducted online finally, so we developed a simple program to simulate the failures of the UV box control cabinet. In each task, there was a recorded video clip briefly describing the equipment's problems.

2.4 Experiment Design

The experiment was a within-subject design with two independent variables (system performance × empathetic design). There were four levels of system performance from high to all-error. For high performance, the real cause of the failure was always ranked at the first or second place in the suggestion list. For medium performance, the third or fourth position. For low performance, the fifth or sixth position. As for the all-error performance condition, there was no correct option provided in the suggestion list. All four levels of performance were combined with a normal design and empathetic design. The differences between those two designs included *the Welcome and login interface* and *Checked problem interface.* In the *Welcome and login interface*, the empathetic design had an anthropomorphic identity called "Xiaolan (Little Blue)" and provided a thoughtful functional introduction of itself; while the normal design only had the text "Welcome". In the *Checked problem interface*, the empathetic design provided an explanation when encouraging the participants to add other troubleshooting steps they have tried but not listed in the suggestions, indicating it could "help Xiaolan make faster progress"; while the normal system only prompted “Please add other troubleshooting steps”. There are eight (4 × 2) combination versions of system performance and empathetic design in total. Each participant had to use all eight versions to complete the diagnosis tasks. The test sequence of the empathetic design and the normal system is balanced among the participants.

2.5 Measurements

Empathy Manipulation Check. In order to check whether the empathy manipulation was successful, empathy manipulation checks were carried out before and after the experiment. The adapted empathy manipulation check measurement consisted of 4 items [6]. The participants rated their empathy toward the system on a 5-point scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

Perceived Performance. Perceived performance measured the participants’ perception of the system’s performance during the troubleshooting task. This measurement consisted of 4 items, which were adapted from the perceived usefulness subscale of the technology acceptance model questionnaire [10]. The participants evaluated on a 7-point scale (1 represents strongly disagree to 7 strongly agree).

Trust. The measurement of the participants’ trust in the system is adapted from the two subscales of the human-computer trust scale proposed by Madsen & Gregor [11]. The two dimensions were Perceived Technical Competency (PTC) and Personal Attachment (PA). PTC measured how participants evaluated the system in the core technical competency of providing useful troubleshooting suggestions. There were 3 items in the PTC subscale, such as “The fault diagnosis system has built-in reliable knowledge about equipment failures”. The PA measured the participants’ attachment to the system during the troubleshooting tasks. The PA subscale consisted of 3 items, e.g. “I personally prefer to use this fault diagnosis system to troubleshoot”. All items were rated on a 7-point scale (1 = strongly disagree to 7 = strongly agree).

Pleasure. The Self-Assessment Manikin (SAM) valence subscale was used to measure emotional pleasure [12]. Participants were asked to rate the pleasure during the task on a 9-point scale, ranging from1 (feeling unhappy and dissatisfied) to 9 (feeling happy and satisfied).

2.6 Data Analysis

First of all, we checked the success of empathetic manipulation by carrying out paired-sample t-tests between empathetic designa and normal design. Then, in order to investigate the impact of the system performance and empathetic design on the user experience, and meanwhile to consider the role of users’ expertise in maintenance, we conducted a series of repeated measure ANOVAs (System performance × Empathetic design × User expertise), with System performance and Empathetic design as within-subject variables, and User expertise as between-subject variable. The dependent variables are perceived performance, trust (PTC and PA), and pleasure.

3 Results

3.1 Empathetic Manipulation Check

Thirteen subjects completed the empathetic manipulation checks both before and after the tasks. The paired-sample t-tests (see Table 1) showed that participants' empathetic feelings toward the empathetic design and normal system were significantly different both before the tasks ($t(12) = 3.828, p = 0.002$) and in the end of the experiment ($t(12) = 2.347, p = 0.037$).

Table 1. The empathetic manipulation checks

	Mean (SD)		T-test	
	Empathetic	Normal	t	p
Pre-test	3.33 (0.86)	2.29 (0.71)	3.828	0.002^{**}
Post-test	3.19 (0.96)	2.33 (0.83)	2.347	0.037^{*}

$^{*}p < 0.05$, $^{**}p < 0.01$

3.2 Perceived Performance

The repeated measure ANOVA (System performance × Empathetic design × User expertise) on perceived performance revealed a significant main effect of System performance ($F(3, 33) = 34.053, p < 0.001$, partial $\eta^2 = 0.756$). Post-hoc analysis showed that the perceived performance of the all-error system was significantly lower than the other three systems ($ps < 0.001$), but there was no significant difference among the three systems of high, medium, and low performance levels. Empathetic design also enhanced the perceive performance ($F(1, 11) = 5.336, p = 0.041$, partial $\eta^2 = 0.327$). In addition, there was a significant interactive effect between System performance × Empathetic design, $F(3, 33) = 3.879, p = 0.018$, partial $\eta^2 = 0.261$. When the System performance was at a high level, the Empathy design improved the user's perception of the system capacity, but when the System performance is at a low level or completely wrong, the Empathy design had no effect (see Fig. 2. Left panel).

We also observed a significant interactive effect on perceived performance between Empathetic design and User expertise, $F(1, 11) = 6.587, p = 0.026$, partial $\eta^2 = 0.375$. The Empathetic design enhanced the perceived performance for the inexperienced users but not the experienced users (see Fig. 2. Right panel).

3.3 Trust

Perceived Technical Competency (PTC). The repeated measure ANOVA (System performance × Empathetic design × User expertise) on PTC showed a significant main

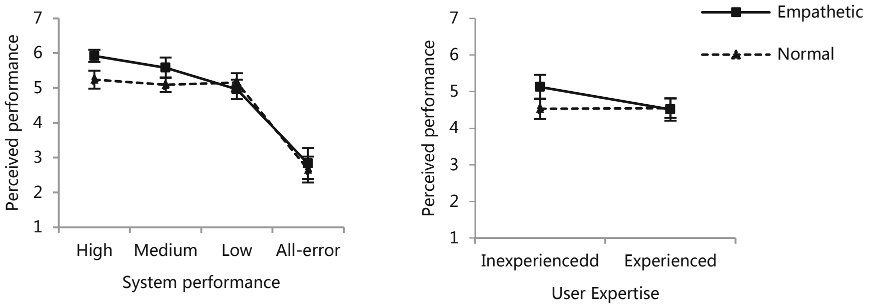


Fig. 2. The effect of System performance and Empathetic design on perceived performance.

effect of System performance ($F(3, 33) = 25.802$, $p < 0.001$, partial $\eta^2 = 0.701$). Similarly, post-hoc analysis showed that the PTC of the all-error system was significantly lower than other three systems ($ps < 0.001$), in the absence of significant difference among the other three levels from high to low. We also found a significant Empathetic design × User expertise interactive effect on PTC, $F(1, 11) = 6.628$, $p = 0.026$, partial $\eta^2 = 0.376$ (see Fig. 3. Right panel). Moreover, there was a trend of System performance × User expertise interactive effect on PTC, $F(3, 33) = 2.553$, $p = 0.072$, partial $\eta^2 = 0.188$. The inexperienced participants rated high on PTC for all system performance levels from high to low (that is, when the system can finally provide the correct troubleshooting items), but rated extremely low for the all-error system. However, the PTC of experienced participants decreased proportionally with the decline of system performance (see Fig. 3. Left panel).

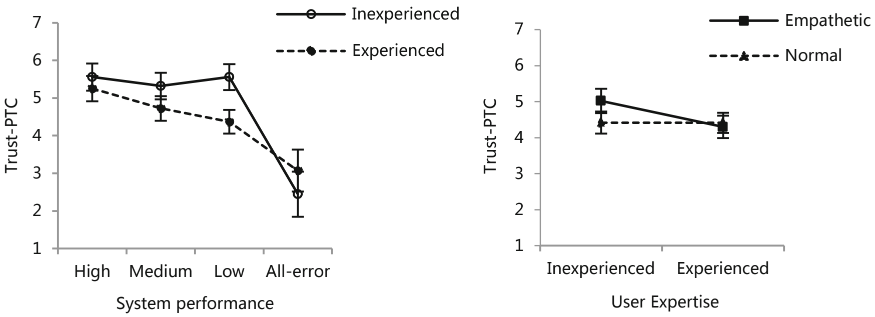


Fig. 3. The effect of System performance and Empathetic design on PTC.

Personal Attachment (PA). The repeated measure ANOVA (System performance × Empathetic design × User expertise) showed the main effect of System performance on PA, $F(3, 33) = 29.173$, $p < 0.001$, partial $\eta^2 = 0.726$. Again, only the PA of the all-error system was significantly different from the other three levels ($ps < 0.001$). There was also a trend that participants held higher PA toward empathetic design than normal design, $F(1, 11) = 4.764, p = 0.052$, partial $\eta^2 = 0.302$. In addition, the System performance × Empathetic design interactive effect was significant, $F(3, 33) = 3.356$, $p = 0.030$, partial $\eta^2 = 0.234$, with Empathetic design only enhancing PA when the System performance was relatively high (see Fig. 4). However, there were no interactive effects between User expertise and other variables for PA (*ps* > 0.1).

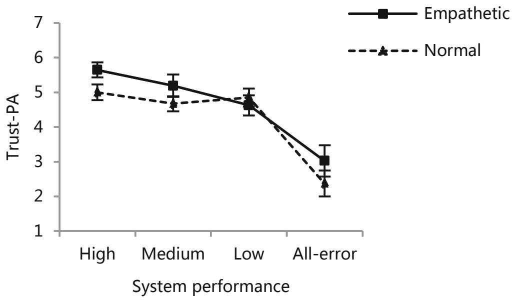


Fig. 4. The effect of System performance and Empathetic design on PA.

3.4 Pleasure

For pleasure, the repeated measure ANOVA (System performance × Empathetic design × User expertise) also revealed a main effect of System performance ($F(3, 33) = 32.580$, $p < 0.001$, partial $\eta^2 = 0.748$), with all-error level significantly lower than all other three (*ps* < 0.001). The ANOVA also showed the main effect of Empathetic design on emotional pleasure ($F(1, 11) = 6.890, p = 0.024$, partial $\eta^2 = 0.385$), with Empathetic design bringing more pleasure than normal design. No significant interactive effects were found for pleasure.

4 Discussion

The present study examined the interactive effect of system performance and empathetic design on user experience. Meanwhile, the modulating effect of user difference in maintenance expertise was also investigated. Participants who majored in automatic control used eight versions of fault diagnosis systems to troubleshoot a UV box control cabinet. They evaluated their perceived performance, trust, and pleasure immediately after experiencing each version of the system. As predicted, the system performance is an essential factor regarding all the evaluated dimensions of user experience. Also, the empathetic design enhanced most aspects of user experience in the fault diagnosis system. Furthermore, the interaction effects among system performance, empathetic design, and user expertise were observed in our study and would be further discussed.

Although system performance impacts all dimensions of user experience, we should notice that the perceived performance did not change proportionally with the real system performance we manipulated. Only the all-error system is rated extremely low compared to the other three levels. However, no obvious decline in the evaluation was observed when the system changed from high performance (correct item ranked at the first or second place) to low performance level (correct item ranked at the fifth or sixth place). Consistently, previous research on the system performance also reported an indirect relationship between objective performance and subjective user evaluation [13, 14]. However, there is a study that found a linear relationship between system performance and subjective user evaluation [15]. This inconsistency in findings may come from the different approaches of evaluation. In the study which reported a linear relationship, the participants were allowed to modify their evaluation of different systems before the last evaluation, which would enable users to make more relative judgments when they looked back. In contrast, in our study, the participants had to evaluate immediately after the use of each system, without the opportunity of post-modification.

The interactive effect between system performance and user experience on personal attachment should also be noticed. While the inexperienced participants showed no preference among the high to low level systems, the experienced participants were more sensitive to the change in system performance.

In addition, our study also observed a positive effect of empathetic design on perceived performance, personal attachment, and pleasure. Therefore, when the system performance is on a plateau and unable to be upgraded at a certain level, the empathetic design would serve as a more economical approach to improve user experience. However, it is worth noting that the interactive effect between system performance and empathetic design reveals that the trick of empathetic design can only work when the system performance is at a qualified level. In other words, an all-error system would never offer users with good experience no matter how fancy the empathetic design is.

Another interesting finding is the interaction between empathetic design and user expertise. The results of this study indicated that not all users buy into the empathetic design in dimensions of perceived performance and perceived technical competency. While the inexperienced users rated a higher score for empathetic design, the experienced users pay more attention to the system performance itself. This result is also reflected in our interview after the tasks. Those participants with previous maintenance experience said they would not be affected by the empathetic design when they made the judgment on the performance of the system, instead they weighted more on the logic behind the suggested items. Even so, the positive side is that in terms of pleasure, the empathetic design enhanced the evaluation of all users regardless of their maintenance expertise.

In conclusion, the results of the present study indicate important effects of system performance and empathetic design on the fault diagnosis system, and also reveal the modulating effect of users' expertise. First of all, the system permeance always matters, both for cognitive and emotional evaluations. The positive perspective is that as long as the system can provide a correct answer in the end, the user experience won't be too disappointing. Secondly, empathetic design is a promising feature to contribute to the user experience, but the premise is that the system performance is at least qualified. It doesn't mean that a system needs to be perfect before deployment. When the system reaches a qualified standard, some design features such as empathetic design would help to enhance the user experience. Lastly, the effect of system performance and empathetic design both interact with users' own characteristics. The trick of empathetic design works out better for users with no prior maintenance experience, while the experienced users tend to be more sensitive to changes in system performance. Further research with more resources could expand the samples and conduct the experiment in factories to verify whether the effects of system performance and empathetic design are still effective in a real working scenario.

Acknowledgment. This research was supported by the National Key Research and Development Plan (Grant No. 2018YFC0831001), and the National Science Foundation of China (U1736220). We are particularly grateful to the support of the Siemens-CAS program "Research in Human-Autonomous System Incorporating with Knowledge." The authors thank Shimeng Xiao, Fanke Meng, Nemin Hu, Yuqing Meng, and Zimo Fan for their contribution to the data collection.

References

1. Angeli, C.: Online expert systems for fault diagnosis in technical processes. Expert. Syst. **25**(2), 115–132 (2008)
2. Verbruggen, H., Tzafestas, S., Zanni, E.: Knowledge-based fault diagnosis of technological systems. In: Tzafestas, S.G., Verbruggen, H.B. (eds.) Artificial Intelligence in Industrial Decision Making, Control and Automation, pp. 449–506. Springer, Dordrecht (1995). https://doi.org/10.1007/978-94-011-0305-3_15
3. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10), 1243–1270 (1992). https://doi.org/10.1080/00140139208967392
4. Shin, D.: How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. Comput. Hum. Behav. **109**, 106344 (2020). https://doi.org/10.1016/j.chb.2020.106344
5. Eisenberg, N., Miller, P.A.: The relation of empathy to prosocial and related behaviors. Psychol. Bull. **101**(1), 91 (1987)
6. Lee, Y., Ha, M., Kwon, S., Shim, Y., Kim, J.: Egoistic and altruistic motivation: how to induce users' willingness to help for imperfect AI. Comput. Hum. Behav. **101**, 180–196 (2019). https://doi.org/10.1016/j.chb.2019.06.009
7. Shin, D.: Empathy and embodied experience in virtual environment: to what extent can virtual reality stimulate empathy and embodied experience? Comput. Hum. Behav. **78**, 64–73 (2018)
8. Lee, J.M., Baek, J., Ju, D.Y.: Anthropomorphic design: emotional perception for deformable object. Front. Psychol. **9**(1829) (2018). https://doi.org/10.3389/fpsyg.2018.01829
9. Ko, D.G., Dennis, A.R.: Profiting from knowledge management: the impact of time and experience. Inf. Syst. Res. **22**(1), 134–152 (2011). https://doi.org/10.1287/isre.1090.0247
10. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag. Sci. **46**(2), 186–204 (2000)
11. Madsen, M., Gregor, S.: Measuring human-computer trust. In: 11th Australasian Conference on Information Systems, vol. 53, pp. 6–8. Australasian Association for Information Systems, Brisbane (2000)
12. Lang, P.J.: The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders. Lawrence Frlbaum, Hillsdale (1985)
13. Al-Maskari, A., Sanderson, M., Clough, P.: The relationship between IR effectiveness measures and user satisfaction. In: Proceedings of the 30th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 773–774 (2007)

14. Huffman, S.B., Hochster, M.: How well does result relevance predict session satisfaction? In: Proceedings of 30th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 567–573 (2007)
15. Kelly, D., Xin, F., Shah, C.: Effects of position and number of relevant documents retrieved on users' evaluations of system performance. ACM Trans. Inf. Syst. **28**(2), 1–29 (2010). https://doi.org/10.1145/1740592.1740597

The Effect of Diagonal Bar Position in Prohibition Signs on Recognition Efficiency of the Signs

Meng Liu, Xiangling Zhuang(✉), Yuhan Huang, Huiling Yang, and Guojie Ma

Shaanxi Key Laboratory of Behavior and Cognitive Neuroscience, School of Psychology,
Shaanxi Normal University, Xi'an, China
zhuangxl@snnu.edu.cn

Abstract. Prohibition signs warn people of certain behaviors to avoid danger or risk, so they should be carefully designed. ISO regulated that prohibition signs should be designed with the diagonal bar as foreground of the pictogram. In China, although general guidelines for prohibition signs follow this rule, the guidelines for prohibition traffic signs adopt an opposite design with the bar as background. This study attempts to evaluate the effect of bar position on the recognition efficiency of prohibition signs from the perspective of warning processing. Using paired prohibition signs that had opposite bar positions and controlled familiarity, we conducted the experiment with a within-subject design. In the experiment, participants judged whether the pictures of prohibition signs matched the textual explanation when they were simultaneously displayed on a blank screen. Considering the differences between different participants and signs, we modelled reaction time with a mixed-effects linear model containing random intercepts. The results showed that participants recognized familiar signs more quickly, and for familiar signs, participants recognized the signs with bar as foreground faster than the signs with bar as background. In practice, the findings suggest designing the prohibition signs with bar as foreground to improve the speed of signs recognition.

Keywords: Prohibition sign · Sign recognition · Warning processing · Bar position

1 Introduction

1.1 Design of bar in Prohibition Signs

Safety signs, especially prohibition signs are very important to our daily life, as they provide crucial information in potentially dangerous scenarios. Prohibition signs are used to warn staff and visitors that certain behaviors or actions are not allowed [1]. The International Standardization Organization (ISO) claims that prohibition sign should always consist of a red circle and a red diagonal on a white background with a black standardized pictogram.

But there are some inconsistent specifications on the design of the bar in China. The Graphical Symbols Safety Colours and Safety Signs of China (in Chinese) [2] regulated

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 368–376, 2021.
https://doi.org/10.1007/978-3-030-77932-0_29

that prohibition signs must position the diagonal bar as foreground, while the Code for Layout of Urban Road Traffic Sign and Markings of China (in Chinese) [3] regulated that the pictogram should be placed as foreground in prohibition traffic signs (see Fig. 1). That is, in China prohibition signs with bar as foreground were usually used in daily life, but specially in traffic setting prohibition signs with bar as background were used more often. There comes the question: Should the diagonal bar placed as foreground or background from human factors perspective?

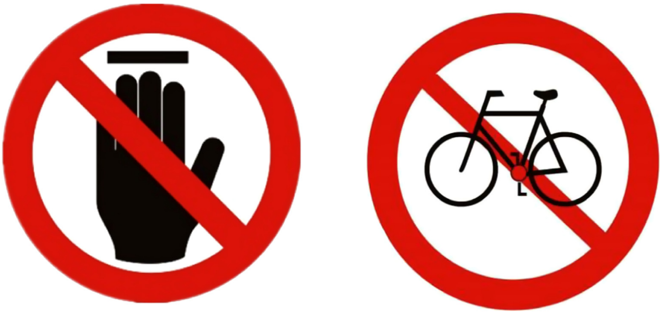


Fig. 1. Prohibition sign with bar as foreground (left) and prohibition sign with bar as background (right).

Previous findings were inconsistent on the role of bar position in prohibition signs. Some researcher thought that bar as foreground obscured the pictogram and made signs recognition more difficult [4, 5]. There were also researchers find out there were no differences between bar as foreground and bar as background [6, 7]. However, most of the previous tasks and measurements focused on the accuracy in sign recognition, while speed of sign recognition was neglected. Prohibition signs are special in that people need to react fast to these signs beyond simple recognition. The more quickly people can recognize signs, the more time they have for hazard information processing. Therefore, we focus on speed of sign recognition in this study.

1.2 Evaluation of Bar Position Based on Warning Process

To evaluate the two types of prohibition signs, we need to resort to how people process warning signs. Rogers et al. [8] have summarized the framework of warning process included noticing, encoding, comprehending and complying stage, and they have proposed several person variables and warning variables influencing the warning process at these stages. Noticing stage means to be attracted to the sign. Encoding stage involves translation from the external sign to the internal representation. The comprehending and complying stage were more likely to be influenced by the content of the sign and other personal variables (e.g. ability, attitude) rather than secondary features like the position of the bar. As a result, the effect of bar position may mainly affect noticing and encoding stages. For now, we will focus on evaluating the two types of prohibition signs on these early stages.

At the noticing stage, Rogers et al. [8] took color as one of the important warning variables that influenced whether people noticed the warning. Compared with prohibition signs with bar as background, signs with red diagonal bar as foreground have more red areas. Red stimuli have been shown to receive an attentional advantage [9] and can promote the efficiency of visual search [10]. Researchers have done many works on visual search task of colored stimuli, most research have found out that compared with other colors, red stimuli can be found easier [11, 12]. Electrophysiological research has shown that red had special arousal on some ERP components which suggest particular selectivity for red and attention may be deployed to red more quickly than to other colours [12]. Therefore, the prohibition signs with bar as foreground, which have large red center areas, are more likely to capture attention and increase alertness and physiological arousal.

Encoding of a sign mainly refers to translating the information of signs into internal representation. The graph coding category model proposed by Reinitz et al. [13] pointed out that category information can improve the speed of acquiring graph information. If the category of a graph is known, the speed of obtaining information from the graph is faster than that from the graph of unknown category. For prohibition signs, “prohibition” is one of its categories, and the red circle and the red diagonal bar usually represent prohibition, which can be processed as the category clue of the sign. As soon as the bar appear, people will immediately perceive that the graph is conveying the information of “prohibition”, allowing them to extract further details more quickly. That is, for the signs with bar as foreground, the circle and bar are shown as the emphasis of categories of the prohibition signs, which may lead to faster encoding of the signs.

Based on the above, this study assumes that people would react faster while recognizing the prohibition signs with diagonal bar as foreground, and this effect is stronger in the familiar signs.

2 Methods

2.1 Experimental Design

The experiment has a 2 × 5 within-subject design. The two independent variables are position of bar (bar as foreground vs. bar as background) and familiarity (five levels from 1: “I have never seen this sign before” to 5: “I often see this sign”). The dependent variable is the speed of sign recognition. It is measured by reaction time from the time a sign and its interpretation are presented to the moment participants response.

2.2 Participants

There were 54 undergraduate students with normal or corrected-to-normal vision from Shaanxi Normal University participating this study. Data from five participants were excluded due to more than 20% of missing data. The final valid data come from 49 participants, including 14 males and 35 females aged from 17–27 (M = 19.04, SD = 1.6). All participants were compensated for their participation. Informed consent was obtained from each participant.

2.3 Materials

We selected 38 prohibition signs from Graphical Symbols Safety Colors and Safety Signs of China (in Chinese) [2]. The sign pictures were edited with Photoshop CC 2017 to make sure that each sign had two forms: bar as foreground and bar as background (see Fig. 2). Finally, we got 38 pairs of different prohibition signs, which all had one version with bar as foreground and the other had signs with bar as background. All pictures were consistent in color, saturation and brightness.

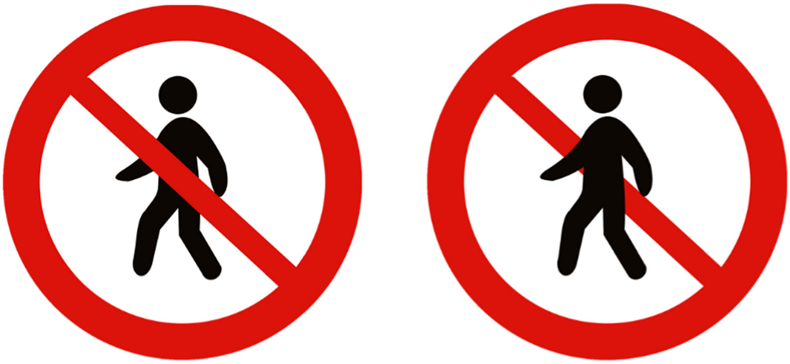


Fig. 2. Example of a pair of prohibitions signs with different bar positions: “No Entry” signs with bar as foreground (left) and bar as background (right).

2.4 Procedure

Participants sat in a well-lit and quiet room in a comfortable position. They faced a 21-in. colored display 0.8 m from them. The pictures of signs appeared on a black background in the center of the display. The diameter of each picture is 10 cm, with a visual angle of 7.2° × 7.2°.

We used E-Prime 2.0 to present stimulus and record the response data. The procedure was demonstrated in Fig. 3. In each trial, a fixation was presented in the center of the screen for 500 ms, followed by a picture of prohibition sign displayed in the center of the screen and its corresponding textual explanations at the bottom for 5 s. Participants were asked to determine whether the meaning of the sign was consistent with the presented text as quickly as they can. Participants pressed ***J*** if they thought that words and meaning of the sign matched, otherwise pressed ***F***. Their reaction time and accuracy were recorded. After participants pressed the required keys, the picture and the words disappeared immediately, followed by a blank screen for 250 ms. If participants did not press the keyboard, a blank screen would be presented 5 s later and the next trial began. Each sign appeared twice, once with the correct meaning and once with incorrect meaning.

Before the formal experiment, participants practiced on 3 pairs of signs presented randomly for ten times. The formal experiment included 140 random trials on the remaining 35 pairs of signs. Participants had a brief break after they finished 70 trials.

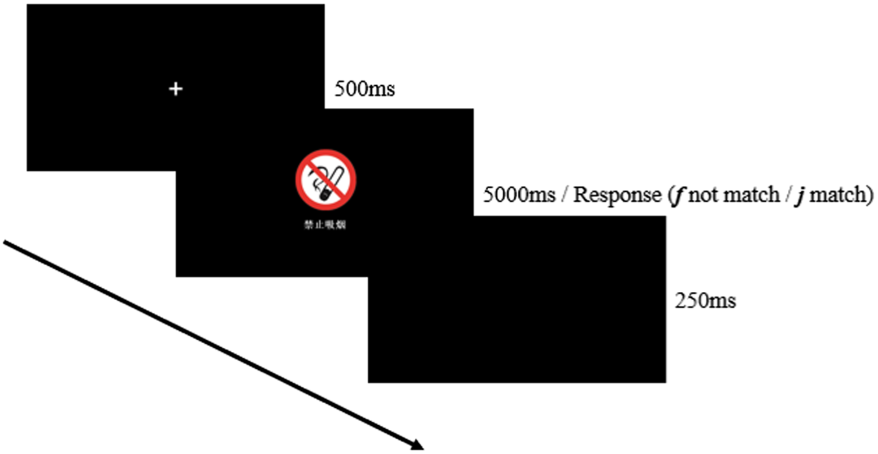


Fig. 3. Procedure of the experiment. The stimulus of this trial is the prohibition sign "No Smoking" and the textual explanations at the bottom is "No Smoking" in Chinese, which is the correct meaning.

After the recognition task was finished, the standard signs were presented on the screen one by one with their correct meanings. Participants were asked to evaluate the familiarity with each sign on the Likert scale of 1 to 5 (1 means "I have never seen this sign before", 5 mean "I often see this sign").

2.5 Statistical Analysis Methods

Reaction time was defined as the time from the presentation of stimulus to the response of participants. We removed outliers beyond 2.5 standard deviations from the data for each participants' reaction time, 6.15% of the data was deleted.

With the remaining data, we built a linear mixed-effect model to analyze how participants' reaction time was affected by bar position and familiarity with different prohibition signs. We modeled individual differences by assuming different random intercepts for each participant [14]. We used R [15] and its lme4 package to estimate all the models [16].

3 Results

The descriptive statistics of accuracy and reaction time are shown in Fig. 4. It shows that the signs with bar as foreground led to higher accuracy and less reaction time than signs with bar as background.

To further analyze how bar position and familiarity affected reaction time, we built two models. We first put the bar position, familiarity of signs and the interaction in the initial model as fixed effect and considered random intercept for participant to account for the interdependent response from the same participant (Model 1). The core syntax is: *Reaction time ~ bar position * familiarity + (1|participant).*

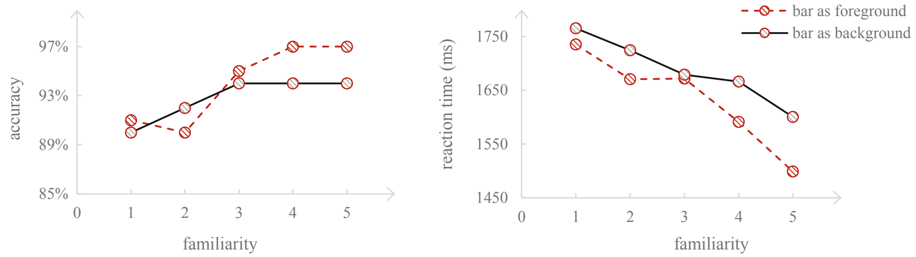


Fig. 4. Accuracy (left) and average reaction time (right) of sign recognition of all responses. Familiarity: 1 means "I have never seen this sign before", 5 means "I often see this sign".

Similarly, given that the reaction time for the same prohibition signs were also interdependent, we add a random effect for signs in our model to account for the different random intercepts for each sign as we did for each participant. We then built another model (Model 2). The core syntax is: *Reaction time ~ bar position * familiarity + (1|participant) + (1|sign).*

Considering the speed-accuracy trade-off, we use the data of correct responses to test and compare the two models, and found out that Model 2 was significantly better than Model 1 (χ^2 (1) = 229.53, $p < 0.001$). That means the random effect of different signs should be included in the interpretation. Therefore, we finally chose Model 2 to interpret the association between reaction time, familiarity and bar position. The results are shown in Table 1.

Table 1. Fixed effect of the linear mixed models.

	Model 1 (with random effect of participants)		Model 2 -Final (with random effects of participants and signs)	
	Estimated Value (SE)	*t* value	Estimated Value (SE)	*t* value
Intercept	1810.1 (41.01)	44.14***	1766.2 (47.64)	37.07***
Bar position (bar as background)	-17.7 (29.60)	-0.60	-12.6 (28.85)	-0.44
Familiarity with signs	-61.2 (7.11)	-8.60***	-42.7 (8.33)	-5.13***
Bar position (bar as background) × familiarity with signs	22.0 (9.87)	2.23*	20.8 (9.62)	2.16*

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The result revealed that the time participants recognized prohibition signs was not significantly different across two bar positions ($t = -0.44$, $p = 0.66$). Familiarity with signs significantly affected reaction time. Participants identified familiar signs more quickly ($t = -5.13$, $p < 0.001$).

The interaction between bar position and familiarity with signs was significant ($t = 2.16$, $p = 0.031$) (see Fig. 5).

The results indicated that for familiar prohibition signs, the signs with bar as foreground was recognized more quickly. Nevertheless, for the prohibition signs that participants did not know well, the bar position did not affect the final reaction time.

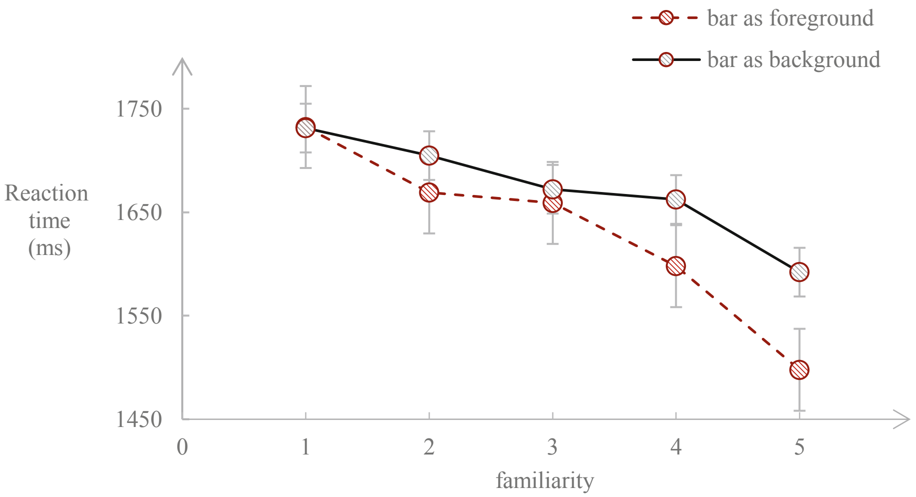


Fig. 5. Reaction time of sign recognition of correct responses. Familiarity: 1 means "I have never seen this sign before", 5 means "I often see this sign".

4 Discussion

Overall, the experiment showed when people were asked to recognize the prohibition signs, they made decisions more quickly when judging the signs with bar as foreground, and this quick reaction only occurs when they faced familiar signs.

As we reviewed in Introduction Sect. 1.2, Rogers et al. [8] divided signs processing into four stages: noticing, encoding, comprehending and complying, and summarized several person variables (such as familiarity) and warning variables (such as color) influencing the warning process at these four stages. Different designs would influence the processing of signs.

In this experiment, participants need to judge the meaning of the simple picture of prohibition signs presented on blank screen, so the time spent on searching for sign is negligible. Therefore, the final reaction time mainly reflects the process of encoding the sign, comprehending the sign and response by pressing the keyboard. In these stages, we found a quicker reaction for more familiar signs. The finding is consistent with previous studies which found that people's recognition and understanding of safety signs are related to familiarity. More familiar the signs are, greater recognition efficiency will be [17].

More importantly, we found a bar-as-foreground advantage only for familiar signs, just as we expected. Although previous study has shown that bar as foreground can obstruct processing of signs [4, 5], we think the obstruction can be diminished by strengthened category cues [13]. Bar as foreground can significantly help people elicit the category information from prohibition signs, so participants took shorter time to encode and recognize the signs with bar as foreground. Besides, for familiar signs, even the bar is placed as foreground, people can integrate the complete pictogram automatically by gestalt principle and previous experience, which may counteract the effect of obstruction. For unfamiliar signs, people don't have enough knowledge about the pictogram in signs, so bar as foreground breaks the integrity of the pictogram, and people cannot encode the whole pictogram quickly. As a result, even the category information facilitated the encoding to some extent, the final reaction time become similar with signs with bar as background.

The results of the experiment suggest that the prohibition signs with bar as foreground are better than bar as background in terms of processing efficiency. Our findings suggest using the prohibition signs with bar as foreground to improve the speed of processing signs.

Our study still has many aspects to improve. In this study, we just presented the pictures of prohibition signs on a blank screen and asked participants to judge the meaning of the signs, ignoring the scenarios where people come across the prohibition signs. In the future considering the signs recognition in complex surrounding may have different consequence with that in a laboratory.

5 Conclusion

The position of diagonal bar and familiarity both affect speed of prohibition signs recognition. People recognize the prohibition signs with bar as foreground more quickly than bar as background and this effect was stronger on familiar signs. Based on this, we suggest using prohibition signs with bar as foreground to speed up its recognition.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (31970998), and Youth Innovation Team of Shaanxi Universities.

References

1. International Standardization Organization: ISO 7010 Graphical symbols - Safety colours and safety signs - Registered safety signs (2019)
2. General Administration of Quality Supervision, Inspection and Quarantine of China, China National Standardization Management Committee: Tuxing fuhao - Anquan se he anquan biaozhi [Graphical symbols - Safety colours and safety signs of China] (2011)
3. Ministry of Public Security of China, Housing and Urban-Rural Construction Bureau of China: Chengshi daolu jiaotong biaozhi he biaoxian shezhi guifan [Code for layout of urban road traffic signs and markings of China] (2015)
4. Dewar, R.E.: The slash obscures the symbol on prohibitive traffic signs. Hum. Factors **18**, 253–258 (1976)
5. Hoffmann, E.R., Smith, G., Kalnins, C., Sanders, B.: Legibility of symbolic turn-restriction signs of various formats. In: Australian Road Research Board Conference Proceedings, pp. 127–134 (1982)
6. Murray, L.A., Magurno, A.B., Glover, B.L., Wogalter, M.S.: Prohibitive pictorials: evaluations of different circle-slash negation symbols. Int. J. Ind. Ergon. **22**, 473–482 (1998)
7. Wogalter, M.S., La Murray, T.A., Glover, B.L., Shaver, E.F.: Comprehension of different types of prohibitive safety symbols with glance exposure. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 46, pp. 1753–1757 (2002)

8. Rogers, W.A., Lamson, N., Rousseau, G.K.: Warning research: an integrative perspective. Hum. Factors **42**, 102–139 (2000)
9. Elliot, A.J.: A historically based review of empirical work on color and psychological functioning: content, methods, and recommendations for future research. Rev. Gen. Psychol. **23**, 177–200 (2018)
10. Elliot, A.J.: Color and psychological functioning: a review of theoretical and empirical work. Front. Psychol **6**, 368 (2015)
11. Lindsay, D.T., Brown, A.M., Reijnen, E., Rich, A.N., Kuzmova, Y.I., Wolfe, J.M.: Color channels, not color appearance of color categories, guide visual search for desaturated color targets. Psychol. Sci. **21**, 1208–1214 (2010)
12. Pomerleau, V.J., Fortier-Gauthier, U., Corriveau, I., Dell'Acqua, R., Jolicœur, P.: Colour-specific differences in attentional deployment for equiluminant pop-out colours: evidence from lateralized potentials. Int. J. Psychophysiol. **91**, 194–205 (2014)
13. Reinitz, M.T., Wright, E., Loftus, G.R.: Effects of semantic priming on visual encoding of pictures. J. Exp. Psychol. Gen. **118**, 280–297 (1989)
14. Winter, B: Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. Accessed 20 Aug 2020
15. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/, last accessed 2020/8/20.
16. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**, 1–48 (2015)
17. McDougall, S.J.P., Curry, M.B., de Bruijn, O.: Measuring symbol and icon characteristics: norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. Behav. Res. Methods Instrum. Comput. **31**, 487–519 (1999). https://doi.org/10.3758/BF03200730

A Survey Study of Factors Influencing Smart Phone Fluency

Xinze Liu[1,2], Weina Qu[1,2], Cheng Wang[3], Qian Zhang[1,2], and Yan Ge[1,2](✉)

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China
gey@psych.ac.cn

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

[3] School of Psychological and Cognitive Sciences, Peking University, Beijing, China

Abstract. When users interact with smart phones, the issue of fluency can influence their experience. However, there has been no previous clear definition and structure of smart phone fluency. The objective of this study is to confirm the factors influencing smart phone fluency and identify operational problems that affect smart phone fluency. A large-scale questionnaire survey was conducted, and 637 questionnaires were collected. We found that fluency is the most important feature of an easy-to-use smart phone. Moreover, the results confirmed the six factors of smart phone fluency: response delay, simplicity, operation error rate, connection, hardware, and visual experience. We also found that there were significant differences in the importance evaluation of visual experience and operation error rate among different age groups. In addition, opening the application, using the application, and opening the browser were the three operations in which people had the least fluent experiences. This study successfully confirmed the elements of smart phone fluency and contributed substantially to future smart phone design to shape a better user experience.

Keywords: Smart phone fluency · User experience · Satisfaction · Human-computer interaction

1 Introduction

Smart phones have developed at an exponential rate in recent years because they serve as important tools for communication (Dufau et al. 2011). In examining the whole industry, the focus of smart phone evaluation tends to be product-centered rather than user-centered, whether from the perspective of terminal equipment manufacturers or application developers (Si and Liu 2015). These studies focus on product design, application service quality, application optimization efficiency, etc. However, with the current relatively mature development of software and hardware technology, customers expect experiences more than utility because the technology has become a means to improve human experiences in many aspects of life (Hassenzahl 2018; Walsh et al. 2010). Thus, not only the function of the products but also the user experience should be considered. User experience is defined as a person's perceptions and responses that result from the use

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 377–388, 2021.
https://doi.org/10.1007/978-3-030-77932-0_30

or anticipated use of a product, system or service (Norman et al. 1995). It is a momentary, primarily evaluative feeling (good-bad) (Hassenzahl 2008). Therefore, a good product design should satisfy users' needs, trigger a feeling of anticipation regarding the use of the product, implicate considerations related to user experience and integrate affect into the design (Agost and Vergara 2014; Law et al. 2014; Seva et al. 2007; Seva et al. 2011; van Kuijk et al. 2015). As smart phone adoption grows rapidly, there is still much to be done to improve the user experience of smart phones.

In previous studies focusing on smart phones, researchers held different views about the dimensions of user experience. It was identified that the performance and appearance of smart phones are the most important factors for users (Si and Liu 2015). Li (2015) proposed a smart phone user experience evaluation system and gave the evaluation indexes from six aspects, including system responsiveness, stability, compatibility and power consumption. However, these two types of categorization were limited because they did not take the users' subjective attitudes and feelings into account in assessing the interactive experience but rather focused too closely on the product characteristics from the technological perspective. Another previous study indicated that the factors influencing users' smart phone experience can be divided into three categories: appearance, operation, and comprehensive satisfaction (Yan et al. 2015). Appearance includes modeling factors, material factors, and button layouts, while operation factors refer to the ease of learning and ease of use of the smart phone system. These categories are related to how easily new users can complete tasks and how convenient the experience is (Yan et al. 2015). In fact, current smart phones are relatively uniform in terms of appearance and operation, and their hardware capabilities are sufficient to meet the functional requirements. For example, at a viewing distance of approximately 30 cm, the 960 × 640 screen resolution of iPhone4 had already reached the limit of human retinal resolution.

Although the objective performance of mobile phones met the industry standard, many complaints about mobile phones could still be heard that reflected users' dissatisfaction. In the model of consumer behavior proposed by Lewis, the role of users' subjective perceptions in consumer behavior is particularly emphasized. Thus, in addition to appearance and operation, another important factor that influences user experience is customer satisfaction. Customer satisfaction refers to the degree to which the product meets a customer's needs (Shi 2005). It is a positive subjective response that is associated with aspects such as comprehensive, friendly, enjoyable, fun, efficient, comfortable, controllable, and fresh experience. In real life, negative user evaluations are often heard that use terms such as "stalling", "flash quit", "inactivity" and "slow" (Si and Liu 2015). These disadvantages, which are apparently related to "smart phone fluency", can strongly negatively affect user satisfaction and purchase aspiration.

There is no clear definition of smart phone fluency, but speed and people's perception of speed are indispensable aspects of smart phone fluency. As the hardware and data communication speed continue to develop rapidly, people may not want to be bothered by slow response times and system performance; thus, a slow system response time could result in users' dissatisfaction and discontinued use of an application (Schleifer and Iii 1989). In addition, studies on the worldwide web showed that web pages with shorter delays and faster retrieval times were judged to be more interesting than slower pages (Ramsay et al. 1998), and if delays extended to 4 s or longer, people's behavioral intentions decreased substantially (Galletta et al. 2004). Another study indicated that the instant connectivity of a device could have an impact on smart phone user satisfaction because people use smart phones not only to call other people but also to access the internet anytime and anywhere. This feature, known as instant connectivity or ubiquitous access, could affect users' mood and hence influence their satisfaction (Park and Lee 2011). All the above mentioned studies indicated that the "fluency" of the interaction process with the products was related to response time, which could play an important role in user satisfaction.

There are some related studies about fluency that could help us to understand this concept. For example, processing fluency objectively refers to the dynamic characteristics of the internal information processing process of the human brain (mainly refers to the processing speed and accuracy) (Winkielman and Cacioppo 2001) and subjectively refers to the individual's subjective experience of the difficulty of processing information (Oppenheimer 2008). In addition, Csikszentmihalyi and Larson defined flow as a mental state in which one is fully immersed in an activity (Csikszentmihalyi and Mihaly 2014). Flow occurs with a high degree of excitement and fullness. Different from the above, the smart phone fluency in the current study emphasizes the subjective feeling of the user when operating the smart phone. When we design future interactive techniques for smart phones, it is important to capture humanity and then make the smart phone no longer feel like a machine that adds burden to users but an extension of users themselves and give users pleasure so that they like to use.

According to the results of a focus group interview conducted by us earlier, fluency refers to the degree of ease with which an individual is able to perform some tasks, and it typically implies a high level of experience or practice. A previous focus group interview explored the dimensions of smart phone fluency (Liu et al. submitted) and found 6 elements: connection, response delay, nonoperational system failure, visual experience, operation error rate and simplicity. Connection refers to a visual sense of connection when switching frames or tasks (i.e., the rate of switching between applications and the stability of the system). Response delay is defined as the time interval from completing an operation to the device providing a recognizable response. Nonoperational system failure is defined as mobile breakdowns caused by hardware configuration, software and the compatibility of the two rather than faulty operations (i.e., a device does not work when the user is typing). Visual experience refers to the sharpness, distinguishability and color matching of the screen and image quality. The operation error rate is defined as the error rate when individuals operate a mobile phone, and the lapses are not caused by the users' lack of mobile-related knowledge. The aspect of simplicity is defined as the number of operations needed to complete a task. As the focus group was only a qualitative study and just on behalf a small sample of users, we designed a questionnaire to confirm factors of fluency in a larger sample and determine operation problems affecting fluency experience.

2 Method

2.1 Participants

We used a convenient sampling method to collect the data. All the questionnaires were distributed through friend circles and survey companies. All the respondents participated in this study voluntarily and anonymously. All participants were informed that all their information and the results would remain confidential and received 10 yuan as a reward. Overall, 637 valid surveys were returned. Table 1 summarizes the demographics of the respondents.

Table 1. Age distribution by gender of participants.

Age	Gender	Number (%)
Under 18 years old	Male	132 (20.7%)
	Female	115 (18.1%)
31–40 years old	Male	124 (19.5%)
	Female	106 (16.6%)
Above 41 years old	Male	82 (12.9%)
	Female	78 (12.3%)

2.2 Questionnaire

A set of questionnaires in three sections was developed for a larger survey. The first section focused on the basic information of participants and their mobile phones, including demographic questions, usage of mobile phones, features of easy-to-use smart phones and so on. The second part concerned users' attitude to the six dimensions of smart phone fluency summarized from previous focus group interview, including response delay (response speed of smart phone), operation error rate (the frequency of false touch), visual experience (the clarity, resolution and color matching of screen picture quality), connection (the connection between two tasks such as app switching and screen page turning), simplicity and convenience (the minimum number of steps required to complete a task), and system failure (failure frequency of smart phone hardware such as battery and CPU). The participants needed to evaluate the significance of each dimension on smart phone fluency using a five-point Likert scale anchored by "0 = not important" and "4 = very important". The third section focused on the specific situations that affect smartphone fluency, including two questions. First, the researchers listed 13 problems that may affect the fluency experience of smart phones, including input feedback delay, slowness of software startup, abnormal call process, restart the phone again and again, slow down after clicking, cannot find the needed function, touch screen failure and so on. Participants needed to choose three problems that most influenced smart phone fluency according to their use experience. If there are other important problems that are not listed, the users can fill in the questionnaire. Next, participants needed to evaluate the frequency of unsmooth experience in 38 common smart phone operation scenarios using a five-point Likert scale anchored by "0 = never have the unsmooth experience" and "4 = always have the fluency experience". Smart phone operation scenarios covered most of the smart phone function, including power on, navigating, opening the application, playing audio and video, browsing pictures and so on.

3 Results

3.1 Features of an Easy-to-Use Smart Phone

In this part, participants need to choose the three most important features of an easy-to-use smart phone in a list. Table 2 shows the frequency of each feature mentioned. Fluency is the most important feature of an easy-to-use smart phone. Convenient operation and sensitive touch screens are also related to fluency.

Table 2. The frequency of important features of easy-to-use smart phones.

Features	Frequency	Features	Frequency
Fluency	268	Light and thin phone body	147
Long standby time	248	Sensitive touch screen	136
Good hand feeling	224	High resolution camera	86
Beautiful interface	183	High resolution screen	69
Convenient operation	178	Fast response camera	42

3.2 The Experience of Smart Phone Fluency Among Different Age Groups

In the questionnaire, the participants rated the significance of the six dimensions on smart phone fluency using a five-point Likert scale ranging from 0 to 4, with 2 as the median. Table 3 shows that the influence of the six factors on smart phone fluency was more than moderate, indicating that the factors are indeed closely related to smart phone fluency. A GLM repeated measure was used to test the difference among each dimension. The main effect was significant ($F = 3.06$, $p < 0.01$). The pairwise comparisons showed that only the operation error rate was higher than the connection, system failure and visual experience scores.

To explore the effect of demographic variables on smart phone fluency, we compared the differences among each group. No difference was found between males and females, but the difference among age groups was significant. The Kruskal-Wallis test showed a significant difference between the age groups in terms of "operation error rate" ($p < .05$) and "visual experience" ($p < .001$). Participants 31–40 years old were significantly more concerned about the operation error rate than people under 30 years old ($p < .05$). Participants under 30 years old were significantly less concerned about visual experience than participants 31–40 years old (p < .01) and participants above 41 years old (p < .001). The descriptive results of factors of concern regarding smart phone fluency in different age groups are shown in Table 3.

Table 3. Importance of each dimension on smart phone fluency in different age groups (M ± SD).

Age	Total	Under 30	31–40	Above 41
Operation error rate	2.19(1.11)	2.03(1.13)	2.23(1.14)	2.38(1.03)
Simplicity and convenience	2.11(1.23)	1.95(1.25)	2.05(1.24)	2.44(1.13)
Response delay	2.08(1.54)	2.02 (1.62)	1.88(1.61)	2.47(1.22)
Connection	2.06(1.23)	1.92(1.31)	1.97(1.22)	2.41(1.05)
System failure	2.05(1.45)	2.35(1.24)	1.97(1.47)	2.39(1.20)
Visual experience	2.03(1.26)	1.77(1.30)	2.00(1.29)	2.49(1.01)

3.3 The Problems that Affect Smart Phone Fluency Experience

In this part, the participants were asked to choose three problems that most affected their smart phone fluency experience. The results showed that the top three were "slow down after use for a period of time", "slowness of software startup", and "touch screen failure"; see Fig. 1 for details.

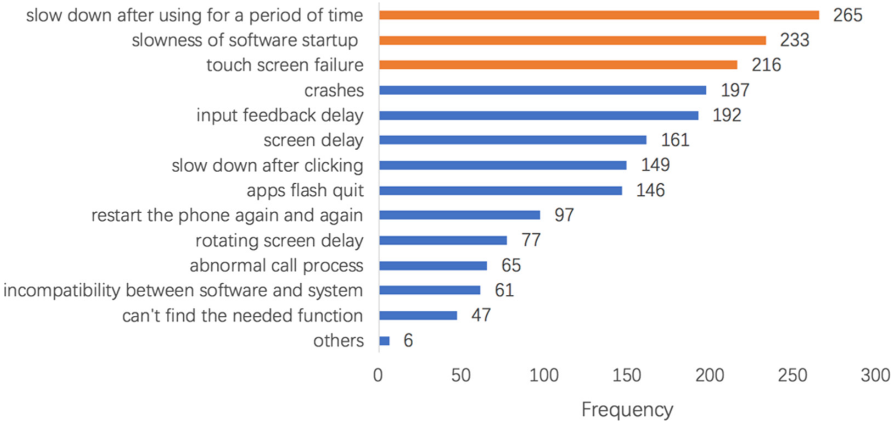


Fig. 1. Problems that most affect smart phone fluency experience.

3.4 The Unfluency Experience in Different Operations

In Fig. 2, we list the top 10 operations in which people have the most unfluency experiences when using smart phones. Among them, opening the application, using the application, and opening the browser were the three operations in which people had the least fluent experiences.

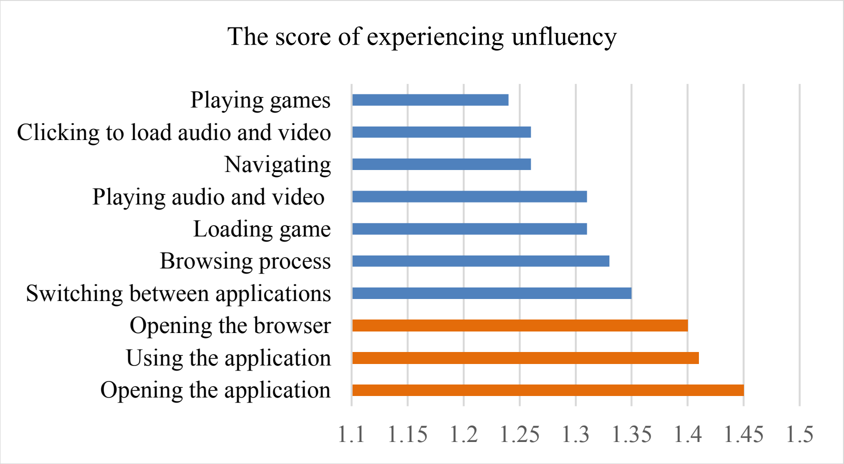


Fig. 2. Top 10 operations in which people have the most unfluency experiences.

4 Discussion

In conclusion, a large-scale questionnaire survey was conducted, which investigated the easy-to-use features of smart phones, proved the six dimensions that affect smart phone fluency, and analyzed the problems and operations that contribute to a lack of fluency on smart phones.

This study successfully confirmed the six factors that influence smart phone fluency, including response delay, simplicity, operation error rate, connection, system failure, and visual experience. The results showed that people regard the six factors as important aspects affecting smart phone fluency, as all the participants rated the influencing degrees as more than mediating. Response delay can bring users a sense of stutters and reduce use experience (Nguyen et al. 2014), which affects the fluency of smart phones. Simplicity makes it easier for users to complete various operations. A higher error rate and lower connectivity can cause more difficulties in task completion and task switching. System failures such as battery failures will seriously influence the normal use of smart phones. Higher visual experience, such as higher screen definition and image quality, helps to improve user satisfaction (Sauer and Sonderegger 2009) and fluency experience. In general, the six dimensions of smart phone fluency cover a series of processes from application startup and application operation to application function switching, which can make a more comprehensive assessment of mobile phone fluency. Moreover, the analysis of GLM repeated measures showed that the importance of the operation error rate on smart phone fluency was significantly higher than that of connection, system failure and visual experience. With the update of mobile phone and hardware facilities, the performance of mobile phone devices is increasingly better, and there are fewer system errors, which make the operation errors more intolerable. Ren et al. (2017) proposed a user experience evaluation model from four dimensions, including performance experience, usability experience, emotional experience and sensory experience, and took fault tolerance as one of the subdimensions of mobile phone usability experience and found that fault tolerance had the greatest impact on mobile phone usability. Compared with the operation error rate, connection and visual experience mainly affect the user's sensory experience and the attractiveness and satisfaction of the smart phone product. Therefore, in the fluency experience, the user's evaluation of the importance of the operation error rate, which affects usability, is higher than that of system failure, which occurs less frequently, and connection and visual experience, which affect sensory experience. It also suggests that operation errors should be reduced as much as possible in product design. We can reduce user operation errors and increase the fluency of mobile phones by designing user-friendly interface layouts (Jain et al. 2013; Nakagawa and Uwano 2012), efficient touch screen interaction methods (Garcia-Ceja et al. 2014), detailed operation guidance (Hagiya et al. 2015; Wang et al. 2014) and so on.

Furthermore, we investigated the problems encountered when using smart phones that most affected users' fluency, and the results showed that slowness and touch screen failure frequently occurred, and these two problems were closely related to response delay and operation error rate. In addition, different age groups had both similarities and differences in terms of their opinions on smart phone fluency. Specifically, participants under 30 years old were significantly more concerned about the frequency of touch error and less concerned about the quality of screens or images than 31–40-year-old people. It is well known that fine movements, motor control and cognitive functions, such as perception, decline with age (Verwey 2010; Voelcker-Rehage 2008). Compared with older people, younger people are more proficient in smart phone operation tasks and have more knowledge and experience in using mobile phones (Al-Showarah and IEEE 2019). Moreover, they also have better cognitive functions and motor control. Therefore, operation errors have a greater impact on the fluency experience of older people.

This study also investigated the problems that affect the fluency experience in smart phones. The results showed that the slowness of software startup, the slow speed after use for a period of time and touch screen failure are the three problems that most affect the fluency experience. The slowness of software startup and the slow speed after use for a period of time are the main response delays, which will increase users' negative emotions and reduce task efficiency and user satisfaction (Szameitat et al. 2009). Touch screen failure is a common problem in the use of mobile phones, which may be related to operation identification and interface design. Regarding touch screen operation identification, if the identification is too sensitive, it will cause many troubles to users, such as dialing the phone by mistake; if the identification is too slow, it will increase the difficulty of user operation. In interface design, an appropriate control size (Jung and Im 2015; Kolly et al. 2012) and icon layout (Jain et al. 2013) should be set to reduce the cognitive load of users and improve the usability of products. Therefore, smart phone manufacturers should improve the touch screen accuracy, shorten the system response time and improve the user experience.

In the last part of the questionnaire, we investigated the fluency experience in different mobile phone operation situations and summed up the 10 operations that people have the most unsmooth experiences. Four of them are not fluent when starting, including opening the application, opening the browser, loading the game and clicking to load audio/video, and one of them is not fluent when switching tasks. Mobile product designers can use some methods to reduce the unsmooth sense when opening and switching, such as showing interactive animation, progress bars and other ways to show that "things are happening", to help reduce users' perceived waiting time when opening and switching applications and improve product satisfaction (Hohenstein et al. 2016; Thomas and Calder 2005). The other five operations are not fluent when using, which may be related to phone memory and hardware facilities. Increasing fluency experience in the use process depends on mobile phone manufacturers continuously improving product performance and managing the background software to improve the response and operation speed of the mobile phone.

5 Limitations and Future Directions

This study proposed six factors of smart phone fluency, but future efforts should be dedicated to making up for the limitations of this study. The present study used a questionnaire method to confirm factors affecting smart phone fluency, but the results were easily influenced by subjectivity, and the respondents were likely to conceal their real opinions. In the future, the findings of the current study could be verified by smart phone operating tasks, which are more objective and accurate. Moreover, it is necessary to develop measurement methods for each factor of smart phone fluency. In addition, the structure of smart phone fluency should be tested by other persuasive techniques, such as confirmatory factor analysis.

6 Conclusion

The present study contributed significantly to confirming the six factors of smart phone fluency based on the survey results in a large sample. We also found features of an easy-to-use smart phone and opening the application, using the application, and opening the browser were the three operations in which people had the least fluent experiences. The important factors of smart phone fluency are response delay, simplicity, operation error rate, connection, system failure, and visual experience. Moreover, there were significant differences in the importance evaluation of visual experience and operation error rate among different age groups. Although these factors may be common in other areas of user experience, they are unique in smart phone user experience. In practice, it provides guidance to smart phone manufacturers for improving user experience. Only by paying more attention to smart phone fluency and user experience can smart phone manufacturers optimize their products and gain more recognition among customers around the world.

Financial Disclosure. This study was supported by grants from the National Key Research and Development Plan of China (2017YFB0802800); the National Natural Science Foundation of China (Grants No. 32071066, 32071064, 31771225, U1736220, 71971073).

References

Al-Showarah, S.A.: Dynamic recognition for user age-group classification using hand-writing based finger on smartphones. In: 2019 10th International Conference on Information and Communication Systems, pp. 140–146. IEEE (2019)

Agost, M.J., Vergara, M.: Relationship between meanings, emotions, product preferences and personal values. Application to ceramic tile floorings. Appl. Ergon. **45**(4), 1076–1086 (2014). https://doi.org/10.1016/j.apergo.2014.01.008

Csikszentmihalyi, M., Mihaly, C.: Flow and the Foundations of Positive Psychology. Springer, New York (2014). https://doi.org/10.1007/978-94-017-9088-8

Dufau, S., et al.: Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. Plos One **6**(9), e24974 (2011). https://doi.org/10.1371/journal.pone.0024974

Galletta, D.F., Henry, R., McCoy, S., Polak, P.: Web site delays: how tolerant are users? J. Assoc. Inf. Syst. **5**(1), 1 (2004)

Garcia-Ceja, E., Brena, R., Galván-Tejada, C.E.: Contextualized hand gesture recognition with smartphones. In: Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Olvera-Lopez, J.A., Salas-Rodríguez, J., Suen, C.Y. (eds.) MCPR 2014. LNCS, vol. 8495, pp. 122–131. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07491-7_13

Hagiya, T., Yazaki, T., Horiuchi, T., Kato, T.: Typing tutor: automatic error detection and instruction in text entry for elderly people. In: MobileHCI 2015 - Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, pp. 696–703 (2015). https://doi.org/10.1145/2786567.2793690

Hassenzahl, M.: User experience (UX): towards an experiential perspective on product quality. Paper presented at the IHM (2008). https://doi.org/10.1145/1512714.1512717

Hassenzahl, M.: The thing and i: understanding the relationship between user and product. In: Blythe, M., Monk, A. (eds.) Funology 2. HIS, pp. 301–313. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-68213-6_19

Hohenstein, J., Khan, H., Canfield, K., Tung, S., Perez Cano, R.: Shorter wait times: the effects of various loading screens on perceived performance. In: Conference on Human Factors in Computing Systems – Proceedings, vol. 7, no. 12, pp. 3084–3090 (2016)

Jain, R., Bose, J., Arif, T.: Contextual adaptive user interface for android devices. In: 2013 Annual IEEE India Conference. IEEE (2013)

Jung, E.S., Im, Y.: Touchable area: an empirical study on design approach considering perception size and touch input behavior. Int. J. Ind. Ergon. **49**, 21–30 (2015). https://doi.org/10.1016/j.ergon.2015.05.008

Kolly, S., Wattenhofer, R., Welten, S.: A personal touch: Recognizing users based on touch screen behavior. In: Proceedings of the 3rd International Workshop on Sensing Applications on Mobile Phones, PhoneSense 2012 - In Conjunction with ACM SenSys (2012). https://doi.org/10.1145/2389148.2389149

Law, E.L.C., van Schaik, P., Roto, V.: Attitudes towards user experience (UX) measurement. Int. J. Hum. Comput. Stud. **72**(6), 526–541 (2014). https://doi.org/10.1016/j.ijhcs.2013.09.006

Li, J.W.: Evaluation system research and implementation of the smart phone user experience. Beijing University of Posts and Telecommunications, Beijing (2015). (Chinese)

Liu, X.Z., Ge, Y., Wang, C., Zhang, Q., Qu. W.N.: Dimension of smart phone fluency: a focus group interview (Submitted)

Nakagawa, T., Uwano, H.: Usability differential in positions of software keyboard on smartphone. In: 1st IEEE Global Conference on Consumer Electronics 2012, GCCE 2012, pp. 304–308 (2012)

Nguyen, D.T., Zhou, G., Xing, G.: Video: study of storage impact on smartphone application delay. In: MobiSys 2014 - Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, p. 389. ACM (2014)

Norman, D., Miller, J., Henderson, A.: What you see, some of what's in the future, and how we go about doing it: HI at Apple Computer. Paper presented at the Conference Companion on Human Factors in Computing Systems (1995)

Oppenheimer, D.M.: The secret life of fluency. Trends Cogn. Sci. **12**(6), 237–241 (2008). https://doi.org/10.1016/j.tics.2008.02.014

Park, B.-W., Lee, K.: A pilot study to analyze the effects of user experience and device characteristics on the customer satisfaction of smartphone users. In: Kim, T.-h, Adeli, H., Robles, R.J., Balitanas, M. (eds.) UCMA 2011. CCIS, vol. 151, pp. 421–427. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20998-7_50

Ramsay, J., Barbesi, A., Preece, J.: A psychological investigation of long retrieval times on the World Wide Web. Interact. Comput. **10**(1), 77–86 (1998). https://doi.org/10.1016/S0953-5438(97)00019-2

Ren, H.B., Zhou, J., Feng, W.Q.: Evaluation of the smart phone user experience based on MTS. Packag. Eng. (2017)

Sauer, J., Sonderegger, A.: The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. Appl. Ergon. **40**(4), 670–677 (2009). https://doi.org/10.1016/j.apergo.2008.06.006

Schleifer, L.M., Iii, B.C.A.: System response time and method of pay: stress effects in computer-based tasks. Int. J. Hum.-Comput. Interact. **1**(1), 17 (1989)

Seva, R.R., Duh, H.B.L., Helander, M.G.: The marketing implications of affective product design. Appl. Ergon. **38**(6), 723–731 (2007). https://doi.org/10.1016/j.apergo.2006.12.001

Seva, R.R., Gosiaco, K.G.T., Santos, M.C.E.D., Pangilinan, D.M.L.: Product design enhancement using apparent usability and affective quality. Appl. Ergon. **42**(3), 511–517 (2011). https://doi.org/10.1016/j.apergo.2010.09.009

Shi., F.P.: Research on the model of consumer satisfaction. Jinan University (2005). (Chinese)

Si, Y.Q., Liu, L.: Study on the Influencing factors of the intelligent mobile phone user experience. Software **36**(3), 111–115 (2015). (Chinese)

Szameitat, A.J., Rummel, J., Szameltat, D.P., Sterr, A.: Behavioral and emotional consequences of brief delays in human-computer interaction. Int. J. Hum. Comput. Stud. **67**(7), 561–570 (2009). https://doi.org/10.1016/j.ijhcs.2009.02.004

Thomas, B.H., Calder, P.: Supporting cartoon animation techniques in direct manipulation graphical user interfaces. Inf. Softw. Technol. **47**(5), 339–355 (2005). https://doi.org/10.1016/j.infsof.2004.09.003

van Kuijk, J., van Driel, L., van Eijk, D.: Usability in product development practice; an exploratory case study comparing four markets. Appl. Ergon. **47**, 308–323 (2015). https://doi.org/10.1016/j.apergo.2014.10.007

Verwey, W.B.: Diminished motor skill development in elderly: indications for limited motor chunk use. Acta Physiol. **134**(2), 206–214 (2010). https://doi.org/10.1016/j.actpsy.2010.02.001

Voelcker-Rehage, C.: Motor-skill learning in older adults—a review of studies on age-related differences. Eur. Rev. Aging Phys. Act. **5**(1), 5–16 (2008). https://doi.org/10.1007/s11556-008-0030-9

Walsh, T., Nurkka, P., Walsh, R.: Cultural differences in smartphone user experience evaluation. Paper presented at the International Conference on Mobile & Ubiquitous Multimedia (2010)

Wang, C.Y., Chu, W.-C., Chen, H.-R., Hsu, C.-Y., Chen, M.: EverTutor: automatically creating interactive guided tutorials on smartphones by user demonstration. In: Conference on Human Factors in Computing Systems – Proceedings (2014). https://doi.org/10.1145/2556288.2557407

Winkielman, P., Cacioppo, J.T.: Mind at ease puts a smile on the face: psychophysiological evidence that processing facilitation elicits positive affect. J. Pers. Soc. Psychol. **81**(6), 989–1000 (2001). https://doi.org/10.1037//0022-3514.81.6.989

Yan, Z., Niu, S., Wang, S.: Research on user experience evaluation methods of smartphone based on fuzzy theory. Paper presented at the International Conference on Intelligent Human-machine Systems & Cybernetics (2015)

Pseudo Human Sense in the Loop: Proposal of a Presentation Support Method by Pseudo Feedback of Audience Sense

Kureha Noguchi, Yoshinari Takegawa(✉), and Keiji Hirata

Future University Hakodate, Hakodate, Japan
{g2120038,yoshi,hirata}@fun.ac.jp

Abstract. In recent years, attention has been focused on information presentation methods that take into account the user's situation by utilizing wearable computing technology. Most of the existing information presentation methods present users with information that gives them a choice of actions, and encourage them to take actions. However, users may not be able to control their own actions appropriately. It is important not only to encourage users to act, but also to forcibly control their behavior. Such behavioral control methods use actuators (e.g. displays, speakers etc.) to stimulate senses such as vision and hearing, and then control behavior by reversing the reaction to maintain consistency with the usual senses. However, most of these studies have not yet examined how to determine the intensity of the stimuli when introducing behavioral control methods into the real world. To solve this problem, we focused on the influence of the presence of others on human behavior control. We propose a method to determine the amount of stimuli to be fed back to the user based on the results of simulating the sensations of others. We define this method as Pseudo Human Sense in the Loop (referred to as "PHSIL" in this paper), and conduct cognitive psychology experiments on presentations, applying PHSIL to both auditory and visual stimuli, to verify the effectiveness of PHSIL.

Keywords: Cognitive psychology · Voice control · Mixed reality · Hearing · Sight

1 Introduction

In recent years, attention has been focused on information presentation methods that take into account the user's situation by utilizing wearable computing technology. For example, wearable devices such as Google Glass and Apple Watch are used to support user behavior by providing useful information such as navigation to a destination, sale information when shopping, and health management based

This work was supported by JSPS KAKENHI Grant Number JP19H04157.

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 389–407, 2021.
https://doi.org/10.1007/978-3-030-77932-0_31

on step count and heartbeat. In this way, many existing information presentation methods encourage users to take actions by presenting them with information that gives them options. However, there are cases where users cannot immediately control their own behavior, such as in a state of tension, intoxication, or confusion. Therefore, it is important to use wearable systems not only to encourage users to act, but also to control their actions.

Several behavioral control methods have been proposed in the past. All of these methods use actuators that stimulate sensory organs (e.g. displays, speakers etc.) to stimulate senses such as vision and hearing, and control behavior by taking advantage of the response that tries to maintain consistency with the usual senses. However, the method of determining the intensity of the stimulus has not yet been studied when introducing these behavior control methods in the real world.

For example, the voice volume control method developed by the author's research group succeeded in amplifying/attenuating voice volume by feeding back white noise and the speaker's own voice to the speaker's auditory system [1]. The stronger the feedback to the auditory system, the stronger the amplification/attenuation of the voice volume. One of the applications of the proposed method is the control of voice volume in presentations. However, the proposed method was only applied to a laboratory experiment in a quiet soundproof room, which is quite different from the actual presentation environment. In an actual presentation, for example, the speaker's voice may be low at the beginning of the presentation due to nervousness, but it may become louder in the latter half of the presentation as the speaker gets used to talking. Also, if there is a noisy audience, the presenter may not know how loudly to speak. The size of the presentation venue also has a great impact on the volume of speech. Therefore, when supporting presentation-making, it is difficult to control behavior with a fixed amount of stimuli.

As a method to solve this problem, this study focused on the fact that the presence of others has an effect on the control of human behavior [21]. We propose a method to determine the intensity of stimuli to be fed back to the user based on the results of simulating the sensation of others, using a microphone and a video camera. We call this method Pseudo Human Sense in the Loop (referred to as PHSIL in this paper), and conduct a cognitive psychology experiment in which we apply PHSIL both auditorily and visually to a presentation, to verify its effectiveness.

2 Related Research

2.1 Presentation Support

There have been many studies on presentation. Among them, Maeda [2] and his colleagues proposed a presentation style in which the presenter stands in front of a screen, so that the presenter and the screen are viewed as one. In this style, consolidating the information presented by the presenter focuses the audience's attention and non-verbal information, which is difficult to convey in conventional presentations, can be conveyed to the audience. In addition, by dynamically arranging objects to avoid the presenter, Maeda et al. solved the problem of shielding, which is a detriment to the proposed presentation style. By mapping the presenter's gestures to the animation of the objects, interaction with the screen is possible, and the presenter's expression can be made more effective. The aspect of this research that supports presentation is related to our research, but the aspect that supports the presenter using the screen is different from our research. Kurihara et al. [3] proposed a presentation training system that combines audio and image information processing. The developed "Presentation Sensei" analyzes the presenter's voice and behavior obtained from a microphone and a Web camera, and provides real-time feedback to the presenter on indicators such as speaking speed, voice inflection, and the degree of eye contact with the audience, in conjunction with presentation tools. Kurihara et al.'s system does not improve the content of the presentation itself, but suppresses unintentional inappropriate behaviors during the presentation, thereby improving the presentation skills of the presenter alone. Our presentation experiment differs from Kurihara et al. in that it aims to control the volume of the presenter's voice and the magnification rate of the slides to an appropriate size in real time.

2.2 Behavior Control Methods

There are many research examples of controlling human behavior using wearable devices. For example, emotion recognition control in children with ASD by providing emotion classification [4], appetite control by visual presentation using virtual reality (VR) [5], behavior control using vibration motors [6,7], and performance tempo control using visualization of heart rate information [8]. In this paper, we introduce a new approach to human behavior control.

2.3 Auditory Feedback

There are a number of studies related to auditory perception in our research, such as the Speech Chain, a vocalization model that has long been used in the fields of articulatory phonetics, acoustic phonetics, and auditory phonetics. The principle of the Speech Chain is that the speaker visualizes a speech message in his or her mind, and then the signaled speech message moves the language development muscles through the central nervous system to produce speech. The articulated sound is then emitted from the speaker's mouth and reaches the ears of the listener. The sound that reaches the listener vibrates the eardrum and reaches the cerebrum from the inner ear through the cochlear nerve (auditory nerve), and the content of the speaker's speech is transmitted to the listener. Next, the speaker and the listener switch roles, and speech is generated and transmitted, and the conversation proceeds in a circular fashion. In recent years, research applying the Speech Chain has had a significant impact on phonetics [9].

The Lombard effect [10] is a vocalization model that specializes in voice feedback. For example, we naturally tend to speak louder in a noisy environment, but it is known that when we are exposed to noise or white noise, not only do our voices become louder, but various acoustic characteristics change, for example, the fundamental frequency and formant frequency become higher. Hodoshima et al. [11] investigated the intelligibility of vocalizations in quiet, noisy, and reverberant environments. In addition, there is research on the realization of the Lombard effect on a computer system [12]. Our study is different in that we investigate the interaction method when the Lombard effect is applied to the field of presentation.

Kawahara et al. [13] proposed transformed auditory feedback (TAF), in which small perturbations to acoustic parameters are applied via auditory feedback to examine the effect on the subsequent speech production, and conducted experiments using the fundamental frequency of speech as the perturbation target. The results of these experiments showed that when the fundamental frequency of speech fluctuates, the speaker's response acts to cancel the fluctuation with a delay of about 150 ms. In our study, we will propose a voice volume control method or design a proposed system based on the results obtained by TAF using fundamental frequency.

Auditory feedback refers to a person listening to his or her own voice during speech and adjusting the pitch and intensity of speech [14]. If we artificially delay the timing at which we hear the sound of our own speech, it will interfere with subsequent speech. Delayed auditory feedback (DAF) has been shown to reduce the frequency of stuttering when applied to people who stutter [15,16]. As another example of a study using DAF, Kurihara et al. developed a system called Speech Jammer that inhibits a speaker's speech without causing physical pain, by artificially delaying the feedback of the speech [3]. Our research is closely related to Speech Jammer in terms of speech control. However, users to whom Speech Jammer is applied stop speaking when their speech is directly inhibited, whereas the approach of our study is different in that it focuses on vocal volume and controls the volume without interfering with speech.

Terada has developed a system that can keep the volume of speech that the user hears constant. This system can accurately convey audio information to the user by changing the presentation method according to the user's situation, while also maintaining a constant volume that the user hears [17]. In this way the system reduces the discomfort of the speech that the user hears, even when the system is worn in different situations. This related research will be used as a reference to investigate ways to reduce the discomfort of the sounds the user receives.

2.4 Visual Feedback

There are many studies related to vision in our research. There is a device called an Inter Dis-Communication Machine that allows two parties to exchange what they are seeing [18]. This device forces the user to place him/herself in the other person's position by seeing only the other person's point of view. The goal is to blur the boundaries of identity. This research is similar to our research in that it allows the user to see a third party's view, but it differs in that it uses part of the user's own vision.

Next, there is research on linking unmanned aerial vehicle (UAV) control operations to human body movements [19]. When compared with the conventional operation method, the results show that UAV operation becomes easier. In our research, it is important to link the range captured by the camera to the user's field of vision and body movements.

In recent years, there has been a lot of research on AR and VR, and we can imagine that in future society Head Mounted Displays (HMD), such as HoloLens, will be used in our daily lives. For certain scenarios, such as monitoring the situation during ship docking, visualizing objects outside the field of view by using HMD is a useful technique [20]. It is relevant to this study in that it assumes the use of HMD in daily life.

3 Behavior Control Method Based on PHSIL

The PHSIL-based behavioral control method is designed to be applied to presentations. It is rare for people not to feel nervous during job-hunting activities or presentations at academic conferences, and this nervousness increases as the importance of the presentation increases. In this paper, we apply the PHSIL-based behavioral control method to presenters to solve these problems.

3.1 Voice Control Method

By making the presenter listen to white noise during the speech, the volume of the presenter's speech is amplified. In a presentation, the optimal voice volume differs depending on not only factors caused by the presenter (nervousness, anxiety, lack of practice, etc.) but also on the environment of the presentation site (size of the presentation site, buzz of the audience, etc.), so the volume of the white noise fed back to the presenter needs to be dynamically controlled.

The voice volume control model is shown in Fig. 1. In the usual voice volume control model that we use in our daily life, (i) the listener hears the speech of the presenter, (ii) The speech delivered by the presenter reaches the eardrum of the presenter through the presenter's own auricle and skull. (iii) If the listener feels that the presenter's speech is difficult to hear, (iv) the listener can tell the presenter that it is difficult to hear, with a comment such as "please speak more loudly". (v)(vii) When the presenter receives this message, (viii) a change in behavior appears, such as raising of the voice.

In this voice volume control model with PHSIL, (i) a listener listens to the presenter's speech. The speech of the presenter as it sounds to the listener is called the audible speech. (viii) The audible speech is fed back to the presenter to control (vi) and (vii) the presenter's voice volume.

Specifically, the presenter is made to listen to white noise so that the volume of his own speech heard by the presenter is approximately the same as the

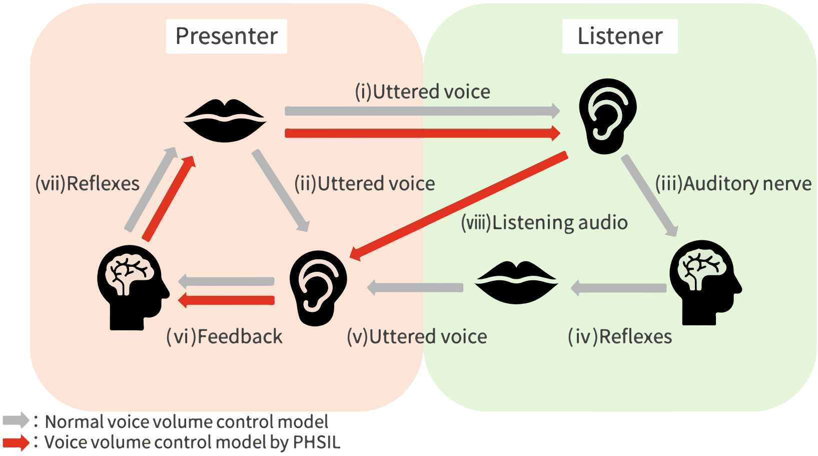


Fig. 1. Voice volume control model

volume of the presenter's speech heard by the listener (e.g., a listener sitting at the back of the presentation hall). In other words, the white noise attenuates the speech sound in (ii). This gives the presenter a simulated experience of the sound heard by the audience. When the audience has difficulty hearing the presenter's speech, the presenter himself also has difficulty hearing his own speech, so he involuntarily increases the volume of his speech. When the volume of the presenter's voice becomes easy for the audience to hear, the presenter's voice also becomes easy for he himself to hear, so the presenter does not increase the volume of his voice unnecessarily.

3.2 Slide Magnification Control Method

When the presenter looks at the screen, if the slide visible is replaced with one smaller than usual, he can be made to increase the magnification rate of the slide. In a presentation, the font size varies depending on the amount of information contained in the slides, and figures and tables are inserted. In addition, the optimum slide magnification rate varies depending on the size of the presentation room.

The slide magnification control model is shown in Fig. 2. In our usual slide magnification control model, (i) the presenter presents the slide to the audience. (ii) When the listener feels that the slide is difficult to see, (iii) the listener reacts by squinting to indicate that the slide is difficult to see. (iv)(v) When the presenter observes this reaction, a behavioral change appears, such as (vi) enlarging the slide.

In this model that incorporates PHSIL into the slide magnification control model, (i) the presenter presents the slides to the audience. (vii) A part of the visual image seen by the audience (screen area only) is (v) fed back to the presenter to (vi) control the slide magnification rate.

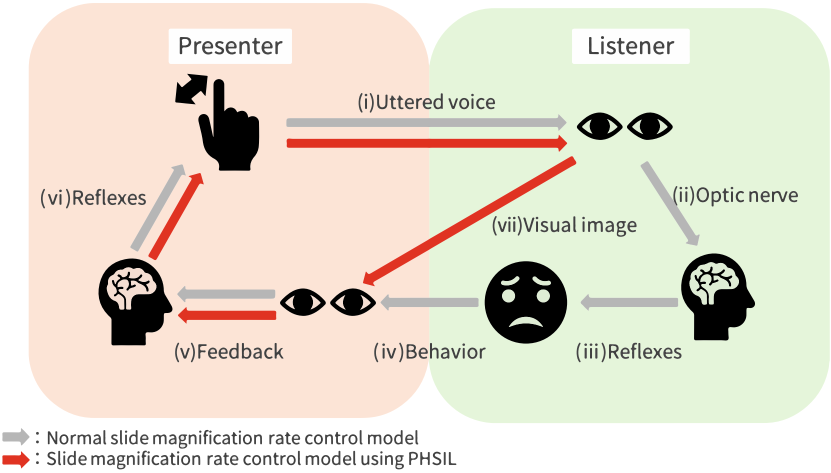


Fig. 2. Slide magnification rate control model

Specifically, the visual image of the screen portion presented to the presenter is controlled so that the visual image seen by the presenter is the same as the image of the presenter seen by the audience (e.g., audience members sitting at the back of the presentation hall). The gives the presenter a simulated experience of how the audience sees the slides projected on the screen. If it is difficult for the audience to see the presenter's slides, it is also difficult for the presenter to see his own slides, so the presenter involuntarily increases the magnification of the slides. When the slides are easy for the audience to see, the slides are also easy for the presenter to see, so the presenter will not increase the size of the screen unnecessarily.

4 Verification of PHSIL Focusing on Voice Volume Control

In order to verify the validity of the PHSIL noise control, we conducted an evaluation experiment. We set up an experimental scenario of a presentation using slides (hereinafter referred to as "presentation"), and the presenter and the listener were the subjects of the experiment. Two comparison methods were used to evaluate the usefulness of the proposed method. The evaluation items were the volume of the presenter's voice during the presentation, the persistence of the volume change, and the subjective audibility of the presentation for the audience.

Subject. There were 20 subjects (10 presenters and 10 listeners), who were healthy university students with no speech, hearing, or visual disabilities. All subjects did not understand the details of this study or the purpose of the experiment. When we recruited the subjects, we told them that they would be asked to perform either a task in which they had to make a presentation with slides in front of a group of people or a task in which they had to listen to a presenter. The presenters were not used to giving presentations, as they gave presentations only once or twice a year.

Comparison Method. As a comparison method, we used the noise-free method, in which the presenter is asked to give a presentation in the same situation as a general presentation without hearing white noise, and the fixed noise method, in which the amplification of white noise is constant. In this paper, the PHSIL noise method and the fixed noise method are collectively referred to as the noise methods.

Assignment Slides. The assignment slides to be presented by the presenter consist of 11 pages, and the presentation time is about 6 min. The same subject slides were used for both the PHSIL noise method and the comparison methods. A web article titled "A Simple Manual for WiFi Users - Towards Safe WiFi Use -"[1], was edited and used as slides.

Instructions for Subjects. The presenters of both the PHSIL noise method and the comparison methods were instructed to convey the contents shown on the slides to the audience in an easy-to-understand manner, and were told that they could freely decide the timing of moving to the next page and the words to be added there. We did not instruct the speaker as to how loud speaking volume should be. The audience was instructed to listen to the presenter's presentation and to answer the questionnaire afterwards.

Experiment Place. The experiment took place in a lecture room at Future University Hakodate. The lecture room is not equipped with ceiling speakers or wall speakers. In a normal lecture, the teacher gives a lecture using slides and a white board. A projector is installed in the lecture room, and the slides delivered from the presentation PC are projected on the screen. The experimental environment is shown in Fig. 3.

Experimental Procedure. The presenters were asked to perform the experiment according to the following procedure. Task items 2 through 4 were administered randomly to each presenter in order to prevent ordering effects. In addition, a 20-minute break was set after each presentation. This was to refresh the effect of the white noise presented during the presentation and to recover physical strength. During the break, the participants were asked to answer the questionnaire shown in Table 1.

1. Subjects understand the contents of the task slides. They also practice the presentation using a presentation PC. This series of tasks is performed until the presenters are satisfied, or for a maximum of 30 min. It is confirmed that slide operation using a PC can be performed without any problem.

[1] http://www.soumu.go.jp/main_sosiki/joho_tsusin/security/cmn/wi-fi/Wi-Fi_manual_for_Users.pdf.

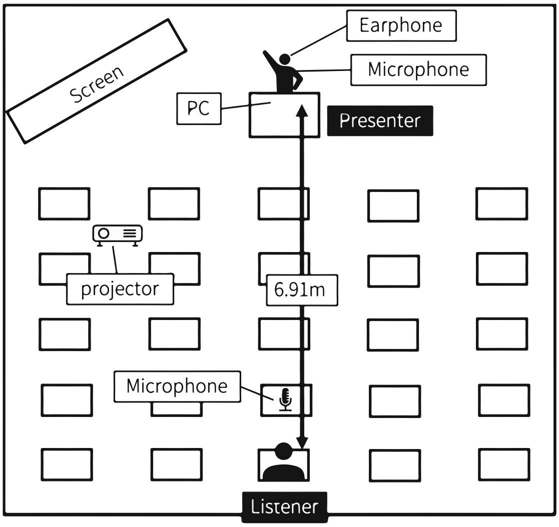


Fig. 3. Experiment environment

Table 1. Questionnaire contents

	Question items
Presenter	How uncomfortable did the white noise feel? (1:Very uncomfortable-5:Not at all uncomfortable)
Listener	How easy was it to hear the presenter's voice? (1:Very difficult-5:Very easy)

2. Subjects give a presentation using the PHSIL noise method.
2. Subjects give a presentation using the fixed noise method.
4. Subjects give a presentation using the noise-free method.

Experimental System. Each presenter wore a pin microphone (for recording the volume of speech) on the chest and earphones (bone conduction earphones for listening to white noise) in the ears. The experimental system for the PHSIL noise method is shown in Fig. 4. The details of the experimental system for each method are described below.

PHSIL noise method: A microphone (hereinafter referred to as a rear microphone) was installed next to a listener seated at the rear of a lecture room. In a preliminary experiment, the optimum voice volume at the rear microphone position was investigated, and the audio signal recorded by the rear microphone at that time was recorded with a sampling frequency of 44.1 kHz and a quantization bit rate of 16 bits. After converting the signal to 4 kHz, the absolute value of the amplitude was obtained and the moving average of 100 samples was calculated. The amplitude varied from 1 to -1, and the absolute value was applied to the

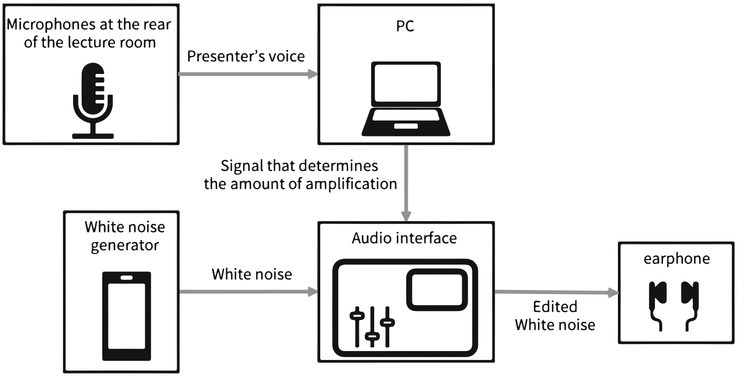

Microphones at the rear of the lecture room
Presenter's voice
PC
Signal that determines the amount of amplification
White noise generator
White noise
Audio interface
Edited White noise
earphone

Fig. 4. Experimental system for PHSIL noise method

amplitude and converted to decibels, which is called X_{prop}. X_{min} is the result of applying the same process to a speech signal that is judged to be difficult to hear. The amplification level of the white noise presented during speech during a presentation ranges from 0 dB to 24 dB, and is controlled in 128 steps. When the presenter is not speaking, the amplification level of the white noise is set to 0 dB. The amplification level (y) of the white noise presented to the presenter is set based on the following equation for the input (x), to which the same processing as above is applied for the audio signal from the rear microphone. The larger y is, the larger the amplification of the white noise presented to the presenter. $y = 0$ when y is less than 0, and $y = 127$ when y is greater than 127.

$y = 127 - (x - X_{min})/(X_{prop} - X_{min}) * 127$ An Apple MacBook was used as a PC for calculating the steps of white noise amplification, and a TOSHIBA dynabook was used as a PC for recording the presenter's voice. An audio technica AT9904 was used as a pin microphone and an audio technica AT-VD3 was used as a rear microphone. In addition, AS400 earphones from AFTERSHOKZ were used. Also, a Roland VR-4HD was used to generate the specified volume for the white noise source. For the projector, we used an EPSON EB-940. Fixed noise method: In the fixed noise method, the presenter is always presented with 24 dB of white noise. The rest of the experimental system is the same as that of the PHSIL noise method. In other words, the fixed noise method corresponds to the experimental condition in which, taking the presenter as a basis, the rear microphone is set at an infinitely distant point. Noise-free method: In the noise-free method, the white noise presented in the PHSIL noise method and the fixed noise method is not presented. Therefore, although the presenter wears earphones, no white noise is output from the earphones. In other words, the noise-free method corresponds to the experimental condition in which a rear microphone is placed at the same position as the pin microphone worn by the presenter.

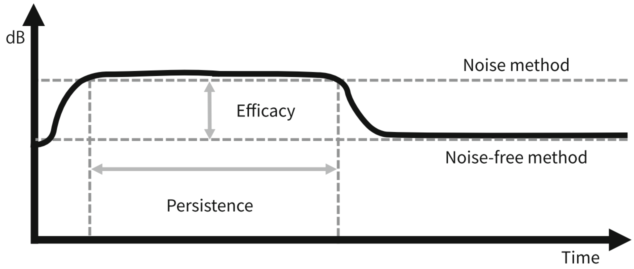


Fig. 5. Evaluation item

Evaluation Items. The evaluation items are efficacy and persistence, as shown in Fig. 5, and the questionnaire items described above. The method of calculating the efficacy and persistence items is described below. The sampling frequency of the audio signal obtained from the microphone is 44.1 kHz and the quantization bit rate is 16 bits. After converting the audio signal to 4 kHz, the absolute amplitude was obtained from the presenter's audio signal, and the moving average of 100 samples was calculated. The amplitude varied from 1 to -1. The absolute value was applied to this amplitude and converted to decibels, and the resulting value was defined as the average speech volume.
Efficacy: The difference between the average speech volume (dB) measured in the noise-free method and the average speech volume (dB) measured in the noise methods.
Persistence: The time (in seconds) that the speaker's vocal volume in the noise methods exceeded the average vocal volume in the noise-free method.

4.1 Result

The results for each evaluation item are shown in Fig. 6. The average amplification of the white noise presented to the presenter in the PHSIL noise method was about 12.8 dB.

Efficacy. The efficacy of the PHSIL noise method averaged 2.6 dB, while the efficacy of the fixed noise method averaged 2.8 dB. When Welch's t-test was applied to the efficacy of the PHSIL and fixed noise methods at a 5% significance level, no significant difference was observed.

Persistence. The persistence of the fixed noise method averaged 99 s and that of the PHSIL noise method averaged 248 s. The persistence of the fixed noise method and the PHSIL noise method are very different. As a result of applying Welch's t-test to the persistence of the fixed and PHSIL noise methods at a significance level of 5%, a significant difference was observed ($t(16) = 4.01$, $p < .05$).

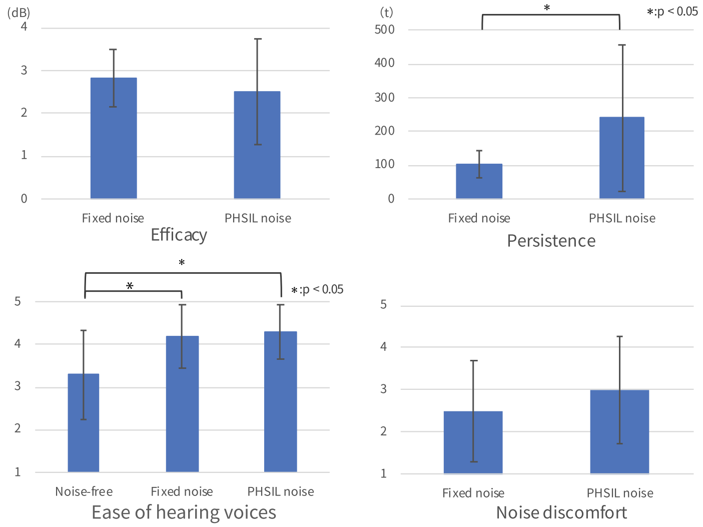


Fig. 6. Voice volume control result

Questionnaire Results. In terms of noise discomfort, the fixed noise method had a mean of 1.8 and the PHSIL noise method had a mean of 3.0. As a result of applying Wilcoxon's signed rank sum test at 5% significance level, a significant difference was observed between the fixed noise method and the PHSIL noise method ($Z = 2.32$, $p < .05$). This means that the fixed noise method is more unpleasant than the PHSIL noise method.

In terms of the audibility of the presenter's voice, the noise-free, PHSIL noise, and fixed noise methods averaged 3.3, 4.3, and 4.2, respectively. As a result of applying Wilcoxon's signed rank sum test at % level of significance, significant differences were observed between the noise-free method and the PHSIL noise method ($Z = 3.37$, $p < .05$) and between the noise-free method and the fixed noise method ($Z = 3.21$, $p < .05$). The higher the value, the more intelligible the voice is, thus speech is more intelligible in the noise methods than the noise-free method.

4.2 Discussion

Noise Methods vs Noise-Free Method. It was observed that the PHSIL noise method and the fixed noise method increased the volume of speech more than the noise-free method. In addition, the questionnaire results of the listeners showed that the volume of the presenter's voice was easier to hear in the methods with noise than in the method without noise. By presenting white noise to the presenter, the volume of the presenter's voice was amplified, and the amplified volume of the presenter's voice changed clearly until the listeners perceived it as easy to hear.

Fixed Noise Method vs PHSIL Noise Method. No difference in speech volume was observed between the PHSIL noise method and the fixed noise method. One of the major differences observed between the PHSIL and fixed noise methods was the persistence. In the fixed noise method, 24 dB of white noise was always presented, which caused the presenter to speak more loudly than necessary, even if the vocal volume was loud enough to be heard at the back of the room. As a result, the presenter became exhausted and was not able to speak loudly at the end of the presentation. In addition, since there was no change in the loudness of the white noise, some presenters found that their ears became accustomed to the white noise environment and their voice volume decreased.

As for the discomfort of white noise, the fixed noise method was found to be more uncomfortable than the PHSIL noise method. When the presenters were asked why they felt the fixed noise method was uncomfortable, all the subjects commented that the amount of amplification of the white noise presented by the fixed noise method was basically large and that they felt the white noise was noisy. On the other hand, the PHSIL noise method can flexibly respond to the loudness of the natural voice, which varies from person to person, and the amplification of the white noise was small for the three presenters with relatively loud natural voices. The average amplification of the white noise was 12.4 dB, so the PHSIL noise method did not present the presenters with the same 24 dB average white noise amplification as the fixed noise method.

5 Verification of PHSIL Focusing on Slide Magnification Control

We conducted an evaluation experiment to verify whether the proposed method with an HMD is useful or not. We set up an experimental scenario of a presentation with slides, with a presenter and an audience as subjects. We also prepared a comparison method and evaluated the usefulness of the proposed method by using the maximum magnification rate of the slides during the presentation and the viewability of the presentation slides as evaluation items.

Subject. There were 20 subjects (10 presenters and 10 listeners), who were healthy university students with no disabilities in speech, hearing, or vision. All subjects did not understand the details of this study or the purpose of the experiment. When we recruited the subjects, we told them that they would be asked to perform either a task in which they had to make a presentation with slides in front of a group of people or a task in which they had to listen to a presenter. The presenters were not used to giving presentations, as they gave presentations only once or twice a year.

Comparison Method. As a comparison method, we used the method without an HMD, in which the participants were asked to give a presentation in the same situation as a general presentation, without wearing an HMD.

Assignment Slides. We used different slides for the method with HMD and without HMD. The slides consisted mainly of text, maps, and graphs. This is because there are many presentations that use text and graphs in general, and, when using a map in a presentation, it is thought that zooming in and out will occur frequently, such as when presenting the entire map then zooming in on a destination. Compared to text and graphs, the presentation method of maps is expected to differ greatly depending on the presenter. The explanation of the airplane[2] and the explanation of IT[3] were used as text slides. Access from Kofu Station to the Science Museum[4] and access from Nagoya Station to Nagoya Castle[5] were used as map slides. The number of typhoons approaching[6] and the estimated production of sweets in Japan[7] were used as graph slides.

Instructions for Subjects. Both presenters with and without HMD were instructed to convey the contents shown on the slides to the audience in an easy-to-understand manner, and were told that they could freely decide the timing of moving to the next page and the words to be added at that time. The presenters were not instructed on the magnification of the slides. The audience was instructed to listen to the presenter's presentation and to answer the questionnaire afterwards.

Experiment Place. We used the same lecture room as in Sect. 4 as the experimental site. The experimental environment is shown in Fig. 7.

The presenter wore an HMD (to insert the slides from the audience's perspective at the rear of the lecture room). The size of the screen was 1.54 × 2.04 (m). The distance between the screen and the projector was 2.76 (m), and the distance between the screen and the listener was 7.72 (m).

Experimental Procedure. The presenters were asked to perform the experiment according to the following procedure. Task items 2 to 3 were administered randomly to each presenter in order to prevent ordering effects. In addition, a 10-minute break was set after each presentation. This was to refresh the effect of the magnification rate of the slides presented during the presentation and to recover physical strength. After the experiment, the participants were asked to respond to an interview about whether the method with HMD helped their presentation. After the experiment, the listeners were asked to answer the questionnaire shown in Table 2.

[2] https://ja.wikipedia.org/wiki/%E9%A3%9B%E8%A1%8C%E6%A9%9F.

[3] https://www.internetacademy.jp/it/programming/programming-basic/it_usage_something_field.html.

[4] https://www.kagakukan.pref.yamanashi.jp/.

[5] https://www.nagoyajo.city.nagoya.jp/.

[6] https://www.data.jma.go.jp/fcd/yoho/typhoon/statistics/accession/index.html.

[7] http://www.chocolate-cocoa.com/statistics/domestic/confectionery_j.html.

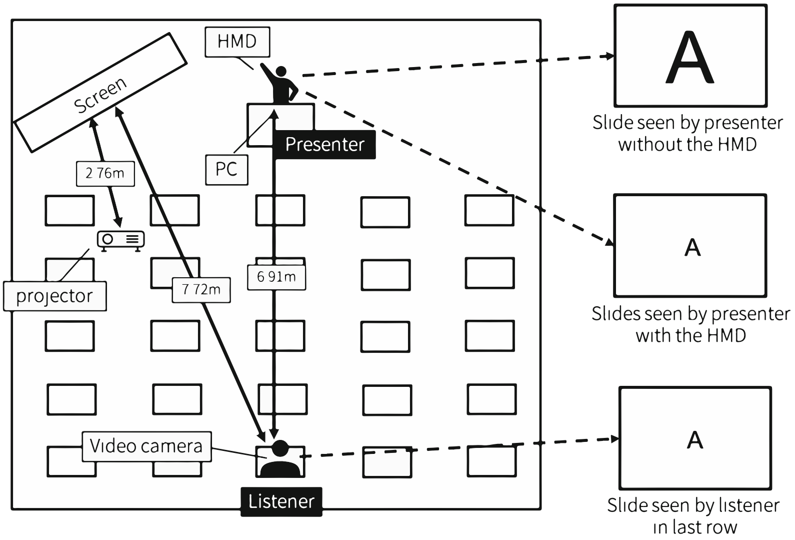

HMD
Screen
2 76m
PC
Presenter
projector
7 72m
6 91m
Video camera
Listener
A
Slide seen by presenter without the HMD
A
Slides seen by presenter with the HMD
A
Slide seen by listener in last row

Fig. 7. Experiment environment

Table 2. Questionnaire contents

	Question items
Listener	How easy to see was the slide being presented? (1: Very hard to see-5: Very easy to see)

1. Subject studies the contents of the task slides then practices the presentation using a presentation PC. This series of tasks is performed until the presenter is satisfied or for up to 30 min. It is confirmed that slide operation using a PC can be performed without any problem.
2. Subject gives a presentation using the method with HMD.
3. Subject gives a presentation using the method without HMD.

Experimental System. Method with HMD: To conduct the evaluation experiment, we constructed the experimental system shown in Fig. 8. The experimental system uses video calling between the PC and the HMD to share screens. We used the Microsoft Teams application on the PC and the Remote Assist application on the HMD to change the magnification of the slides by pinching in and out on the touchpad of the PC. The presenter wore an HMD, into which was inserted a part of the visual image seen by a person looking at the screen from the back seat of the lecture room.

The size of the slide inserted in the HMD was adjusted to be the same as the size of the slide on the screen, with the experimenter seated in the back seat of the lecture room before the start of the method with HMD. The size of

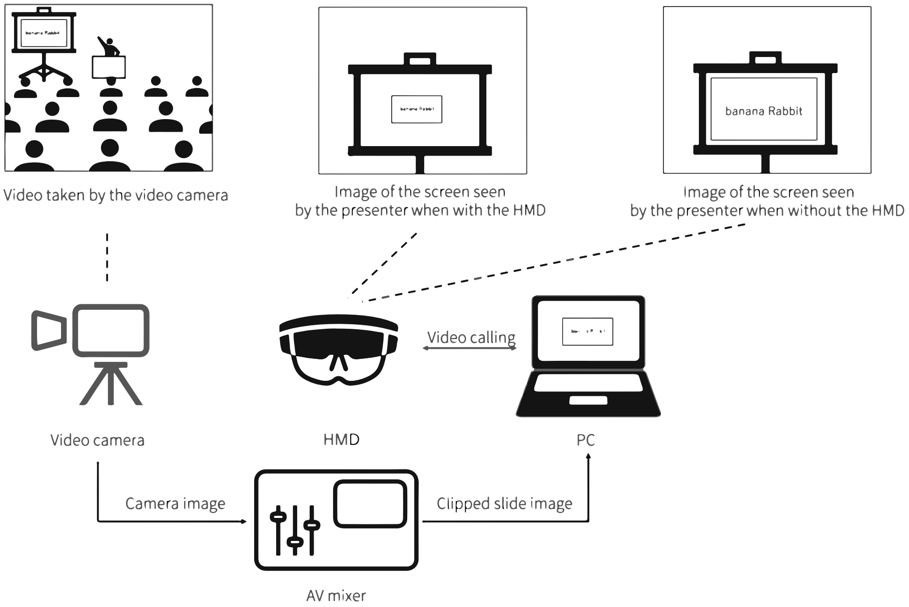

Video taken by the video camera
Image of the screen seen by the presenter when with the HMD
Image of the screen seen by the presenter when without the HMD
banana Rabbit
Video calling
Video camera
HMD
PC
Camera image
Clipped slide image
AV mixer

Fig. 8. Experimental system

the slides in the HMD was adjusted so that they did not appear to be floating away from the screen. The magnification of the slides displayed on the projector at the beginning of the trial was 100%, and the presenter could change the magnification freely during the trial.

We used an Apple Macbook as the PC, which was used to display the slides and to talk to the HMD, and a Microsoft HoloLens as the HMD, which was used to insert a part of another person's visual image into the presenter's vision. As a camera, a SONY HANDYCAM HDR-CX680 was used. The camera was used to record the experiment. The EPSON EB-940 of was used as a projector to project the slides on the screen.

Method without HMD: The method without HMD does not present the visual images of others that were presented in the method with HMD.

5.1 Result

The average magnification ratio for each graph is shown in Fig. 9. For the three types of slide (text, map, and graph), the average magnification was larger for the method with HMD than for the method without HMD. Especially, the average magnification of the map was the largest. In addition, the standard deviation of the map slide was larger than that of the text and graph slides.

For the text slides, the average magnification for the method without HMD was 114%, and the average magnification for the method with HMD was 157%, with the average magnification for the method with HMD being 1.4 times larger.

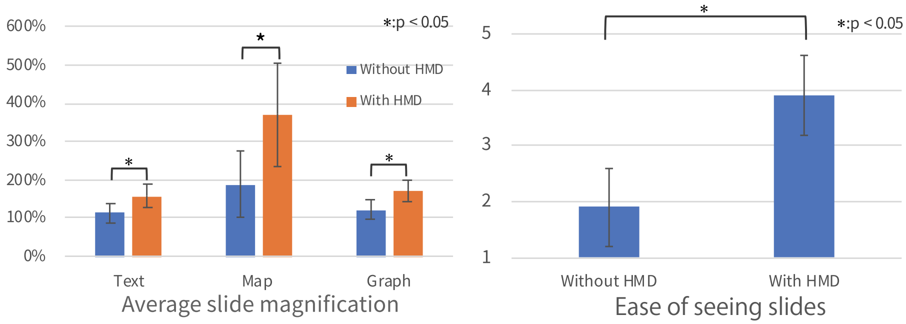


Fig. 9. Average slide magnification

When a t-test was applied at a significance level of 5%, a significant difference ($t(5) = 8.37, p = 0.0004$) was observed between the method with HMD and the trial without HMD.

For the map slide, the average magnification for the method without HMD was 189%, and the average magnification for the method with HMD was 372%; the average magnification for the method with HMD was 2.0 times larger. When a t-test was applied at a significance level of 5%, a significant difference ($t(5) = 5.04, p = 0.0040$) was observed between the method with HMD and the method without HMD.

For the graph slide, the average magnification for the method without HMD is 122%, and the average magnification for the method with HMD is 173%, which is 1.4 times larger than the average magnification for the method with HMD. When a t-test was applied at the 5% level of significance, a significant difference ($t(5) = 5.30, p = 0.0032$) was observed between the method with HMD and the method without HMD.

In terms of the ease of viewing the slides during the presentation, the mean for the method without HMD was 1.9 and the mean for the method with HMD was 3.9. As a result of applying the Mann-Whitney U test, a significant difference was observed between the method without HMD ($M = 1.9, SD = 0.7$) and the method with HMD ($M = 3.9, SD = 0.7$) ($U = 3.00, p < 0.01$). The higher the value, the more difficult it was to see the slides, meaning that the slides were easier to see in the method with HMD than in the method without HMD.

The presenters were interviewed about whether the HMD method helped their presentations. They answered that they enlarged the slides because the text was smaller than they expected (Presenter 1, Presenter 2, Presenter 4, Presenter 6, Presenter 7, Presenter 8, Presenter 9, Presenter 10), and that they wanted to enlarge the text on the map slides more because they felt the text was particularly small (Presenter 3, Presenter 5).

5.2 Discussion

For all three slides, the average magnification ratio of the method with HMD was larger than the average magnification ratio of the method without HMD, and the presentation method was changed by wearing an HMD into which was inserted the visual image seen by the audience seated in the back seats. For the map slide, the magnification rate was 189% for the method without HMD and 372% for the method with HMD, which was the largest among the three types of slides. The standard deviation of the map slide was larger than that of the other two slides for both the method without HMD and the method with HMD. This is thought to be because the presenters used different methods to enlarge the map. For example, some presenters enlarged the starting point, destination, and landmarks, in particular, while others showed the entire route from the starting point to the destination. In addition, as a result of interviewing about whether the method with HMD helped the presentation, two subjects answered that they thought to enlarge the map slides in particular. This provides another reason for the standard deviation of the map slides becoming larger. In contrast, many of the presenters enlarged the text slides and graph slides within the range where the whole image could be seen, rather than enlarging one part in particular. Therefore, the standard deviation was smaller.

As a result of interviewing about whether the method with HMD supported the presentation, four subjects answered that they enlarged the slides because the text looked small. Therefore, it can be said that the method with HMD supported the presentation.

6 Conclusion

In this study, we proposed PHSIL, a method of controlling user behavior by feeding back to the user the results of simulating the sensations of others. We also conducted a cognitive psychology experiment in which we controlled the volume of the presenter's voice and the magnification rate of the slides. Although no difference was observed between the fixed noise method and the existing method for controlling the volume of speech, the PHSIL noise method increased the duration of the increased speech volume more than the comparison method. For slide magnification control, the average magnification rate of the method with HMD was significantly higher than that of the method without HMD.

In the future, we plan to conduct evaluation experiments in a more realistic experimental environment, such as with a larger audience, with different presentation contents and presentation skills, and to develop techniques for estimating the volume of the presenter's voice and the text in the slides received by the audience.

References

1. Hara, F., Takegawa, Y., Hirata, K.: Design and implementation of a voice feedback device for voice loudness control. In: Munekata, N., Kunita, I., Hoshino, J. (eds.) ICEC 2017. LNCS, vol. 10507, pp. 81–87. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66715-7_9
2. Maeda, H., Kurihara, K., Homei, M.: Proposal for presentation style of presenter standing in front of the screen. In: Information Processing Society of Japan, Entertainment Computing (Published in Japanese), pp. 431–434 (2011)

3. Kurihara, K., Tsukada, K.: Speechjammer: a system utilizing artificial speech disturbance with delayed auditory feedback. arXiv preprint arXiv:1202.6106 (2012)
4. Daniels, J.: Feasibility testing of a wearable behavioral aid for social learning in children with autism. Appl. Clin. Inform. **9**(1), 129 (2018)
5. Narumi, T., Ban, Y., Kajinami, T., Tanikawa, T., Hirose, M.: Augmented perception of satiety: controlling food consumption by changing apparent size of food with augmented reality. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 109–118 (2012)
6. Rekimoto, J.: Traxion: a tactile interaction device with virtual force sensation. In: ACM SIGGRAPH 2014 Emerging Technologies, p. 1 (2014)
7. Watanabe, J., Ando, H., Maeda, T.: Shoe-shaped interface for inducing a walking cycle. In: Proceedings of the 2005 International Conference on Augmented Tele-Existence, pp. 30–34 (2005)
8. Suzuki, D., Takegawa, Y., Terada, T., Tsukamoto, M.: A heart rate presentation system for keeping music tempo in live performance. In: 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE), pp. 177–181. IEEE (2013)
9. Tjandra, A., Sakti, S., Nakamura, S.: Listening while speaking: speech chain by deep learning. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 301–308. IEEE (2017)
10. Lane, H., Tranel, B.: The lombard sign and the role of hearing in speech. J. Speech Hear. Res. **14**(4), 677–709 (1971)
11. Hodoshima, N., Arai, T., Kurisu, K.: Intelligibility of speech spoken in noise and reverberation. In: Proceedings of ICA, vol. 151 (2010)
12. Kogure, K., Yoshinaga, M., Suzuki, H., Kitahara, T.: A spoken dialogue system for noisy environment. In: Stephanidis, C. (ed.) HCI 2014. CCIS, vol. 435, pp. 577–582. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07854-0_100
13. Kawahara, H.: Transformed auditory feedback: effects of fundamental frequency perturbation. J. Acoust. Soc. Am. **94**(3), 1883–1884 (1993)
14. Lee, B.S.: Effects of delayed speech feedback. J. Acoust. Soc. Am. **22**(6), 824–826 (1950)
15. Stuart, A., Kalinowski, J., Rastatter, M.P.: Effect of monaural and binaural altered auditory feedback on stuttering frequency. J. Acoust. Soc. Am. **101**(6), 3806–3809 (1997)
16. Ryan, B.P., Van Kirk, B.: The establishment, transfer, and maintenance of fluent speech in 50 stutterers using delayed auditory feedback and operant procedures. J. Speech Hear. Disord. **39**(1), 3–10 (1974)
17. Yataka, S., Tanaka, K., Terada, T., Tsukamoto, M.: A context-aware audio presentation method in wearable computing. In: Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 405–412 (2011)
18. Hachiya, K.: Inter dis-communication machine. In: The Virtual Reality Society of Japan [Published in Japanese], vol. 1, pp. 117–118 (1996)
19. Higuchi, K., Fujii, K., Rekimoto, J.: Flying head: a head-synchronization mechanism for flying telepresence. In: 2013 23rd International Conference on Artificial Reality and Telexistence (ICAT), pp. 28–34. IEEE (2013)
20. Gruenefeld, U., Ennenga, D., Ali, A.E., Heuten, W., Boll, S Eyesee360: designing a visualization technique for out-of-view objects in head-mounted augmented reality. In: Proceedings of the 5th Symposium on Spatial User Interaction, pp. 109–118 (2017)
21. Denes, P.B., Pinson. E.N.: The speech chain: The physics and biology of spoken language. bell telephone laboratories. Inc. Baltimore, Maryland: Waverly Press, Inc (1963)

The Design Challenges of Drone Swarm Control

Fabrice Saffre[1](✉), Hanno Hildmann[2], and Hannu Karvonen[3]

[1] VTT – Technical Research Centre of Finland Ltd.,
MIKES bldg, Tekniikantie 1, 02150 Espoo, Finland
fabrice.saffre@vtt.fi

[2] TNO – Netherlands Organisation for Applied Scientific Research,
Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands
hanno.hildmann@tno.nl

[3] VTT – Technical Research Centre of Finland Ltd.,
FutureHub bldg, Tekniikantie 21, 02150 Espoo, Finland
hannu.karvonen@vtt.fi

Abstract. It is widely recognised that swarms are the likely next step for Unmanned Aerial Vehicle (UAV) or drone technology. Although substantially increased autonomy for navigation, data collection and decision-making is very much part of the *"collective artificial intelligence"* vision, this expected development raises questions about the most productive form of interaction between the swarm and its human operator(s). On the one hand, low-level *"micro-management"* of every unit clearly nullifies many of the advantages of using swarms. On the other, retaining an ability to exercise some control over the swarm's objectives and real-time behaviour is obviously paramount. We present two families of control methods, direct and indirect, that we believe could be used to design suitable, i.e. simultaneously intuitive, easy to use, powerful and flexible, Graphical User Interfaces (GUI) that would allow a single operator to choreograph a swarm's actions. Simulation results are used to illustrate the concept and perform a quantitative performance analysis of both control methods in different scenarios. Human factors aspects related to drone swarm control are identified and both control methods are discussed from the human operator's usage point of view. We conclude that the direct approach is more suitable over short time-scales (*"tactical"* level), whilst indirect methods allow to specify more abstract long-term objectives (*"operational"* level), making them naturally complementary.

Keywords: Drone swarms · Direct control · Indirect control · Human factors · Autonomous systems · Self-organisation

The Academy of Finland is acknowledged for financial support of project 'Finnish UAV Ecosystem' 337878 (FUAVE, project grant number 337878). This research was also partly funded by the Scientific Advisory Board for Defence (MATINE) in the CounterSwarm project.

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 408–426, 2021.
https://doi.org/10.1007/978-3-030-77932-0_32

1 Introduction

Drones (or unmanned aerial vehicles, UAVs) are increasingly becoming an everyday tool [9] in a growing number of application areas [17], as a convenient and cost-efficient way of remote sensing [4] or to gather information best acquired from a vantage point not easily accessible by other means (e.g., [2,18]).

In the vast majority of cases, these benefits are achieved by combining remote control by a human operator with relatively simple autonomous features, such as attitude control [22,24], obstacle avoidance, and return-to-base functions. In the case of beyond visual line of sight (BVLOS) operations, the assumption is often that the human in the loop uses a real-time video feed to pilot the drone as if he/she were on board (see Fig. 1).

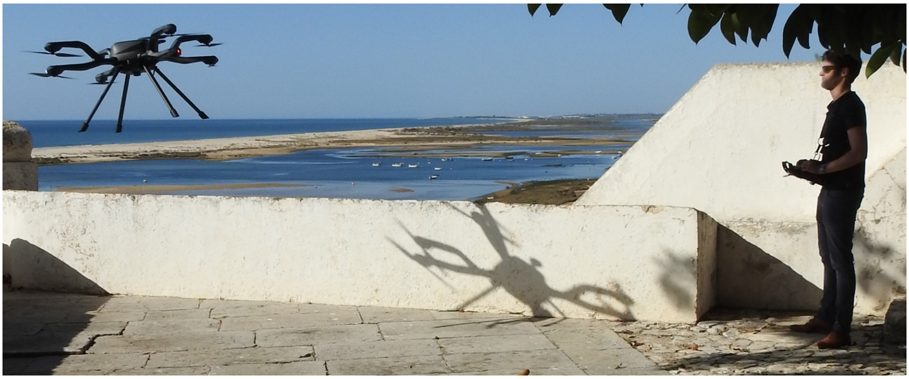


Fig. 1. A TNO reconnaissance drone departing for BVLOS flight during the (now concluded) EU funded ALFA (Advanced Low Flying Aircrafts Detection and Tracking) project [5] where it facilitates automatic threat evaluation for border security and surveillance. The pilot uses the control interface to *see* what the drone can see.

However, another paradigm is gaining momentum that is poised to challenge this mode of operation: fleets or swarms of drones. Simple common sense suggests that it is possible to achieve more with several units working as a team than with a single UAV and, accordingly, that an ability to leverage drone swarms would have a multiplicative effect on the usefulness of the technology. However, this trivial statement hides the considerable underlying complexity of orchestrating or choreographing the joint operation of a collective dozens or hundreds strong.

In the case of exploration or surveillance, for instance, it is obvious that n identical drones have the potential to cover an area roughly n times larger in the same amount of time as a single unit (notwithstanding restrictions imposed by battery life or shared point of origin). However, this implicitly assumes division of labour, i.e., no or limited overlap between flight paths, to avoid the duplication of effort. Even in this simplest of cases and without any real-time change of objectives, this would require assigning each drone to a particular zone in the area of interest. Furthermore, short of having one human operator per drone, each one would have to fly its respective patrol route autonomously, without real-time supervision. This in itself poses various safety concerns [20].

In the area of drone swarms, there are a number of concrete research questions [9]. For example, what interface design would best allow the user to monitor and operate a drone swarm? Furthermore, is there a suitable trade-off between micro-management (i.e., directly piloting every unit in real-time) and assigning global objectives to the entire fleet? And if the answer to the previous question is 'yes', what autonomous features are required and what kind of useful collective behaviour is possible? Finally, what control functions would allow to the operator of a drone swarm to achieve maximum utility for minimum complexity?

As a starting point, it does seem useful and necessary to create a *"taxonomy"* of control functions and associated distributed algorithms for drone swarms. A first distinction would be between those that involve directly piloting a subset of units (possibly a single one) and those that instead specify abstract collective goals. It is worth noting that these two approaches, although functionally different, are not mutually exclusive and could govern different aspects of fleet operations in parallel/at the same time.

The former (direct control) could be used on a short time-scale to coordinate the movement of the swarm into a chosen direction through formation flying. In this scenario, the human operator pilots the remote-controlled *"leader"* unit and the other members of the swarm use simple autonomous features (relative positioning) to spread out around it, forming a pattern. This would result in the type of collective behaviour most often cited as an example of swarming, i.e., flocking. More advanced functions could be introduced in the form of a simple ability to update key parameter values in real-time. For instance, how tight or loose the formation is could easily be controlled by fine-tuning the separation distance. Other basic commands could involve, e.g., instructions to follow, spread around or align with the leader perpendicularly to the direction of movement. Similar commands could be used to control altitude, determining, for instance, whether *"subordinate"* units should distribute themselves in the same horizontal plane as the leader or form a 3D lattice (with upper and lower bounds).

The latter (indirect control through abstract collective goals) is likely to be more useful on a longer time-scale and/or when the swarm is expected to fulfil its mission without direct human control or supervision. For instance, the drone *"colony"* could be tasked with patrolling a region of interest over an extended period (hours, days, or more). This would require much more complex autonomous features in the form of decentralised resource-management and collective decision-making to ensure that the airborne contingent balances the need to recharge with that to visit every part of the target area regularly. In this scenario, the challenge is to create an interface that allows the user to specify and subsequently update such abstract goals intuitively. For example, the ability to *"paint"* a region of arbitrary size and shape on a digital map to designate it as being of interest and communicate this information to the swarm would be paramount. In both direct and indirect control, human factors aspects need to be considered in the design of the system and its related user interfaces.

2 Human Factors Aspects in Drone Swarm Control

As has occurred in many other highly automated domains (e.g., industrial process control, ship navigation, and traditional aviation), conducting drone operations will also eventually evolve to become more of a supervisory task than an active manual control. Consequently, similar human factors problems that have been identified in these other domains will then become prevalent. These issues include, but are not limited to, operator trust in automation [23], excessive mental workload in exception situations [25], situation awareness issues [12,13], operator boredom [11], work vigilance [6] and the integrity of the artificial intelligence [26]. Additionally, with higher automation levels supported with AI, the supervisory control of multiple UAVs will become possible. To approach the human factors problems of controlling a drone swarm, various approaches have been discussed in the literature, only a fraction of which is discussed here.

To analyse, optimise and divide the tasks to be conducted by humans and by the AI/automated system in drone swarm control, task/work analyses are a key approach. They are especially suitable to consider the human limitations and support the definition of human operator's meaningful tasks and decision-making in the design phase of the system [1]. As a one guideline for design when considering human-automation task allocation in drone swarm control, it has been found out that instead of *management by consent* (automation as an assistant to the operator), *management by exception* improves the operator's performance [10]. According to [16], management by exception means that the automation decides to take an action based on some set of predetermined criteria and gives operators only a chance to veto the automation's decision.

In systems engineering, defining a Concept of Operations (ConOps) for the drone swarm control in the start of the system design is essential [14]. From the human factors point of view, this ConOps typically includes a clear description of, for example, the division of tasks between human and the automation, operator tools, roles/responsibilities, and procedures [3]. This ConOps should work as a boundary object (see, e.g., [29]) and allow the relevant stakeholders (e.g., engineers, users, and designers) to discuss about the system under development and the related aims in a manner understandable for all involved parties.

On a methodological level, for example, ecological interface design (EID) has been successfully applied to the control of UAV swarms by [15]. The results of this study showed that EID-inspired interface design enabled operators to control a drone swarm and successfully resolve failures during mission execution [15]. Particularly, the ecological interface designs promoted creative problem-solving activities to scenarios that could not have been solved by following a fixed procedure (see details in [15]).

Regarding detailed control systems, for example, [19] have provided a review of human-system interface (HSI) solutions for the management of swarms of drones. Their main conclusion from this review was that allowing user and mission-specific customization to user interfaces and raising the swarm's level of autonomy to reduce the cognitive workload shouldered by the operator are beneficial and improve operators' situation awareness [19].

There have also been some studies about different interaction modalities for drone swarm control, such as gesture and touch [21]. However, the benefits of user interfaces with novel input paradigms remains unclear compared to traditional point-and-click interfaces. In general, in drone swarm control HSI design it is important to ensure that the operator has adequate means to first observe and then direct the automation's functioning in order to be responsive to potential situation specific changes [8]. Therefore, both the hardware and software solutions developed to monitor and control the swarm need to be suitable for the specific situation and task at hand without cumbersome interaction solutions that may hinder safe and efficient operations.

3 Direct Control Methods

The most straightforward method for controlling a fleet of semi-autonomous UAVs, at least over short time-scales (of the order of a drone's battery life), is to pilot one or a few participating units directly and use parameterised formation flying to orchestrate the collective behaviour of the rest of the swarm. This "leader" may very well differ from the other members of the *"flock"*. For instance, it could be a special drone with enhanced capabilities (e.g., for reliable long-range communication and telemetry), or a manned aircraft that the swarm is meant to escort or *"extend"* (acting as a network of distributed sensors).

Notwithstanding such specifics, the difficulty lies in identifying the right balance between precise control and ease of use. The human operating the swarm may have limited time or cognitive resources to allocate to this task, as other activities may require his or her urgent attention. Therefore, to minimise mental workload, orchestrating the UAV collective should be as easy and seamless as possible instead of a fastidious exercise with real time fine-tuning. To use a common metaphor: the swarm should *"feel"* like a mere extension of the user's own sensing and actuation capabilities. This aim means that identifying the right parameters, a range of suitable values for them, and a suite of intuitive, user-friendly tools to pick or change one or more of these parameter values is paramount. Clear and illustrative visualisations of the swarm's behaviour as well as *'what if'* scenarios in the user interfaces are also essential for fluent operations.

There is no proven or *"one size fits all"* approach to solving this conundrum. The framework we present here is meant as an illustration of a possible *"swarm interface"* design, not as a final product. Different applications will undoubtedly require specific additional functions that we do not discuss here, as they would require restricting our findings to a particular mission-specific domain.

UAVs, as physical devices, are defined by hardware characteristics that act as constraints over what can and cannot be achieved. Some of these traits, such as, for example, battery life, are of critical importance when considering longer time-scales (days or weeks) but are not particularly relevant for short periods (which we previously argued is the context in which direct control methods are likely to be most useful).

Other limiting factors also play an essential role in formation flying, such as:

- Positioning accuracy
- Maximum speed
- Maximum acceleration
- Sensor/communication range

Here, it is good to note that fixed-wing aircraft represent a special case in that they rely on lift to stay airborne (and so have a minimum speed threshold too) and cannot accelerate in any direction, but rely on course correction. What follows assumes that the swarm is comprised of rotary-wing drones (e.g., quadcopters or (remote controlled or autonomous) helicopters [7]).

Positioning accuracy is a tricky parameter to take into account, but it is mostly relevant for close formations such as the ones used for aerobatics display. In most other applications, the target separation distance between units will be considerably higher (sometimes by over one order of magnitude) than the average positioning error. For instance, GPS is typically accurate down to a few meters, so if drones are attempting to maintain a separation of 50+ m even based solely on broadcasted GPS coordinates, the error will already not have much impact. Furthermore, other methods such as dead reckoning, radio signal triangulation and attenuation analysis, or even real-time computer vision can be used to improve accuracy of (relative) positioning. In the remainder of this section, the coordinates are assumed to be the ones reported by the UAVs, which means that the actual location might slightly differ, but not to the extent that it will significantly affect the configuration of the swarm.

Speed and acceleration are another matter. Not only do their maximum possible values, imposed by aerodynamic characteristics, represent an upper bound to which instructions given to the swarm must conform, they also less directly but significantly affect flying patterns.

Our simulation results suggest that active/deliberate modulation of the target speed (V) and acceleration (*delta-*V) by the swarm operator can be one of the simplest and most efficient way to control collective behaviour through a single pair of parameters. It is important to understand here that, whereas *maximum* values are hardware constraints, *target* acceleration and speed are not. For instance, a quadcopter may not be capable of accelerating at a faster rate than $10\frac{\mathrm{m}}{\mathrm{s}^2}$ but there is nothing preventing on-board software from targeting a *lower* value if so instructed. Since lower acceleration means greater inertia, under the influence of the same *"forces"* (e.g., a tropism), different target *delta-*Vs will result in different movement patterns.

If we consider the case of a single rotary wing UAV, acceleration is the result of its tilting into the direction of intended movement, converting some lift into thrust. The angle of the tilt, combined with a possible change in the speed of the rotors/blades, is what determines the value of the *delta-*V. If we hypothesise that the rotors automatically adjust their speed to keep lift (and so altitude) constant by default, then acceleration can be controlled via tilt angle only.

For instance, a drone might have a maximum *"safe"* tilting angle (determined by physical constraints such as the top speed at which rotors can spin) allowing for a maximum horizontal acceleration of $10\frac{\mathrm{m}}{\mathrm{s}^2}$. At this angle, it will take such a device 1 s to go from hovering (stationary) to traveling at a target velocity of $10\frac{\mathrm{m}}{\mathrm{s}}$, during which time it will have travelled 5 m into the intended direction. It will take another 1 s (and another 5 m) for it to slow down and return to hovering mode (N.B. these figures assume a negligible drag coefficient).

Let us consider the situation in which a drone traveling East at a cruising speed of $10\frac{\mathrm{m}}{\mathrm{s}}$ is instructed to change direction and go North, using the maximum allowable acceleration. Such a device will first need to stop its Eastward movement (1 s, 5 m), then accelerate North (again 1 s, 5 m). The change of course takes 2 s and the drone will reach its new heading and cruising speed at a point 5 m East (slowing down) and 5 m North (reaccelerating) of the position where the instruction was received. Had the tilting angle been such that the acceleration rate was only $5\frac{\mathrm{m}}{\mathrm{s}^2}$ instead of $10\frac{\mathrm{m}}{\mathrm{s}^2}$, the same manoeuvre would have taken twice the time (4 s) and the drift east during deceleration would have been 10 m.

3.1 Two Drones Interaction

These relatively trivial calculations become more subtle when the intended direction of travel depends on the relative location of multiple UAVs attempting to coordinate their movements to achieve a certain objective (e.g., spatial distribution pattern). Figures 2 and 3 visualize these interaction dynamics.

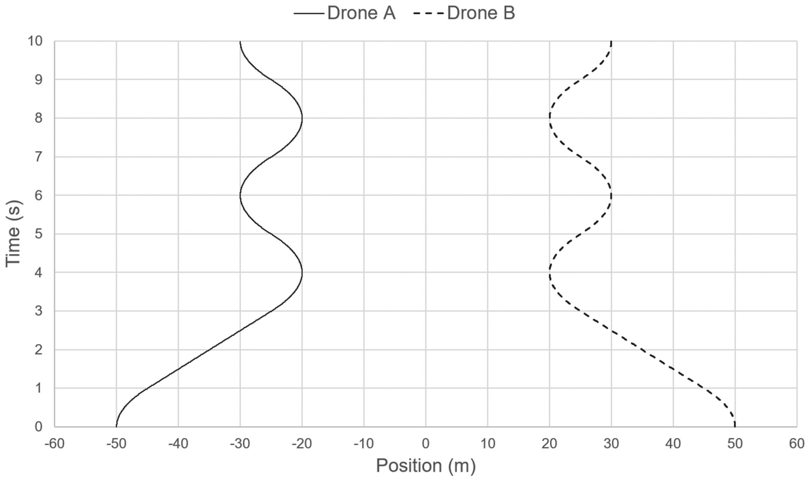


Fig. 2. Visualisation of the two drones interaction dynamics: the evolution of the position over time on the axis the origin of which is the mid-point between the two units (cf. Fig. 3 for speed vs. position). The oscillatory regime after 2.5 s is clearly visible.

For example, if two drones start hovering 100 m apart and the target separation distance is 50 m, they will accelerate toward each other, say at the maximum allowable *delta-V* $(10\frac{\mathrm{m}}{\mathrm{s}^2})$. They will reach their cruising speed after 1 s, having travelled 5 m each, in opposite directions, i.e., 90 m apart. Another two seconds later, they will have travelled another 20 m each and be 50 m apart as instructed. However, at this point, they are flying toward each other at $10\frac{\mathrm{m}}{\mathrm{s}}$. If they start decelerating immediately, they will be only 40 m apart by the time they are stationary (overshooting the target separation). Intuitively, this means that they should now accelerate in the opposite direction to open the gap. If they do so, they will be 50 m from each other again one second later, but traveling at $10\frac{\mathrm{m}}{\mathrm{s}}$. This will initiate an oscillation around the target separation distance, with an amplitude of 20 m. This is illustrated in Figs. 2 and 3.

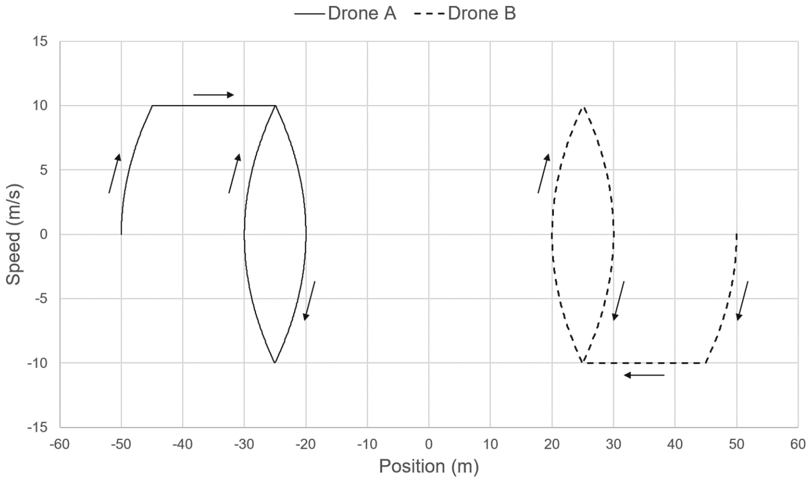


Fig. 3. Visualisation of the two drones interaction dynamics: the speed vs. position (for comparison, cf. Fig. 2 for the evolution of the position over time). The *"spindle-shaped"* part corresponds to the oscillation, when the two drones travel back and forth around the equilibrium point. Arrows indicate the passage of time (the closed and indefinitely repeating loop corresponding to the oscillatory regime).

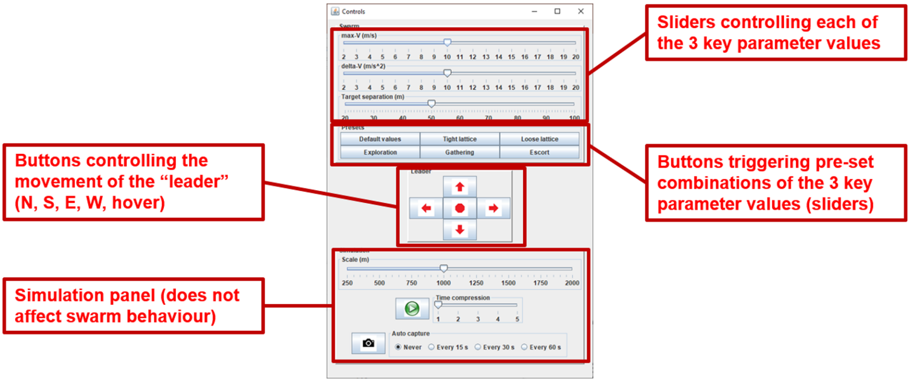
Had they been limited to a tilt angle and acceleration of $5\frac{\mathrm{m}}{\mathrm{s}^2}$, the scenario would be qualitatively identical but quantitatively different. Indeed, the two UAVs would have reached their cruising speed $(10\frac{\mathrm{m}}{\mathrm{s}})$ not one but two seconds later and 80 m apart instead of 90. Having reached the target separation distance 1.5 s later instead of 2, they would require another two seconds to come to a halt, closing the distance between them to 30 m (instead of 40). Accelerating away from each other, still at the same reduced rate of $5\frac{\mathrm{m}}{\mathrm{s}^2}$, then slowing down after reaching the 50 m separation target, they will be 70 m apart before they have returned to hovering mode. The oscillation is now double the amplitude, the distance between the two drones varying between 30 m and 70 m. The perhaps somewhat counter-intuitive conclusion is that a higher acceleration rate results in a reduced deviation from the target separation at steady state.

These effects could of course be counteracted by simply anticipating the overshoot and decelerating pre-emptively, but what would be an easy calculation in the above example (with just two drones traveling along a single dimension) becomes almost impossible in a swarm dozens of units strong, *"pushing and pulling"* each other in 2D or 3D space.

3.2 Multiple Drones Interaction

We used simulation extensively to investigate the collective dynamics of a swarm controlled by combining direct piloting of a single unit (*"leader"*) with real-time modification of the aforementioned parameter values (target speed, acceleration and separation). The results of these numerical experiments are summarised in this section. We present the user interface (Fig. 4) as well as screenshots of the swarm in various configurations together with the corresponding parameter settings (in Figs. 6 and 7), to emphasise the link between them.

In order to ensure reproducibility of our findings, we must first disclose the *"hidden"* rules of interaction, i.e., those that are, so to speak, *"hard-coded"* into the rules governing drone behaviour and not tied to any modifiable parameter. It is important to understand that these rules are ad-hoc in nature and that we make no claim of having investigated them in any meaningful way: they were simply found to be suitable to illustrate the type of collective activity pattern that the end user can elicit in the swarm by altering the value of the controllable parameters. The tunable parameters used are shown in Fig. 4.



Sliders controlling each of the 3 key parameter values

Buttons controlling the movement of the "leader" (N, S, E, W, hover)

Buttons triggering pre-set combinations of the 3 key parameter values (sliders)

Simulation panel (does not affect swarm behaviour)

Fig. 4. The user interface with the tunable parameters. Figures 6 and 7 show the sliders to indicate the varied parameters and their respective settings. This could be regarded as an early prototype for a real-world user interface, with the exception of the *"Leader"* sub-panel, which is a very crude approximation for an actual remote-control station.

We simulated a group of 37 drones, one of which is assumed to be remote-controlled and the other 36 semi-autonomous (in the sense that they plan and execute their own movement based on interaction rules and parameter values). The remote-controlled unit (swarm *"leader"*) depends entirely on the human pilot's instructions to perform any action other than hovering in position.

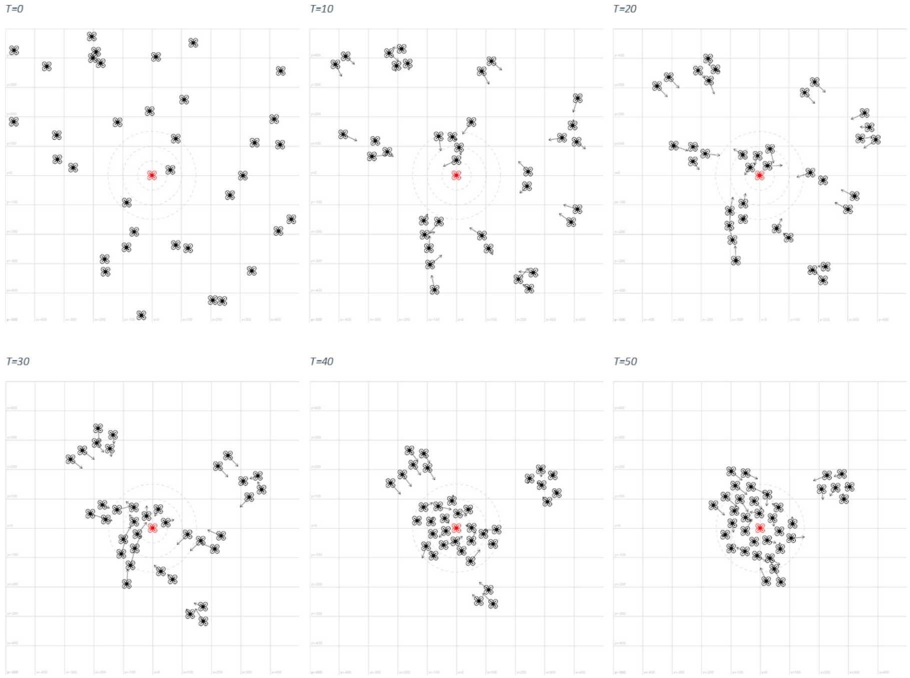


Fig. 5. Clustering process for the default parameter values ($maxV = 10\frac{\text{m}}{\text{s}}$, *delta-V* $= 10\frac{\text{m}}{\text{s}^2}$, separation = 50 m) from random initial locations (top-left corner). Screenshots are 10 s apart. Arrows are the speed vectors; the screen/display is always centered on the swarm leader, indicated in red. (Color figure online)

We hypothesise that all drones are constantly broadcasting their location and that they can all reliably communicate with each other (so every member of the swarm can potentially use the location of any other as input for planning its own movement). In the chosen rule-set, every drone is only using the location of the leader and of its two nearest neighbours (the identity of the two nearest neighbours may of course change over time as the swarm reconfigures itself).

The location of these three other units influences path planning as follows:

- Each one exerts a force inversely proportional to the distance (i.e., $\frac{1}{r}$).
- For distances < target separation this is repulsive, otherwise attractive.
- If the distance falls below half the target separation, this force is multiplied by 10 and a flag is raised (similar to a collision avoidance mechanism).

- The three attraction/repulsion vectors are then added up and the resulting vector is normalised (indicating the direction of acceleration, not intensity).
- This vector is then multiplied by the chosen acceleration rate (*delta-*V) unless the *"collision avoidance"* flag is raised, in which case the maximum value is always used ($20\frac{\mathrm{m}}{\mathrm{s}^2}$ in our experiment). NB: because of the $\frac{1}{r}$ rule and of the multiplicative factor applied to the repulsion force exerted by a neighbour less than half the target separation away, when in danger of collision, this vector tends to point directly away from the nearest neighbour.

The resulting acceleration vector is then used to update the drone's airspeed, until the target velocity is reached. This is done synchronously (i.e., all speeds and positions are updated simultaneously), with an integration step of 0.01 s.

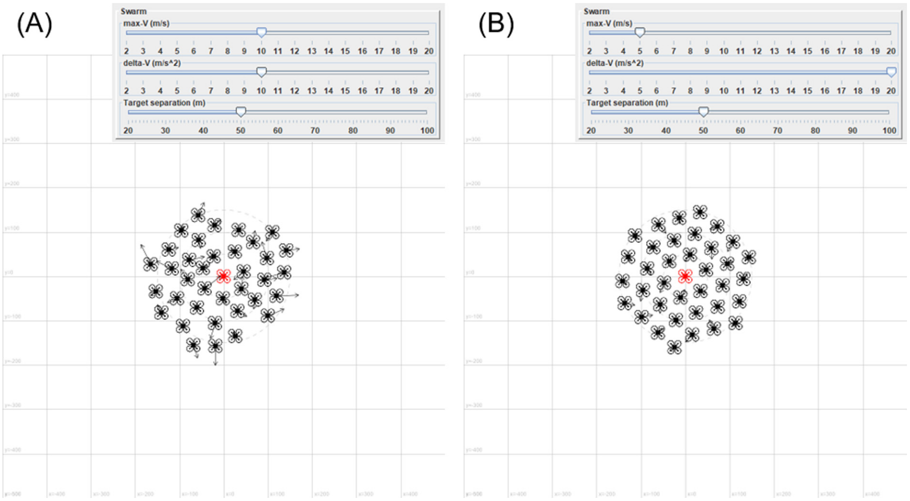


Fig. 6. Controlling the swarm through global parameters only. Each screenshot is a typical illustration of the type of distribution pattern observed at steady state for different combinations of values. Shown are **(A):** *"Flocking"*, **(B):** tight regular lattice.

Unless specified otherwise, units start at a random location within one square kilometre centred on the leader (coordinates origin $x = 0$, $y = 0$), with the only constraint that no two drones can be closer to each other than half the default target separation (25 m). The default values for acceleration and maximum (or target) speed are $10\frac{\mathrm{m}}{\mathrm{s}^2}$ and $10\frac{\mathrm{m}}{\mathrm{s}}$. The sequence shown in Fig. 5 illustrates the typical clustering process taking place if no actions are taken by the operator (i.e., all parameters are at their default value and the leader remains stationary).

As illustrated in Figs. 6 and 7 it is possible to control the behaviour of the swarm by changing the values of the three previously identified key parameters for relative movement and positioning ($maxV$, *delta-*V and separation).

Some changes will result in a *"quantitative"* difference (e.g., tighter or looser lattice, cf. Fig. 6(B), 7(A)) others may bring about a *"qualitative"* change (e.g., from holding position, Fig. 7(A), to adopting a random search pattern, Fig. 7(B)).

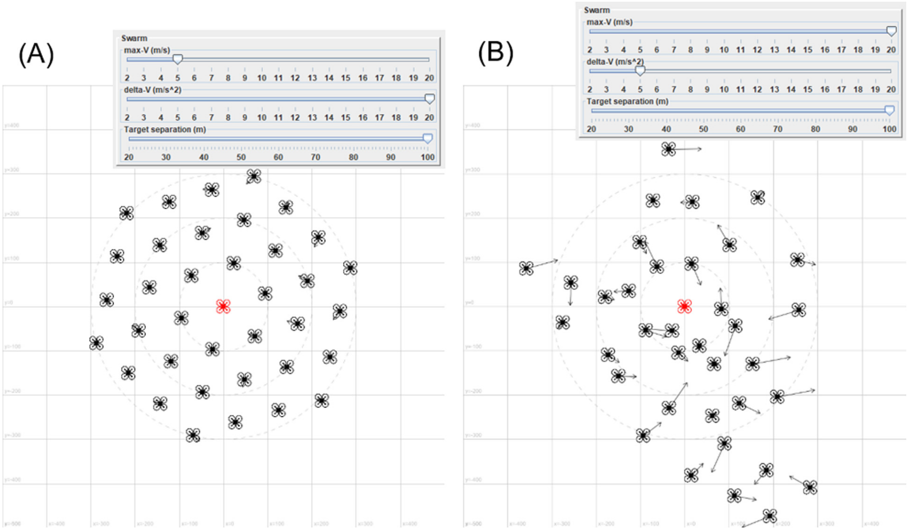


Fig. 7. Controlling the swarm through global parameters only (continuation of Fig. 6). Each screenshot is a typical illustration of the type of distribution pattern observed at steady state for different combinations of values. Shown are **(A):** loose regular lattice, **(B):** *"Exploration"*. The swarm will continously self-organize and transition, seemingly spontaneously, between these states as the human operator updates parameter value(s).

Figures 6 and 7 show typical examples. The precise influence of each parameter can be analysed using statistical methods, which should inform the design of the graphical user interface (e.g. by identifying suitable maximum and minimum parameter values). Such principled investigation is also critical to discover possibly *"unsafe"* combinations (e.g. if the ratio between $maxV$ and $delta\text{-}V$ is such that it increases the risk of collision beyond an acceptable level).

To illustrate this approach we present an analysis of a numerical experiment designed to study how changing the acceleration rate ($delta\text{-}V$) affects the swarm's ability to form and maintain a regular lattice. Figure 8, shows the resulting frequency distribution.

The results shown in Fig. 8 provide at least two useful pieces of information:

1. The distributed algorithm being used is capable of reliably generating a close approximation of a regular hexagonal lattice, which means that, by positioning the *"leader"* at its centre and letting the swarm self-organise, homogeneous coverage of a region of interest (the area of which is controlled solely

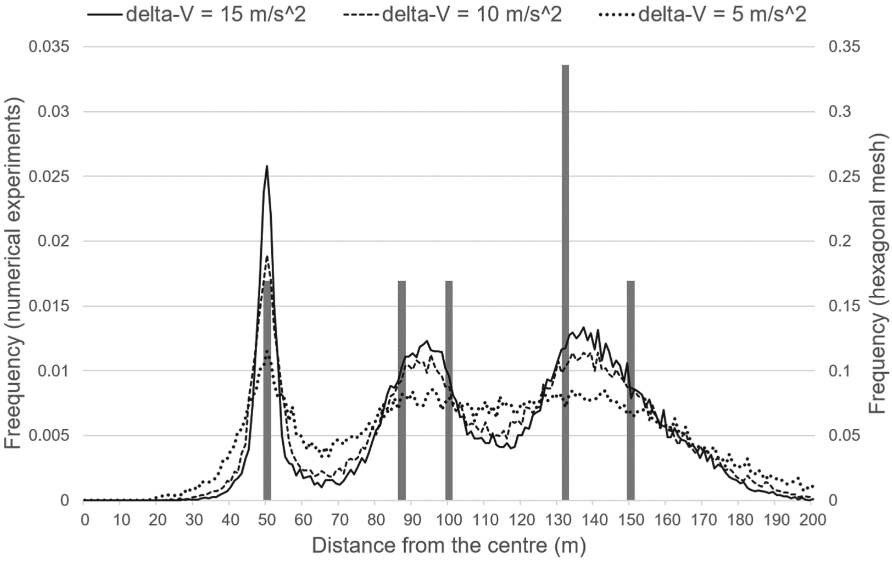


Fig. 8. Frequency distribution of the distance from the centre (*"leader"*), at or close to steady state ($10'$ after take-off), for a fleet of 37 drones and 3 different values of the *delta-V* parameter and for a target separation of 50 m (max. velocity = 10 $\frac{\text{m}}{\text{s}}$). The bars indicate the corresponding distribution in a perfect hexagonal mesh. 1000 independent realisations from randomised initial conditions per parameter value.

by the target separation distance) can be obtained. See Figs. 6(B) and 7(A) for an illustration.

2. *Delta-V* has a strong influence on the accuracy and stability of the lattice, with lower values ($5\frac{\text{m}}{\text{s}^2}$) resulting in the near disappearance of a discrete second and third *"rings"*, replaced by a more diffuse *"cloud"* in the same annular region (≈80–160 m). One can also observe that the effect of the acceleration parameter is nonlinear, with the difference between $10\frac{\text{m}}{\text{s}^2}$ and $15\frac{\text{m}}{\text{s}^2}$ being much lower than between $10\frac{\text{m}}{\text{s}^2}$ and $5\frac{\text{m}}{\text{s}^2}$.

3.3 Direct Control: Summary and Conclusions

The practical implication for the operator of a swarm is that by modulating the value of three global parameters shared by all drones, target separation distance, speed and acceleration, he/she can trigger a variety of collective movement patterns. A high *delta-V* and low target speed will tend to *"lock"* individual units in a regular mesh (in 2D, a hexagonal one) the density of which is controlled by the separation distance. Conversely, a high target speed and low acceleration rate will favour a much more dynamic and less predictable behaviour in which UAVs follow complicated *"orbits"* around the geometric centre of the swarm. Clearly, the former may be suitable for certain mission types (e.g., homogeneous coverage or systematic survey), the latter for others (e.g., escorting a potential target by scouting all possible attack vectors).

Critically, alternating between these two types of collective behaviour can be achieved simply by broadcasting updates to the corresponding parameter values without any need for *"micro-management"* or path-planning.

All this requires adaptivity from the user interface designed to supervise and control the swarm. For example, mission-specific customizations are needed particularly for the acceleration control. Visualisations of the effects of different parameter values are essential for the user to understand their effect in practice. Furthermore, to predict the effects before making the actual changes, simulation capabilities are needed to show the operator *'what if'* type of scenarios in order for optimal decision-making. Finally, to minimise mental workload, the user interface needs to support the operational situations in a sufficient level of abstraction for the human operator.

4 Indirect Control Methods

By indirect control methods, we refer to any algorithmic framework designed to make the swarm work towards achieving a certain goal without any real-time intervention by a human operator being necessary. In that sense, it is analogous to *"management-by-exception"* [10], whereby a system or organisation is capable of operating smoothly by default and only needs new instructions or temporary takeover occasionally, when circumstances or objectives change.

The chosen scenario to illustrate this concept is long-term surveillance, specific cases of which could be patrolling a remote border or protecting a wilderness area by deterring damaging human activity (e.g., poaching or illegal logging). Because such a mission would typically vastly exceed a drone's battery life, this implies the presence of at least one and probably several *"bases"* at which individual units can land and recharge. Simple navigation methods can be used to guarantee that a drone's flight path never exceeds its autonomy and that it can safely land either at its point of origin (round-trip) or at another base located at the end of a one-way flight.

The challenge is elsewhere and consists in finding ways of leveraging cooperative effects to ensure good coverage (i.e., no *"blind spots"*) and avoid duplication of effort (i.e., several drones patrolling the same area simultaneously).

We found that a suitable way to achieve cooperation with minimal need for explicit coordination and communication is to let all drones access a shared real-time simulation of the world, also known as so-called *Digital Twin* (DT), and modify it to exchange relevant information (a method of interaction known as *stigmergy*). However, the focus of the present paper being user interfaces and methods for orchestrating or choreographing swarms, we will not go into a detailed description and performance analysis here. This is the subject of a separate publication by the same authors team [28].

Our opinion is that, in the long-term surveillance scenario, interaction between the swarm and its human master should also take the form of modifications to the DT. For instance, the obvious first step would likely consist

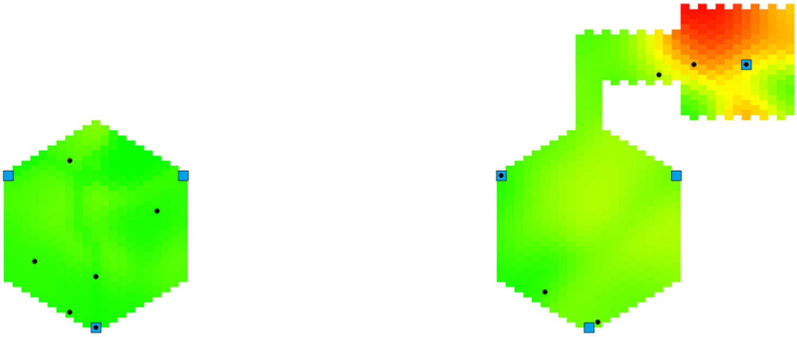


Fig. 9. Images from the user interface for a simulation of the proposed *control through area designation* paradigm. White indicates areas not under consideration, the colour spectrum indicates how close/recently a particular zone has been patrolled by at least one drone (green ≈ nearby and/or a short time ago, red ≈ far and/or a while ago. The panel on the left shows some initial area for the swarm to operate in. Shown are three bases (square, blue) and six drones (one of which is currently at a base). The panel on the right shows the same area after it has been extended towards the North East, using a *"drag-click"* tool of the interface. (Color figure online)

in designating the default target area for which the user wants the swarm to *"take responsibility"*. The easiest and most intuitive way to do so is to access the real-time simulation through a graphical interface in which standard *"paint"* or *"click-and-drag"* functions can be used to seamlessly add or subtract geographic zones to or from the region of interest. This is illustrated by Figs. 9 and 10.

Using this functionality, the human user has added a rectangular area to the North East as well as a connecting corridor. While the number of drones has remained the same, an additional location for a base has also been designated in the middle of the new area.

In essence, when preparing its upcoming flight plan, a departing drone accesses the DT and uses the information it contains to determine the most desirable patrol route at this point in time, based on a suitable utility function.

In the surveillance scenario, the objective is to ensure that no part of the designated area of interest (the boundaries of which can be modified at runtime as shown in Figs. 9 and 10) remains unobserved for too long, which is represented in the DT by a *"heat-map"*. This heat-map indicates the concentration of a diffusing virtual signal that is created when no drone is present and is removed when visited. So in order to maximise its contribution to the swarm's mission, a departing unit only has to *"climb"* the gradient to ensure that it is heading toward an area that needs inspection.

Accordingly, a human operator can also influence the swarm priorities by manipulating the signal. Removing a region from the swarm's territory is effectively done by removing and stopping production of the signal in the corresponding area. Symmetrically, adding a region takes the form of starting production of the signal where it was absent before. Obviously, dropping a large *"quantity"*

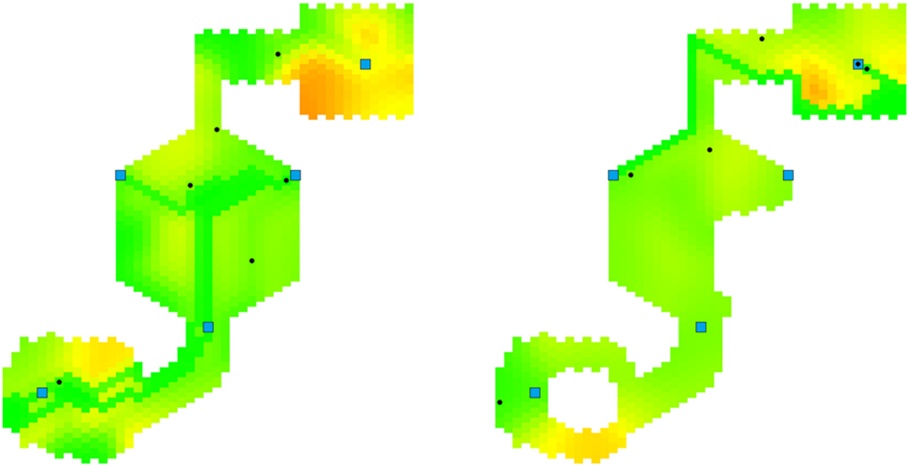


Fig. 10. *Control through area designation*: the area from Fig. 9 (right panel) is amended even further. The panel on the left further modifies the area using *"paint" toadd* an irregular region to the South-West. The panel on the right shows the area after using *"ctrl-paint" to remove* certain areas from the swarm's *"territory"*.

of the diffusing signal at a given location (*"honeypot"*) will result in a gradient that will attract drones to this area, a method that can be used to *"bias"* the swarm in favour of patrolling a particular zone. This can be regarded as the indirect control equivalent of an explicit command to *"go there"*.

5 Comparison and Conclusion

To conclude this paper we reflect on what as discussed and proposed. In line with the rest of the article, we are considering two main view points: that of someone concerned primarily with the control aspect of a swarm and the accompanying considerations regarding control, from a human factor point of view.

From the control methods point of view, perhaps the most useful lesson to be learned is that it is relatively straightforward to foster the emergence of desirable collective behaviour by leveraging self-organisation. This paradigm is exceedingly familiar in complexity science and its potential will come as no surprise to experts in the field, from Physics to Biology, but it still appears under-exploited in technology and engineering in general and in distributed robotics in particular. This could be about to change with the increasingly realistic prospect of drone swarms performing various useful tasks in the physical world. Quite simply, if we want to make the most of this opportunity to delegate various complex tasks to machines, self-organisation and collective artificial intelligence will cease to be an optional design choice to become a practical requirement.

It should however be remembered that *undesirable* emergent properties are as common as desirable ones, which is why the principled study of global dynamics in a system governed by local rules is extremely important, not least because they can be counter-intuitive or at least not obvious. We encountered this kind of unexpected behaviour during the course of the present investigation, which led to some adjustments. For instance, in the direct control algorithm, the decision to always include the *"leader"* in the trio of influencers was made after it was noticed that not doing so could result in the formation of isolated clusters. After modification, this effect is still present but now transient because the corresponding configuration is usually unstable (see intermediate stages in Fig. 5). Retrospectively, this could have been easily anticipated, but some emergent dynamics are more subtle.

This simulation-based *"trial-and-error"* approach, where candidate algorithms are tested, modified, then re-evaluated using quantitative and qualitative measures of performance, may appear ad-hoc in nature, but it is surprisingly effective. The reason is that, after being trained in complex systems modelling, a human designer can gain an intuitive understanding of what particular technique could solve a particular problem, much as a skilled craftsman knows what tool is most suitable to perform a certain task. We expect this type of expertise to become increasingly in demand over the coming decade, not only at design time but also for everyday operations. Indeed, the end user of a drone swarm will rely upon the same familiarity with emergent properties and self-organisation to orchestrate collective behaviour effectively and efficiently, particularly in the face of unexpected events or circumstances. This has clear implications for the design of user interfaces that are fit for purpose.

From a human factors perspective, some conclusions can be drawn related to both control methods. It is clear that direct control requires more active and hands-on user operation than indirect control. The direct control mode of operation may allow the operator to stay better *"in-the-loop"* as regards to what is happening with the swarm at each moment.

However, for efficient control of large swarms in complex operations, indirect control is needed to mitigate, for example, potential human operator workload issues. One clear future research issue in indirect control is in how to achieve the necessary level of operator situation awareness about what the swarm is currently doing. In addition, questions such as how to design the user interface to support operator SA at an appropriate level, how to support the calibration of user trust in the user interface, and how to make the indirect drone swarm control to be an engaging activity are relevant. When combining the direct and indirect control modes to a single user interface, the issues of mode transitions (see e.g., [27]) and handover implementation from indirect to direct control become essential. Clearly, more human factors research is needed in this challenging area.

References

1. Adams, J.A.: Cognitive task analysis for unmanned aerial system design. In: Valavanis, K.P., Vachtsevanos, G.J. (eds.) Handbook of Unmanned Aerial Vehicles, pp. 2425–2441. Springer, Dordrecht (2015). https://doi.org/10.1007/978-90-481-9707-1_26

2. Al-Kaff, A., Madridano, A., Campos, S., García, F., Martín, D., de la Escalera, A.: Emergency support unmanned aerial vehicle for forest fire surveillance. Electronics **9**(2) (2020). https://doi.org/10.3390/electronics9020260. https://www.mdpi.com/2079-9292/9/2/260
3. Bilimoria, K.D., Johnson, W.W., Schutte, P.C.: Conceptual framework for single pilot operations. In: Proceedings of the International Conference on Human-Computer Interaction in Aerospace, HCI-Aero 2014. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2669592.2669647
4. Boubeta-Puig, J., Moguel, E., Sánchez-Figueroa, F., Hernández, J., Preciado, J.C.: An autonomous UAV architecture for remote sensing and intelligent decision-making. IEEE Internet Comput. **22**(3), 6–15 (2018). https://doi.org/10.1109/MIC.2018.032501511
5. van den Broek, B., van der Velde, J., van den Baar, M., Nijsten, L., van Heijster, R.: Automatic threat evaluation for border security and surveillance. In: Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III, vol. 11166, pp. 113–122. International Society for Optics and Photonics, SPIE (2019). https://doi.org/10.1117/12.2532308
6. Cabon, P., Coblentz, A., Mollard, R., Foulliot, J.: Human vigilance in railway and long-haul flight operation. Ergonomics **36**(9), 1019–1033 (1993). https://doi.org/10.1080/00140139308967974, pMID: 8404831
7. Calero, D.P., et al.: Low altitude remote sensing. In: Meynart, R., Neeck, S.P., Shimoda, H., Habib, S. (eds.) Sensors, Systems, and Next-Generation Satellites XII, vol. 7106, pp. 353–363. International Society for Optics and Photonics, SPIE (2008). https://doi.org/10.1117/12.799809
8. Calhoun, G.L., Draper, M.H.: Display and control concepts for multi-UAV applications. In: Valavanis, K.P., Vachtsevanos, G.J. (eds.) Handbook of Unmanned Aerial Vehicles, pp. 2443–2473. Springer, Dordrecht (2015). https://doi.org/10.1007/978-90-481-9707-1_88
9. Coops, N., Goodbody, T., Cao, L.: Four steps to extend drone use in research. Nature **572**, 433–435 (2019). https://doi.org/10.1038/d41586-019-02474-y
10. Cummings, M.L., Mitchell, P.J.: Predicting controller capacity in supervisory control of multiple UAVs. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(2), 451–460 (2008). https://doi.org/10.1109/TSMCA.2007.914757
11. Cummings, M., Mastracchio, C., Thornburg, K., Mkrtchyan, A.: Boredom and distraction in multiple unmanned vehicle supervisory control. Interact. Comput. **25**(1), 34–47 (2013). https://doi.org/10.1093/iwc/iws011
12. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Hum. Factors **37**(1), 32–64 (1995). https://doi.org/10.1518/001872095779049543
13. Endsley, M.R.: Automation and situation awareness. Autom. Hum. Perform. Theory Appl. **20**, 163–181 (1996)
14. Fairley, R.E., Thayer, R.H.: The concept of operations: the bridge from operational requirements to technical specifications. Ann. Softw. Eng. **3**, 417–432 (1997)
15. Fuchs, C., Borst, C., de Croon, G.C.H.E., van Paassen, M.M.R., Mulder, M.: An ecological approach to the supervisory control of UAV swarms. Int. J. Micro Air Veh. **6**(4), 211–229 (2014). https://doi.org/10.1260/1756-8293.6.4.211
16. Goodrich, M.A., Cummings, M.L.: Human factors perspective on next generation unmanned aerial systems. In: Valavanis, K.P., Vachtsevanos, G.J. (eds.) Handbook of Unmanned Aerial Vehicles, pp. 2405–2423. Springer, Dordrecht (2015). https://doi.org/10.1007/978-90-481-9707-1_23

17. Hildmann, H., Kovacs, E.: Review: using unmanned aerial vehicles (UAVs) as mobile sensing platforms (MSPs) for disaster response, civil security and public safety. Drones **3**(3) (2019). https://doi.org/10.3390/drones3030059. https://www.mdpi.com/2504-446X/3/3/59
18. Hildmann, H., Kovacs, E., Saffre, F., Isakovic, A.F.: Nature-inspired drone swarming for real-time aerial data-collection under dynamic operational constraints. Drones **3**(3) (2019). https://doi.org/10.3390/drones3030071. https://www.mdpi.com/2504-446X/3/3/71
19. Hocraffer, A., Nam, C.S.: A meta-analysis of human-system interfaces in unmanned aerial vehicle (UAV) swarm management. Appl. Ergon. **58**, 66–80 (2017). https://doi.org/10.1016/j.apergo.2016.05.011. http://www.sciencedirect.com/science/article/pii/S0003687016300989
20. Karvonen, H., Heikkilä, E., Wahlström, M.: Safety challenges of AI in autonomous systems design – solutions from human factors perspective emphasizing AI awareness. In: Harris, D., Li, W.-C. (eds.) HCII 2020. LNCS (LNAI), vol. 12187, pp. 147–160. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49183-3_12
21. Kim, L.H., Drew, D.S., Domova, V., Follmer, S.: User-defined swarm robot control. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, pp. 1–13. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3313831.3376814
22. Koch, W., Mancuso, R., West, R., Bestavros, A.: Reinforcement learning for UAV attitude control. ACM Trans. Cyber-Phys. Syst. **3**(2) (2019). https://doi.org/10.1145/3301273
23. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004). https://doi.org/10.1518/hfes.46.1.50_30392, pMID: 15151155
24. Lotufo, M.A., Colangelo, L., Perez-Montenegro, C., Canuto, E., Novara, C.: UAV quadrotor attitude control: an ADRC-EMC combined approach. Control Eng. Pract. **84**, 13–22 (2019). https://doi.org/10.1016/j.conengprac.2018.11.002. https://www.sciencedirect.com/science/article/pii/S0967066118305148
25. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Making **2**(2), 140–160 (2008). https://doi.org/10.1518/155534308X284417
26. Rahwan, I., et al.: Machine behaviour. Nature **568**(7753), 477–486 (2019). https://doi.org/10.1038/s41586-019-1138-y
27. Sarter, N.B., Woods, D.D.: How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Hum. Factors **37**(1), 5–19 (1995). https://doi.org/10.1518/001872095779049516
28. Saffre, F., Hildmann, H., Karvonen, H., Lind, T.: Drones: new developments and environmental applications, chap. Self-swarming for Multi-Robot Systems deployed for Situational Awareness. Remote Sensing/Photogrammetry. Springer (2021)
29. Star, S.L.: This is not a boundary object: reflections on the origin of a concept. Sci. Technol. Hum. Values **35**(5), 601–617 (2010). https://doi.org/10.1177/0162243910377624

Effect of Red Blue 3D Videos on Visual Fatigue

Guilei Sun[1](✉), Yun Lin[1], Linghua Ran[2], and Yanhua Meng[1]

[1] School of Safety Engineering, China University of Labor Relations, Beijing 100048, China
sunguilei@culr.edu.cn

[2] China Standardization Institute, Beijing 100191, China

Abstract. In order to analyze the visual fatigue caused by Red Blue 3D videos, a combination of subjective questionnaire survey and objective experimental data was adopted. Questionnaire investigation for visual fatigue and eye movement data were collected in the process of viewing with human machine environment synchronous platform. Three segments of 0–15 min, 15–30 min and 30–45 min were intercepted, and the normality of physiological signal data was detected by SPSS 23.0. Kruskal-Wallis test and multiple tests were used to compare the differences between the groups. The results show that the fixation time of eyes on the screen decreases and the number of blinks increases gradually with the increase of watching time. And the pupil diameter gradually decreases while the right pupil diameter is larger than the left one. The subjects are more comfortable with blue lens than red one.

Keywords: Red and blue 3D display · Physiological signals · Multiple test · Pupil diameter · Left pupil

1 Introduction

With the improvement of living standards, 3D display has become a mature and common technology. When watching movies, the left and right eye images are completely separated. In order to maintain the fusion function, the brain is always in a state of high tension, which leads to the appearance of visual fatigue [1]. Therefore, 3D display brings people a real visual experience, followed by the problem of visual fatigue. Color difference 3D technology, combined with the use of passive red-blue (or red-green, red-cyan) filter color 3D glasses [2]. The use of 3D glasses ensures that the viewer's two eyes can see different images, which makes the viewer feel immersive. However, a problem of visual fatigue appears.

At present, some scholars have carried out research on fatigue on 3D display. Sun [3] researched respiratory signals for visual fatigue caused by 3D display. Wang et al. [4] selected the common double parallax 3D stereo image as the research object. Chen et al. [5] used EEG and VEP in the study of VDT visual fatigue measurement, in which the change of EEG reflected fatigue closer to mental fatigue, but the VEP did not change significantly. The eigenvalues of blink frequency and pupil regulation were also used. Li [6], Sakamoto [7], Kim [8] and some other scholars reflected the degree of visual fatigue caused by watching different VDT by blinking frequency, which showed that

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 427–437, 2021.
https://doi.org/10.1007/978-3-030-77932-0_33

the blinking frequency increased with the extension of time, while Wang [9] and Cai [10] used pulse signal in the study of VDT visual fatigue, which shows the pulse value decreased significantly with the deepening of visual fatigue.

In this research, the relationship of Red Blue 3D display and visual fatigue were studied and the questionnaire investigation, the fixation time, blinking times and pupil diameter were used to analyze characteristic of the subjects.

2 Experiments

2.1 Experiment Equipment

Tobii X2-30 screen-based eye tracker; Workstation (NVIDIA Jetson Xavier NX Developer Kitrefresh rate 120 Hz); Display (AOC, 23.8 in., 165 Hz/1 ms/2k); Red Blue 3D glasses (left red and right blue); Flash fusion frequency meter (BD-11–118, red light as the test light source, the background light intensity is 1, the light black ratio is 1:1, and the bright light intensity is 1).

2.2 Subject Selection

30 subjects were selected, including 15 males and 15 females, with an average age of 21.8 years. Before the experiment, the subjects were reminded to get enough sleep, and during the experiment, the subjects were voluntary and had no other emotional influence.

The informed consent was signed before the experiment and the experiment was completed according to the instructions.

2.3 Experimental Design

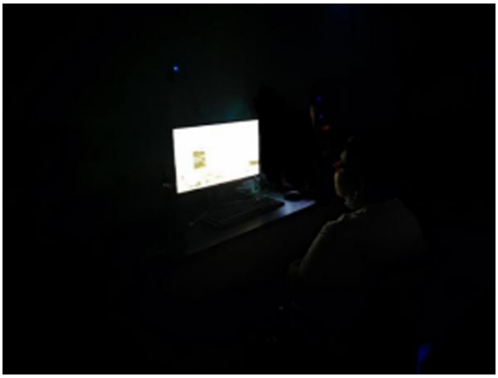


Fig. 1. Scene of the experiment

The subjects were given a questionnaire survey for visual fatigue before watching the video, and the flash fusion frequency meter was used to measure the critical fusion frequency (CFF) in order to obtaining the baseline data. Tobii x2–30 was used to collect the eye movement data during the watching process which lasting for 45 min. After watching the video, questionnaire survey and flash fusion frequency measurement were conducted again. The experiment was carried out in the dark which was shown in Fig. 1.

3 Data Processing

3.1 CFF Analysis

The mean value, median value and standard deviation of CFF before and after watching the video tested by flash fusion frequency meter are compared, as shown in Table 1.

Table 1. Descriptive statistics of CFF

Group		Mean	Median	Standard Deviation
Before video watching	Flash frequency	26.763	26.700	4.132
	Fusion frequency	28.280	28.400	3.493
After video watching	Flash frequency	26.517	27.600	3.769
	Fusion frequency	28.190	28.150	3.388

The bigger the CFF, the higher level of the brain consciousness. And when the human body is tired, the CFF is low [11]. It can be seen from the data in Table 1 that the three descriptive statistical data (mean, median and standard deviation) after watching are less than those before watching, which indicate that the visual fatigue after watching is significantly higher than before watching.

In order to analyze these four sets of data, Skapiro-Wilk method is used for normality test, and the test results are shown in Table 2.

Table 2. Normality test of CFF

Group		P-value
Before watching video	Flash frequency	0.826
	Fusion frequency	0.904
After watching video	Flash frequency	0.027
	Fusion frequency	0.481

It can be seen from Table 2, before watching the video, the flash frequency value and fusion frequency value are greater than 0.05, which conforms to the normal distribution. While p-value of fusion frequency is less 0.05 after watching the video and it does not conform to the normal distribution. The fusion value after watching the video is greater than 0.05, which also conforms to the normal distribution. At this condition, the four groups of flash fusion frequency data do not meet the normal distribution, then Kruskal-Wallis test was used.

Table 3. Kruskal Wallis test results of flash frequencies and fusion frequencies

Chi-square	5.1727
Degrees of freedom	3
Pr > Chi-square	0.1596

It can be seen from Table 3 that the chi square value is 5.1727 and $P = 0.1596 > 0.05$, showing that there is no significant difference in flash fusion before and after watching, that is to say, it is not effective to use CFF data to describe visual fatigue.

3.2 Questionnaire Analysis

Blehm [12] has studied that people's visual fatigue includes ten items, such as blurred vision, sore eyes, tingling eyes, drowsiness, dizziness, ghosting, tears, dry eyes, headache and vomiting. The reliability coefficient of the questionnaire is tested, and the result is 0.761, which is greater than 0.7, indicating that the subjective reliability quality of the questionnaire fatigue is very suitable for the experiment. Based on the reliability of the questionnaire, the fatigue of the subjects before and after watching the video was investigated. Each subjective problem was described in five different degrees. The scoring standard is shown in Table 4.

Table 4. Standard of point score

Score	Fatigue degree
0	Not at all
1	A little
2	Feel it
3	It's a little bit strong
4	Feel strong

The statistics for the scores of the questionnaire before and after watching the video were done and shown in Table 5.

It can be seen from Table 5 that the score after watching is significantly higher than that before watching. That is to say, the scores after watching are increased compared with that before watching in ten aspects, and it is more obvious in four aspects: blurred vision, sore eyes, drowsiness and ghosting. Moreover, drowsiness is the most obvious one. These two sets of data are analyzed, and the results is shown in Table 6.

Table 5. Scores for questionnaire before and after watching video

Content	Before watching video	After watching video
Blurred vision	33	46
Sore eyes	29	44
Tingling eyes	22	28
Drowsiness	30	72
Dizziness	24	34
Ghosting	31	40
Tears	22	25
Dry eyes	31	38
Headache	22	33
Vomiting	21	22

Table 6. Analysis of subjective questionnaire data before and after watching the video

Groups	Mean	95% confidence interval of the difference		Standard deviation	95% confidence interval of the difference	
		Lower	Upper		Lower	Upper
Before watching video	36.500	33.140	39.860	4.696	3.230	8.574
After watching video	49.200	38.666	59.734	14.726	10.129	26.883

Table 6 shows that the mean value and visual fatigue store are within the 95% confidence interval, and the mean value and standard error mean after the video watching are significantly greater than those before video watching. Then the two groups of data were tested for normality, the results showed that the p-value before and after viewing were greater than 0.05, in line with the normal distribution. Analysis of variance (ANOVA) was used on the data, and the results are shown in Table 7.

Table 7. ANOVA of data before and after watching video

Method	Numerator degrees of freedom	Denominator degrees of freedom	F-value	Pr > F
F-test	9	9	9.83	0.002

Table 7 shows Pr > F = 0.002 < 0.05. That is, the variances are homogeneity. T-test is carried out and the result is shown in Table 8. It can be seen that p-values are all less than 0.05 whether the variances are homogeneity or not. If the hypothesis is established, it means that the fatigue test scores by questionnaire before and after watching the video are statistically significant.

Table 8. T test results

Method	Variance	Degrees of freedom	T-value	P-value
Pooled	Equal	18	−2.60	0.018
Satterthwaite	Not equal	10.812	−2.60	0.025

3.3 Eye Movement Data Analysis

Due to the difference between the left and right lenses of Red Blue glasses, the eye movement test data were divided into three periods, 0–15 min, 15–30 min and 30–45 min, in order to analyze the fatigue and changes of the left and right eyes in different time periods.

Proportion of Fixation Time. Descriptive statistics of the total fixation time of the three periods are given, as shown in Table 9.

Table 9. Descriptive statistics of fixation time proportion

Time period	Mean	Standard deviation	Median	Variance	Range
0–15 min	0.482	0.188	0.511	0.036	0.741
15–30 min	0.423	0.188	0.511	0.029	0.741
30–45 min	0.302	0.147	0.270	0.022	0.526

As shown in Table 9, the average value of the data of three groups decrease gradually, indicating that the proportion of total fixation time in the total time shows a downward trend, and the total fixation time becomes less and less with the increase of viewing time. It also indicates the fatigue performance of the subjects are increasing. Among the data of three groups, the mean, standard error, median, variance and range are the smallest in the 30–45 min viewing time, which indicates that the dispersion degree of this group is the smallest. In other words, the total fixation time of 30 subjects in the period of 30–45 min viewing time has different decline, and tends to a certain value.

Skapiro Wilk test was performed on the data of the three groups. It showed that the p-values of fixation time data were all greater than 0.05, which conformed to the normal distribution. The data of three groups were analyzed by ANOVA, as shown in Table 10.

Table 10. ANOVA of the fixation time

Source	Degrees of freedom	Quadratic sum	Mean square	F-value	P-value
Fixation Time	2	0.457	0.229	7.86	0.0007

In Table 10, 0.05 is taken as the test level of hypothesis test, to the three groups of data, $p < 0.05$. It indicates statistical significance. $F = 7.86$, indicating that there are differences among the data of three groups. To sum up, the viewing and fixation time of the three time periods were different and statistically significant. It shows that with the increase of viewing time, the degree of visual fatigue increases, and the fixation time on the screen decreases gradually.

Eye Blinks. Through the analysis of the number of eye blinks recorded by the eye tracker when watching video, the number of blinks in three periods are carried out, as shown in Table 11.

Table 11. Descriptive statistics of blink times

Time period	Mean	Standard Deviation	Median	Variance	Range
0–15 min	291.50	165.00	276.00	15643	477.00
15–30 min	305.83	172.00	287.00	20407	574.00
30–45 min	316.70	181.00	265.00	14376	433.00

It can be seen from Table 11 that the mean value and inter quartile range increased with the increase of watching time. The data of the period 30–45 min is the maximum value in three periods. It indicates that the period of 30–45 min is the most discrete in the data of three periods, and the increase of blinking times is the most significant. That is to say, with the increase of viewing time, the subjects have to relieve the fatigue caused by watching video through the increase of blinking times.

The data of three time periods were tested for normality, and the results are shown in Table 12. In the period of 30–45 min, it does not follow the normal distribution. And then Kruskal Wallis method of multiple rank sum test was used. The results are shown in Table 13.

It can be seen from Table 13 that the difference in the total number of blinks among the three groups is statistically significant ($\text{Chi-square} = 0.7032$, $P = 0.0036 < 0.05$). From Fig. 2, it can be seen that 30–45 min is the most significant data. It indicates that the visual fatigue and the blink times increase with the increase of viewing time while the medium has a little decrease.

Pupil Diameter of Left and Right Eyes. To analyze the data of left and right pupils, the corresponding relationship between visual fatigue and pupil diameter, the descriptive statistics of the pupil diameters in three time periods are shown in Table 14.

Table 12. Skapiro Wilk test

Time period	Statistic		P-value	
0–15 min	W	0.958	Pr < W	0.270
15–30 min	W	0.952	Pr < W	0.189
30–45 min	W	0.915	Pr < W	0.020

Table 13. Kruskal-Wallis test

Chi-square	0.703
Degrees of freedom	2
Pr > Chi-square	0.004

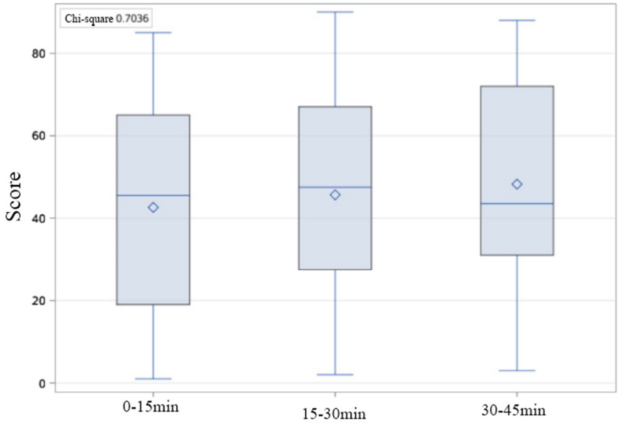


Fig. 2. Kruskal-Wallis test result

It can be seen from Table 14 that the mean value and the pupil diameter of both eyes decreases with the increase of watching time, which indicates that the pupil diameter gradually decreases with the increase of eye fatigue. The mean, standard deviation, variance and interquartile distance of pupil diameter in the right eye were greater than those in the left eye in the three periods of 0–15 min, 15–30 min and 30–45 min. It was assumed that the fatigue degree of the right eye was lower than that of the left eye. Then the normality of pupil data of left and right eyes in three periods is tested, as shown in Table 15.

It can be seen from Table 15 that the p-value of the left pupil is less than 0.05 at the period of 0–15 min, which does not conform to the normal distribution. The P values of the other five groups are greater than 0.05. Therefore, the average pupil diameter data of

Table 14. Pupil diameters of left and right eyes in three period

Descriptive statistics	Left eye			Right eye		
	0–15 min	15–30 min	30–45 min	0–15 min	15–30 min	30–45 min
Mean	4.636	4.317	4.032	4.650	4.347	4.058
Std	0.819	0.941	0.964	0.898	0.977	0.978
Standard deviation	0.671	0.885	0.928	0.807	0.954	0.957
Range	3.530	3.900	3.740	3.570	4.400	3.440
Inter quartile range	0.800	1.240	1.510	0.845	1.470	1.660

Table 15. Normality test for two pupils at different period

P-value	Time/min		
	0–15	15–30	30–45
Left pupil	0.038	0.563	0.467
Right pupil	0.210	0.218	0.420

the six groups do not meet the normal distribution. Kruskal Wallis test is used and the results are shown in Fig. 3 and Table 16.

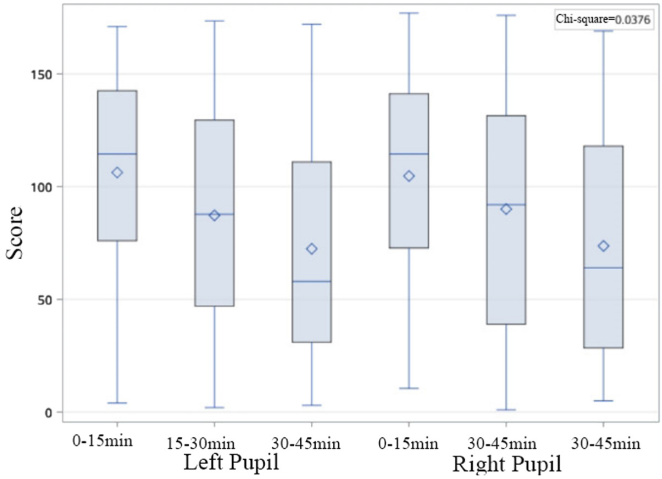


Fig. 3. Kruskal-Wallis test result

In Fig. 3, the scores of left and right eyes are all the highest in the period of 0–15 min, the data is the most significant and the scores of right eye are higher than that the left. Then the previous preliminary hypothesis is verified, the fatigue degree of the right eye is lower than the left, that is to say, under the same stimulation conditions, blue 3D lens in the Red Blue glasses is more comfortable than the red lens. And it are produce less visual fatigue.

Table 16 shows that the average pupil diameter of the six groups is 11.802 and $p = 0.0376 < 0.05$. It indicates that when the visual fatigue is produced, the pupil decreases, and the left and the right eyes are not synchronized. With the external stimulation, the degree of visual fatigue will be different, so the size of the pupil also has different performance.

Table 16. Kruskal-Wallis test

Chi-square	11.802
Degrees of freedom	5
Pr > Chi-square	0.0376

4 Conclusions

1) According to the questionnaire survey analysis, the subjects had obvious feeling of visual fatigue after watching video. Blurred vision, sore eyes, drowsiness and ghosting were more obvious, among which drowsiness was the most obvious one.
2) The visual fatigue and the blink times (mean and quartile) increase with the increase of watching time while the medium has a little decrease.
3) With the increase of viewing time, visual fatigue occurs, the number of blinks increases, the fixation time on the screen, the pupil diameter of the left and the right eyes decrease, and the pupil diameter of the left and right eyes is not synchronized with Red Blue 3D glasses.
4) Under the same conditions, the right eye using the blue lens in 3D glasses is more comfortable than the left eye. That is, the blue lens results in lower visual fatigue than the red one.

Acknowledgments. The presented work has been supported by General project of China University of Labor Relations (21XYJS012) and National key R & D projects (2017YFF0206604).

References

1. Wang, J.H., et al.: 3D Visual Fatigue of Personal Computer Display. Acta Sci. Natur. Univ. Sunyatseni **52**(05), 1–5 (2013)

2. Xu, P.: Analysis on the development and application of 3D TV technology. China Digit. Cable TV, (05), 584–586 (2015)
3. Sun, G.: Research on respiratory signals for visual fatigue caused by 3D display. In: Long, S., Dhillon, B.S. (eds.) Man-Machine-Environment System Engineering: Proceedings of the 20th International Conference on MMESE, pp. 663–672. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-6978-4_76
4. Wang, F.J.. Research on visual fatigue of 3D video. Beijing University of Posts and Telecommunications (2014)
5. Chen, C.M., Wang, H., Zhou, J., Du, Z.Q.: Review of VDT visual fatigue and its measurement methods. Chin. J. Ergon. **19**(02), 92–95 (2013)
6. Lee, E.C., Park, K.R., Whang, M., Min, K.: Measuring the degree of eyestrain caused by watching LCD and PDP devices. Int. J. Ind. Ergon. **39**(5), 798–806 (2009)
7. Sakamoto, K., Aoyama, S., Asahara, S., Yamashita, K., Okada, A.: Measurement of visual fatigue for large-sized TVs in a home viewing environment. In: The 13th IEEE International Symposium on Consumer Electronics, pp. 738–742 (2009)
8. Kim, D., Choi, S., Park, S., Sohn, K.: Stereoscopic visual fatigue measurement based on fusional response curve and eye blinks. In: Digital Signal Processing (DSP), pp. 1–6 (2011)
9. Wang, Y.T.: Research on VDT visual fatigue state recognition based on real-time pulse signal processing. Lanzhou University of Technology, Gan Su (2010)
10. Cai, H.Y.: Research on VDT visual fatigue based on pulse signal. Lanzhou University of Technology, Gan Su (2012)
11. Hsu, B.W., Wang, M.J.J.: Visual fatigue evaluation of PDP, LCD and CRT TV. In: Proceedings of 17th World Congress on Ergonomics (2009)
12. Blehm, C., et al.: Computer vision syndrome: a review. Surv. Ophthalmol. **50**(3), 253–262 (2005)

Human-Centered HMI for Crane Teleoperation: Intuitive Concepts Based on Mental Models, Compatibility and Mental Workload

Felix Top[1(✉)], Sebastian Pütz[2], and Johannes Fottner[1]

[1] Technical University of Munich, Munich, Germany
felix.top@tum.de
[2] RWTH Aachen University, Aachen, Germany

Abstract. Current crane controls result from a machine-centered design process and rely heavily on the operator, as he/she has to perform continuous mental transformations to assess the correct control inputs for all joints, causing considerable mental workload. To facilitate the development of intuitive HMIs, a design framework for human-centered remote crane controls is presented. The framework allows comparison of the traditional machine-centered HMI with other, new HMIs in respect of important design principles. It focusses on supporting the operator in achieving his/her primary goal: moving the load. The framework is used to analyze two new human-centered HMIs for a loader crane, the direction-oriented and the target-oriented HMI. Based on the compatible task/action mappings of the direction-oriented HMI, it is predicted that the HMI facilitates interactions even for operators with minimal prior experience. A mental workload reduction is also expected for the target-oriented HMI as it combines task/action mappings with an increase in automation.

Keywords: Human-machine interaction · Human-centered design · Intuitive control · Crane control · Teleoperation · Mental workload

1 Introduction

Cranes are a means of conveyance that allow a load to be lifted, lowered or moved horizontally by attaching it to a load supporting means such as a chain or a rope [1]. To fulfill the needs of multiple applications, they are available in a range of designs such as overhead cranes, wall-mounted travelling cranes or rotating tower cranes (cf. [2]). The standard human-machine interface (HMI) for crane operations asks the user to separately control the crane's different degrees of freedom (DOFs), which are defined by the crane's slewing gears and linear motions [3]. This general design concept is independent of the chosen input device. For example, when using push buttons, two buttons are used to trigger

D. Harris and W.-C. Li (Eds.): HCII 2021, LNAI 12767, pp. 438–456, 2021.
https://doi.org/10.1007/978-3-030-77932-0_34

each of the two possible movement directions per degree of freedom (DOF). Using linear levers, both movement directions of one DOF are mapped to one input element, and when using joysticks, multiple DOFs are controlled with the same input element. Most cranes are controlled using one of these three input elements [3]. The separate control of the individual DOFs causes two main hindrances for the user's interaction with the crane. First, there is no intuitive mapping between the DOFs and the input elements. Therefore, input elements have to be labelled and new users are required to explicitly learn the mapping. Second, a load movement has to be realized by the superposition of movements along the crane's DOFs. This requires the operator to perform a mental transformation of the intended load movement to select the DOFs involved as well as the respective movement directions.

Hence, the standard HMI for crane operations is based on a machine-centered design perspective. The way the user interacts with the system is defined by the functional principle of the machine since the user controls the movements of the separate crane segments along their corresponding DOFs. This can be considered a strong impediment to effective crane control and is especially demanding for less experienced operators because the operator is continually confronted by several simultaneous tasks. Taking into consideration that both the productivity and safety of crane operations depend on the effectiveness and efficiency of the operator-crane interaction (cf. [4]), the importance of optimizing crane human-machine interfaces (HMIs) becomes apparent.

2 State of the Art: Improving the Crane HMI

Having identified this potential for improvement, several solutions have been developed. The most promising approach shifts the responsibility of transforming the desired load movement into the required crane segment movements from the human operator to the crane system. As a consequence, the operator is able to control the crane's boom tip, and thereby the load, directly without needing to consider the crane's DOFs. This is called boom tip control. The approach has been shown to improve operator performance with regard to learning times [5] as well as cycle times and error rates [6].

Furthermore, research projects have focused on alternative input devices for controlling the load. For example, Peng [7,8] developed several devices that allow the operator to specify the desired path of the load. By tracking various objects that can be carried by the operator, the load follows the operator's movements. Kazerooni et al. [9] chose a glove as an input device and used the manual force applied to the load as a directional input. Both concepts were successful in improving operation efficiency. However, the solutions referred to are not without drawbacks. They all require either expensive technical equipment and infrastructure or impose restrictions on the position of the operator.

These restrictions are in contrast to the current trend of eliminating restrictions on the operator's position by use of teleoperation [10]. Teleoperation improves the user's flexibility and comfort and is also associated with substantial safety benefits (cf. [3]). First, the user can position himself/herself in a way that ensures an unobstructed view of the load, and second, he/she can avoid being too close to hazardous locations such as excavation pits. Thus, the cited approaches do not provide an optimal solution as they fail to combine the benefits of an intuitive interaction concept with those of teleoperation. Due to the numerous advantages of teleoperation over a fixed operating station, failure to use remote controls is neither a targeted nor an optimum solution.

When using a remote input device (i.e. radio control) in combination with boom tip control, a further problem in respect of the operator's interaction with the crane occurs. As long as the operator is positioned in a fixed cab at the crane, the user's and the crane's reference systems are aligned. Therefore, the user's inputs and the crane's movements are always compatible. In contrast, if the user changes his/her position, as he/she can when using radio control, the user's and the crane's reference systems dissociate. Thus, the user is required to perform a mental rotation of his/her own reference system to realign it with the crane's reference system increasing the cognitive demand of the task [11].

To overcome these compatibility problems, Kivila et al. investigate rotating the control coordinate system with the user [12]. This allows the operator to control the hook directly from his/her perspective. Research results are promising as they demonstrate that direct hook control from the operator's own perspective can lead to reduced time spent on the task as well as fewer errors for novice operators. Additionally, operator satisfaction can be increased. However, the solutions presented in [12] have only been tested in indoor environments using gantry cranes and also require costly spatial infrastructure. These drawbacks have to be avoided when developing control concepts for cranes that are for example mobile or have large operating spaces. Nevertheless, the results confirm that the effectiveness of the human-machine interaction is strongly influenced by compatibility and by taking the user's orientation into account.

3 Research Gap and Objective

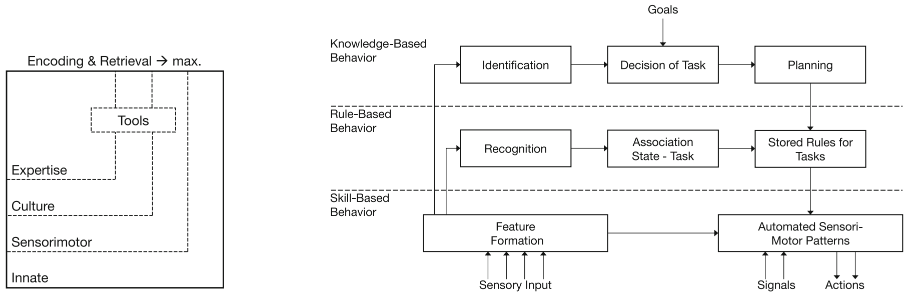
Examining the current state of the art of remote crane control and the corresponding HMIs, it appears that until now, all developments have been based solely on using a different HMI design idea and/or the first-time usage of sensor technologies in this context. Although some solutions show promising improvements with regard to operator performance and operator satisfaction, there exists no systematic consideration of the reasons why certain HMIs can be expected to be superior to others from the perspective of the human operator. Thus, new concepts are the result of different (arbitrary) procedures because they do not rely on an established methodological framework for crane HMI design. Such a framework would support the development of human-centered crane HMIs that on the one hand are suitable for controlling the machine but on the other hand, and most importantly, promote simple and efficient interactions and, thus, reduce the mental demand for the operator. This human-centered design approach (i.e. creating a system to accomodate the human's needs, capabilities and behavior [13]) could systematically improve the productivity and safety of crane operations.

The goal of this paper is to analyze which psychological constructs can be assumed to differentiate the effects of crane HMIs on the operator's mental workload, in order to establish a general framework that can guide the development of improved crane HMIs. We focus on remote-control HMIs because teleoperation is the preferred operating mode in the majority of cases. In the following section, we first present the conceptual foundation for human-centered design with regard to intuitiveness, mental models, compatibility and the effect of automation on mental workload. Then we introduce a design framework that is suitable for comparing crane HMIs in respect of the interaction quality between the operator and the machine. Finally, we present two innovative crane HMI concepts which, based on our framework, are expected to significantly improve human-machine interaction during crane control.

4 Conceptual Foundation: Human-Centered Design in the Light of Intuitiveness, Mental Models, Compatibility and Mental Workload

4.1 Intuitiveness

According to common definitions, intuitiveness can be considered as the result of the unconscious application of prior knowledge during a human's interaction with a certain system [14–16]. Applying pre-existing knowledge allows the user to focus on his task rather than investing mental effort into understanding the system's interface. Consequently, facilitating intuitive interactions is understood as a primary way of reducing the user's mental workload [17].



(a) Continuum of knowledge, adapted from [17, p. 131].

(b) Simplified illustration of the Skills, Rules, and Knowledge model by Rasmussen, adapted from [18, p. 258].

Fig. 1. Conceptual foundation: intuitiveness.

Based on this definition of intuitiveness, Naumann et al. [17] have developed the 'continuum of knowledge' as shown in Fig. 1a. This is a model for differentiating classes of pre-existing knowledge that enable intuitive interactions. The continuum begins with innate knowledge that is based on gene activation in the prenatal stage. The following two levels sensorimotor and culture distinguish between knowledge that is acquired universally in early childhood or that depends on the culture a human grew up in. The most complex level of knowledge is termed expertise and is obtained during education and professional training. By moving down through the continuum, the group of people that possesses the knowledge and is, therefore, able to experience the intuitive interaction becomes larger and increasingly heterogeneous [15]. Moreover, the knowledge in the lower levels is not only more prevalent among users but also more consolidated because it is encoded and retrieved more frequently. As a consequence, users rely on lower levels of knowledge in situations of high mental demand. It is important to emphasize that any interaction will only be intuitive if the pre-existing knowledge can be applied subconsciously (cf. [14,16,17]). Thus, if the user's interaction with a system depends on an initial impression that is formed via unconscious associative processes, the interaction can be labeled as intuitive.

This view on intuitive interaction makes it possible to establish a connection to one of the most prominent theories for explaining human action control: the Skills, Rules, and Knowledge (SRK) model of Rasmussen [18]. This model differentiates human action control into knowledge-based, rule-based and skill-based behavior (s. Fig. 1b). Knowledge-based behavior is required in situations that are unknown. The human has to identify the relevant parameters of the situation and use this information to define a task goal. He/she then has to develop a task rule that allows him/her to accomplish the set goal. If a situation has been experienced before, the human has to recognize the recurrence and carry out the already formed task rule. This is referred to as rule-based behavior. Finally, in the case of skill-based behavior, the action pattern for the situation is already internalized, so it is automatically activated when the human is confronted with the respective conditions. Skill-based behavior usually requires substantial experience in the specific task.

Transferring the discussed insights into the model of intuitive interactions, the unconscious application of pre-existing knowledge from another situation or system state makes it possible to skip the phase of knowledge-based behavior. Instead, rules from other situations are applied or even automatically activated resulting in rule- or skill-based interactions. This relationship between the rule- and skill-based levels of the SRK model and intuitive interactions has been stated by multiple researchers [17,19]. It highlights the advantage of intuitive interactions in providing a head start in learning how to interact with a system.

4.2 Mental Models

A mental model constitutes a structural model that determines how and if environmental cues are processed and information is integrated into the human's understanding of his/her surroundings [20]. The formation of a mental model allows the user to learn how to interact with a new system by enabling him to understand and predict the system's responses [21,22]. An insufficient mental model of the system causes the operator to perform incorrect or unnecessary actions resulting in rising mental workload due to the additional effort and unexpected system responses [23]. Moreover, existing mental models circumvent the necessity to learn novel systems from the beginning because they provide a structure for integrating new information into pre-existing knowledge structures. This enables a more efficient learning process that has been shown to result in faster encoding, better retention, and more accurate and quicker operating performance [24]. Understanding the user's mental model is an optimal approach to understanding why the user interacts with a system in a certain way and for achieving ways in which the performance of the human-machine system can be improved [25].

However, mental models cannot be considered as a homogenous group of internal representations that structure the user's knowledge. Researchers have tried to differentiate different types of mental models that satisfy different requirements. The most common use of the term is to describe "a rich and elaborate structure, reflecting the user's understanding of what the system contains, how it works, and why it works that way" [21, p. 12]. The most important feature of such a model is that it allows the mental simulation of the system's behavior [22]. In contrast to these elaborate internal representations of the target system, there exist more simplified models which Young defined as task/action mappings [26]. Unlike classical mental models, task/action mappings do not contain extensive knowledge about how a system works. Instead, they store the concrete actions that are required to accomplish a certain task. Whether a task/action mapping or an elaborate mental model is more beneficial for the user depends on the application. For example, an elaborate mental model is superior when attempting to explain why a system behaves in a certain way under unusual circumstances as the task/action mapping lacks an explanation of the system's functioning principles [26]. In those situations of knowledge-based behavior, the elaborate mental model is necessary for generating new insights, for example by developing new mappings that fit the current situation. However, since the elaborate mental model exclusively focuses on the system's overall functioning principles, it lacks direct applicability for specific tasks.

Consequently, task/action mappings guide the user's actions in performance-oriented tasks using rule- or skill-based behavior [26,27]. According to Young [26], the elaborate mental model is practically irrelevant under those conditions. Taking everything into account, the answer to the question of which type of mental model should be facilitated via the system design, depends on the type of tasks the user is regularly confronted with. Enabling the user to develop an in-depth understanding of the system enables him/her to deal with novel use cases more easily, whereas consistent task/action mappings are highly beneficial for operating performance in everyday use [26].

4.3 Compatibility

Compatibility is defined as the magnitude of decoding effort in human information processing whereby low compatibility corresponds to high decoding effort [28]. In the 1950s, Fitts and Seeger showed that the spatial correspondence between a set of visual stimuli and a set of manual response elements affects both the speed and accuracy of participants' reactions: The higher the spatial correspondence between the set of stimuli and the response elements, the better the participants' performances [29]. The observed effects are referred to as spatial stimulus-response (SR) compatibility. In a second experiment, Fitts and Deininger only used one corresponding arrangement but manipulated the pairings between the stimuli and the response elements [30]. They concluded that the requirements for SR compatibility are (1) congruent stimulus and response sets and (2) congruent pairings between the individual stimulus and response elements. Today, these two conditions are usually termed set-level compatibility and element-level compatibility, as proposed by Kornblum and colleagues [31]. They also developed a general model for describing SR compatibility [31]. According to them, set-level compatibility is based on the fact that stimulus and response elements vary in a common dimension which is referred to as dimensional overlap. Element-level compatibility is present if stimulus and response elements are matched based on their level in the shared dimension.

Human information processing can be divided into minimum three stages: stimulus identification, response selection and motor execution (cf. [32]). The effect of SR compatibility is attributed to the stage of response selection. Popular models for explaining the observed effects are dual process models that differentiate two routes of response selection (see Fig. 2). In this model, the first route is based on the conscious identification of the presented stimulus and the subsequent intentional selection of the required response alternative (lower branch in Fig. 2). The second route describes the automated activation of a response when perceiving the stimulus, whether it is correct in the current situation or not (upper branch in Fig. 2).

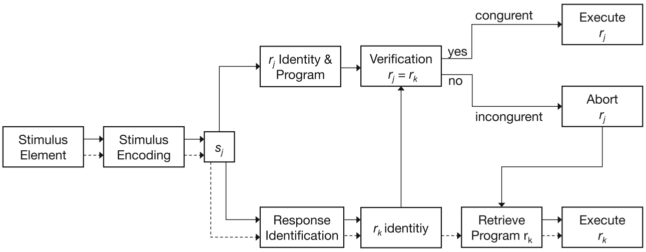


Fig. 2. Illustration of the dual process model for stimulus-response compatibility tasks with (solid lines) and without (dashed lines) dimensional overlap between stimulus and response sets. The top branch shows the automatic and the lower branch the intentional processing route. Adapted from [31, p. 257].

If no dimensional overlap exists between the stimulus and response sets, the human only uses the intentional processing route which results in comparatively slow responses. In contrast, the presence of set-level compatibility also enables the automated processing route, and the response that is compatible with the presented stimulus is activated [33]. In the case of element-level compatibility, the two processing routes yield congruent responses resulting in faster reactions.

Another compatibility effect is referred to as response-effect (RE) compatibility. In an experimental RE paradigm, the participant's response is followed by an effect. The response elements and effect elements are either associated during a learning phase or the response and effect sets have a dimensional overlap [34]. Therefore, response and effect pairings can be either compatible or incompatible. Multiple studies have shown that compatible RE mappings lead to faster reaction times (e.g. [35,36]). Considering that effect follows response, this shows that the anticipation of the effect plays a critical role in the response initiation.

Furthermore, compatibility is also influenced by the human's visual perspective. In most cases, control elements and the view of the controlled system are aligned. However, if the operator uses control elements that are in front of him but has to turn his/her head to one side to see the controlled system, the control can only be either compatible with the system's movements in a global reference system or compatible in the reference system that is defined by the operators visual field. Worringham and Beringer referred to these conditions as control-display compatibility and visual-field compatibility [37]. Their experiments showed that operating performance is strongly affected by visual-field compatibility but not by control-display compatibility.

4.4 Automation and Mental Workload

Considering human information processing when designing a system in line with the user's needs and abilities (i.e. increasing intuitiveness) is one option for decreasing the demands imposed on the operator. An alternative approach is to eliminate certain information processing demands in the first place. This is the approach of automation. Different automation solutions can be classified by considering what fraction of the human's responsibilities are handed over to the system [38]. Where the human operator is concerned, automation is intended to reduce mental workload by freeing up some of the limited cognitive resources [39]. By reducing the task demands, automation reduces the effect of the skill gap between experts and novices, relieving the novice from deliberative processing similar to an expert [40]. However, the effect of automation on mental workload is not solely positive. When automated systems are designed without considering the user, those functions that are well understood are automated while less understood and therefore potentially more difficult tasks are left to the human operator [41]. Common residual tasks for the human are monitoring the system and taking over in the event of system errors which are tasks that humans are not particularly effective at performing. If the user is not involved in controlling the system, both his/her skill and situation awareness will decline. In this case, the user may experience mental underload or mental overload due to unexpected increases in processing demands [42]. To prevent this, human-centered automation should always keep the human operator in the loop [43].

5 Framework for the Design of Human-Centered HMIs for Teleoperated Cranes

Having established the necessary definitions and the conceptual foundation, we can draw conclusions regarding human-centered HMI design for crane teleoperation. To increase intuitiveness compared to the present machine-centered HMI, a human-centered HMI should support the formation of appropriate mental models, assure compatibility and use potentials of system automation, without taking the operator out of the control loop.

5.1 Support Appropriate Mental Model: Task/Action Mappings

The user's mental model is essentially formed by the system and the task that has to be completed. The main task when operating a crane is moving a load from a starting point to its destination. Hence, any human-centered HMI should provide a direct match with the human's goal of moving the load. This is not the case for the machine-centered HMI as the exact movements of the crane segments themselves are not part of the task specification. One could argue, that in the case of the machine-centered HMI, the main input for the formation of the user's mental model is not the load movement but the crane's DOFs instead.

In the vast majority of cases, the operator is not (and should not be) required to consider and control the movements of the separate crane segments in order to reach his/her primary goal (i.e. moving the load). Therefore, there is no need for an accurate mental representation of the system's DOFs. Furthermore, problem solving and consequently knowledge-based behavior do not play a crucial role in crane control. For those two reasons, we conclude that an elaborate mental model is not required for the vast majority of crane operators and applications. In contrast, since cranes are used for one particular task only that is repetitively performed with low variation, it is reasonable to choose an HMI concept that facilitates simple task/action mappings to benefit rule- and skill-based behavior.

Therefore, a human-centered crane HMI should provide direct task/action mappings by design, instead of contributing to more elaborate mental models. Consequently, a human-centered HMI should allow the operator to directly control the movements of the crane's boom tip and thereby the load.

5.2 Assuring Compatibility

When using the machine-centered HMI, the mapping between the user inputs and the load's movements does not correspond to any action-effect mapping that is learned outside the application of cranes. Thus, the knowledge required to establish task/action mappings for this HMI can be classified as expertise knowledge for tools in the continuum of knowledge. In contrast, the task/action mappings for a human-centered HMI should follow the principles of interface compatibility which are located at the sensorimotor level. This means that a certain control input in one direction results in a load movement in a fixed dimension in order to enable set-level compatibility with the load's movement. Set-level compatibility is not present when applying the machine-centered HMI. Not only do none of the common input device options such as push buttons, linear levers or joysticks provide spatial compatibility with the movement axes of the crane segments, the fact that the orientation of these axes changes depending on the orientation of the other crane segments makes set-level compatibility impossible for fixed crane controls. Only the presence of set-level compatibility creates the potential of element-level compatibility and the associated processing speed benefits due to correspondence between the automatic and the intentional processing route. Element-level compatibility is present, if, for example, a control input always results in a load movement parallel to the input direction, regardless of the operator's orientation.

There are some additional factors that should be considered when analyzing the compatibility of an interface. First, RE compatibility only holds when the load movement is in fact the intended effect of the operator's action [34]. Second, compatibility effects depend on visual-field compatibility rather than control-display compatibility. Thus, only when the operator's visual field is aligned with the input device's reference system, will he/she benefit from the full potential of compatibility effects.

To summarize, a human-centered HMI should allow the operator to form compatible task/action mappings based on compatible user inputs and system outputs. As simple spatial compatibility relations are learned independent of cultural background, the required knowledge is located at the sensorimotor level in the continuum of knowledge. What is more, this knowledge is applied unconsciously as highlighted by the dual process model of response selection that explains compatibility effects by via parallel automatic processing route. Therefore, the requirements for facilitating intuitive interactions for a large and heterogeneous group of users are met independent of their prior experience with controlling cranes. The application of compatible task/action mappings that have already been internalized can enable the operator to interact with the crane on a rule-based level even when using it for the first time.

Moreover, these compatible task rules can be expected to be particularly stable because they are constantly encoded and retrieved independent of crane usage. Consequently, they are likely to improve operator performance in rule- and skill-based behavior. In contrast, the task/action mappings for the machine-centered HMI have to be consciously learned by any operator without substantial experience with the respective crane system locating them at the expertise level on the continuum of knowledge. Importantly, even for experienced users, compatible task/action mappings can be expected to be more stable which would benefit operating performance in demanding situations. Taking everything into account, since a human-centered HMI based on compatible task-action mappings should facilitate intuitive interactions at a more basic level of prior knowledge thus enabling the operator to focus on accomplishing the task rather than figuring out the interface, its application is expected to improve human-machine interaction significantly compared to the machine-centered HMI.

5.3 Using Effects of Automation

The interaction between the operator and the crane can also be improved by equipping the crane in such a way that information processing demands on the operator's side are eliminated. If the user specifies his/her task goal (i.e. the destination of the load), the crane's control unit could take over all necessary operations in order to move the load to the desired point. In this case, the operator would not need to specify any crane movements at all. Thus, it would be a more automated interaction concept that allocates more responsibility to the machine since the operator is not required to constantly specify the load's movement. It might be expected that the such an HMI concept would provide an additional reduction in the operator's mental workload, however, the literature review also shows that the effect of automation on mental workload is more complex. When the automation shifts the responsibilities of the operator from active controlling to passive monitoring and failure detection, operating performance has been shown to degenerate. This can be caused by either mental underload or mental overload in cases of unexpected system behavior. However, when applied with care, automation can be another tool for supporting the operator in achieving his/her primary goal.

5.4 Resulting Framework

In summary, we propose three main tools for implementing intuitive crane HMIs with the goal of enhanced intuitiveness:

1. Support an appropriate mental model, i.e. task/action mapping.
2. Ensure compatibility, i.e. set-level & element-level (considering stimulus-response, response-effect and visual-field) compatibility to support the formation of compatible task/action mappings.
3. Increase the level of automation.

These tools form the methodological toolbox for developing human-centered crane HMIs for teleoperation. Figure 3 shows a summary of our methodological framework and also includes the expected results if the method is applied properly. Mainly, we expect increased intuitiveness, which in turn should result in a reduction of the operator's mental workload. Furthermore, as increased intuitiveness is an indicator for less demanding human-machine interaction, we expect other performance indicators to improve, as well. This includes efficiency, effectiveness and satisfaction, all three of which are subsumed under usability.

To conclude this section, we have established the necessary framework of human-centered design and interaction. It can form the methodological basis for developing innovative crane HMIs as it allows the evaluation of different crane HMIs in respect of their contribution to an intuitive human-machine interaction prior to their implementation. Consequently, we consider the presented framework an important guideline for future developments of crane HMIs.

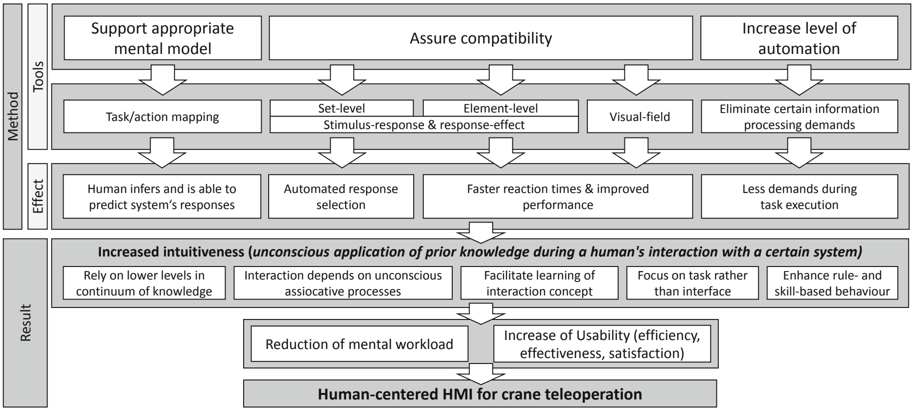
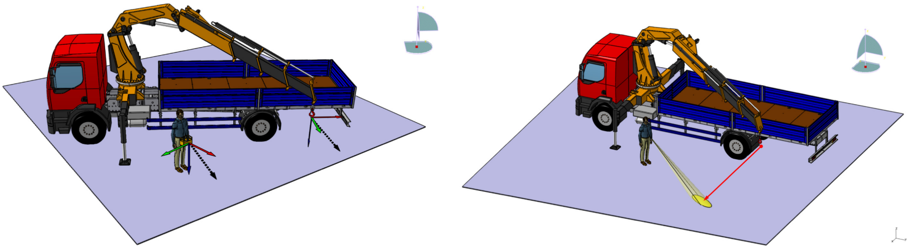

Method
Tools
Effect
Result
Support appropriate mental model
Assure compatibility
Increase level of automation
Task/action mapping
Set-level
Element-level
Stimulus-response & response-effect
Visual-field
Eliminate certain information processing demands
Human infers and is able to predict system's responses
Automated response selection
Faster reaction times & improved performance
Less demands during task execution
Increased intuitiveness (unconscious application of prior knowledge during a human's interaction with a certain system)
Rely on lower levels in continuum of knowledge
Interaction depends on unconscious assiocative processes
Facilitate learning of interaction concept
Focus on task rather than interface
Enhance rule- and skill-based behaviour
Reduction of mental workload
Increase of Usability (efficiency, effectiveness, satisfaction)
Human-centered HMI for crane teleoperation

Fig. 3. Methodological framework for the development of human-centered crane HMIs for teleoperation: tools and known effects (*top*) and expected results (*bottom*).

6 Proposed Human-Centered HMI for Teleoperated Cranes

In this section, we propose two newly developed human-centered HMIs and compare them to the traditional machine-centered HMI using the established framework.


(a) Direction-oriented HMI.

(b) Target-oriented HMI.

Fig. 4. Proposed concepts for remote control of teleoperated cranes.

6.1 Direction-Orientated HMI

Principle. Based on the principle of combining boom tip control and improved HMI design, a solution that combines an HMI that enables the operator to directly control the load's movement from his/her perspective without considering the crane's DOFs, with the prerequisites for teleoperation (due to its numerous advantages) and also the absence of costly infrastructure has been developed [44]. As part of the project, inverse kinematics have been specified that enable the crane to transform the user's input into the required joint velocities for the crane's actuators [45]. Hence, the movements of the individual crane segments that are needed to accomplish the specified load movement are calculated by the crane system and do not have to be considered by the operator. Additionally, the user's orientation (i.e. the relative rotation to the crane coordiante system) is captured by using the inertial sensor technology of a microelectromechanical system (MEMS) in the input device. This enables the crane's control unit to adjust the load movements according to the operator's orientation. Using a joystick controller, one joystick controls the load's movement in the horizontal plane, and a second joystick with one input axis controls the load's vertical movement. As shown in Fig. 4a the load's movement (dashed black line at the hook) is, therefore, always parallel to the tilting direction of the joystick (dashed black line at the user's position).

Consequently, the user can control the load based on his/her own reference system without considering crane related factors. We expect this system to simplify crane control considerably, especially when the user's coordinate system and the crane's coordinate system are not aligned because of the user's orientation (as can be seen in Fig. 4a). As the user specifies the desired direction of the load's movement from his/her own perspective via the input device, this concept is termed a direction-oriented HMI.

Discussion with Regard to Framework. Given that the task of the operator-crane system is to position the load, the concept of the direction-oriented HMI provides a direct match with the human's goal. Furthermore, the direction-oriented HMI allows the operator to form task/action mappings based on compatible user inputs and system outputs.

As the joystick controller directly controls the load's movement, this HMI provides set-level compatibility with the movement of the load. Element-level compatibility for the vertical control is easily achieved by mapping load lifting to pushing the joystick forward and load lowering to pulling the joystick backward, using existing population stereotypes [46]. The compatibility of horizontal load movements depends on the position of the operator relative to the load, also ensuring element-level compatibility. To be more precise, element-level compatibility is present in the sense of both SR and RE compatibility. When the target location as the task defining stimulus is on the left side of the load, the operator is required to respond by pushing the joystick to the left making the interface SR compatible. Similarly, pushing the joystick to the left causes the load as the distal effector to move to the left making the interface RE compatible.

RE compatibility, as elaborated above, only holds if the load movement is in fact the intended effect of the operator's action. However, since the operator's task is to position the load, it is very reasonable to assume that RE compatibility is achieved. Concerning visual-field compatibility, processing benefits can be restricted as the load movement is adjusted based on the orientation of the input device and not of the operator's head. That said, since in most cases the operator is able to turn himself and the controller towards the load aligning his/her visual field with the input device's reference system, visual-field compatibility should be achieved in most situations.

In summary, the direction-oriented HMI is in line with the elaborated framework. Hence, its application is expected to cause a reduction in the operator's mental workload in comparison to the traditional machine-centered HMI.

6.2 Target-Oriented HMI

Principle. In a further attempt to facilitate intuitive interactions between the operator and the crane, a second HMI is implemented that allows the user to specify the desired target location of the load with a pointing device, such as a laser pointer. We term this concept a target-oriented HMI. Like the direction-oriented HMI, the target-oriented HMI enables the operator to control the load directly without considering the movement of the separate crane segments. The implementation of vertical control is equivalent to the direction-oriented HMI with one joystick, with its one input axis controlling load lifting and load lowering. The difference between the two concepts is the horizontal control of the load. Using the target-oriented HMI, the operator specifies the desired target location of the load by pointing with a laser pointer (see Fig. 4b) at the respective location on the ground and saving the input with a press on the confirmation button. While the forward feed lever is tilted, the crane automatically moves the load horizontally towards the specified location until the load is positioned exactly above it. Accordingly, the operator is not required to constantly specify the direction of the load's movement, but he/she can rely on the crane system to move the load directly towards the target location. While doing so, he/she has to monitor the horizontal load movement and control the load's vertical position.

Like the direction-oriented HMI, this solution is compatible with teleoperation and does not ask the user to consider the crane's DOFs. However, it is a more automated interaction concept and allocates more responsibility to the machine.

Discussion with Regard to Framework. The target-oriented HMI mirrors the direction-oriented HMI in that the core principle is controlling the load and not controlling the crane. Thus, it has the same task-centered and thereby human-centered system image. In addition, the task/action mapping for vertical control is equivalent and based on a compatible mapping between joystick and load movement. Consequently, set-level & element-level compatibility are present. In contrast, the horizontal control is not based on a strictly compatible mapping between joystick and load. It is divided into the two phases of target specification (TS phase) and actual load movement (LM phase). The interface in the form of a laser pointer in the target specification phase is inherently compatible since the laser point directly follows the movements of the laser pointer. Using a pointing device such as a laser pointer can, therefore, be classified as sensorimotor knowledge on the continuum of knowledge.

However, the load movement phase will account for the majority of time spent on a task in most cases. As the operator does not explicitly specify the horizontal load movement in this phase, there is also no compatibility between user input and system output. This emphasizes that the human-machine interaction in this phase is mainly determined by the degree of system automation and not by interface compatibility. Applying the terminology of the levels of automation for real-time control systems [38], the machine-centered HMI is a standard example of manual control. The direction-oriented HMI can be classified as action support since the crane system helps the operator to implement the load movements by controlling the separate crane segments. Consequently, the vertical control of the target-oriented HMI is also located at the action support level. In contrast, the horizontal control of the target-oriented HMI is an example of batch processing. The user's task is to generate and select the desired location, but the machine is solely responsible for implementing the movement to reach it. Thus, the target-oriented HMI has a higher level of automation compared to the direction-oriented HMI. This highlights that the horizontal control of the target-oriented HMI does not simplify the human-machine interaction by facilitating intuitive interactions but by transferring additional processing demands from the operator to the crane system.

As one of the main goals of increasing system automation is to reduce the user's mental workload, it might be expected that the target-oriented HMI provides an additional reduction of the operator's mental workload compared to the direction-oriented HMI. However, when the automation shifts the responsibilities of the operator from active controlling to passive monitoring and failure detection, operating performance has been shown to degenerate. Transferring these insights to the application of the target-oriented HMI, it is difficult to anticipate whether operators will feel relieved due to the additional automation and reduced processing demands or will feel under more load because they are out-of-the-loop. Since the direction-oriented and the target-oriented HMI influence the operator's mental workload in different ways, more insight is required to reliably answer the question of which of the two provides a greater reduction in the operator's mental workload. When contrasting the target-oriented HMI to the standard machine-centered HMI, the comparison becomes less ambiguous. Based on its human-centered system image of controlling the load as well as the compatible vertical control and the increase in system automation for horizontal control, in line with our framework, the target-oriented HMI is expected to cause a reduction in the operator's mental workload compared to the machine-centered HMI.

6.3 Comparison

Table 1 shows a comparison between the two human-centered HMIs and the machine-centered HMI with regard to the established framework. It illustrates the superiority of the human-centered HMIs over the traditional HMI in all framework categories.

In a video simulation based online survey with 149 participants, results demonstrated improved operation efficiency, and reduced error frequency and learning time as well as increased input certainty and satisfaction in comparison

Table 1. Comparison of direction-oriented and target-oriented HMI with traditional machine-centered HMI

HMI	App. mental model	Assure compatibility (SR & RE)			Automation	Result (exp.)
	Task/action mapping	*Set-level*	*Element-level*	*Visual-field*	*increased*	*Mental Workl.*
Direction-oriented	✓	✓	✓	✓	⇑ (see [a])	⇓
Target-oriented*	✓	✓ (TS, z)	✓ (TS, z)	✓ (TS, z)	⇑ ⇑ (see [b])	⇓
Machine-centered	×	×	×	×	× (see [c])	*reference*

* TS: target specification phase; LM: load movement phase; x/y/z: movement in corresp. direction during LM
[a] action support [b] batch processing (x, y) & action support (z) [c] manual control (*reference*)

to the traditional machine-centered HMI [44]. These results were further supported in a usability study with a loader crane in which participants made fewer control errors when using a direction-oriented HMI [47]. Since the target-oriented HMI is still under development, there is currently no data available regarding its performance. However, the expected causal relationship between our framework's tools and mental workload as well as the stated empirical evidence lead us to suspect that both human-centered HMIs lead to a considerable reduction in the operator's mental workload.

7 Conclusion and Outlook

Our research aims to develop solutions for remote crane control that are based on a human-centered design perspective in order to increase intuitiveness. To do so, we analyzed which psychological constructs can be assumed to differentiate the effects of the HMIs on the operator's mental workload. Then we established a suitable framework as guidance for comparing the machine-centered and new human-centered HMIs with regard to intuitiveness, mental models, compatibility and the effect of automation on mental workload. For the first time, the framework allows the comparison of different remote crane HMIs based on methodological principles prior to implementation. The presented framework, therefore, forms an important guideline for future developments of crane HMIs.

Subsequent to establishing the framework, we presented two human-centered HMIs for remote control of teleoperated cranes. Comparing the two new HMIs to the traditional machine-centered HMI, using the established framework, leads us to suspect that both HMIs improve remote crane control significantly. The interaction between operator and crane is no longer defined by the functional principle of the machine but by the human's intention in using the system. As the user directly controls the load, his/her actual task of positioning the load is matched closely by the design of the new HMIs. The HMI analysis showed that the user inputs and system responses of the direction-oriented HMI follow the principles of compatibility. This allows operators to form task/action mappings for controlling the crane that are not based on expert knowledge of crane operations but on basic sensorimotor knowledge that is possessed by all people. Consequently, the direction-oriented HMI's potential for facilitating intuitive human-machine interactions for a large and heterogenous group of crane operators was identified. These advantages become particularly obvious in comparison to the original machine-centered HMI, which lacks interface compatibility. For the target-oriented HMI, it was determined that the HMI facilitates intuitive interactions for part of the operator-crane interaction via compatible task/action mappings while removing other interaction responsibilities from the operator by increasing the level of automation. Both the facilitation of intuitive interactions and the increase in automation are expected to reduce the operator's workload.

Hence, we believe firstly that our framework helps in evaluating HMI concepts at an early stage of development. Secondly, based on the framework, we conclude that the proposed human-centered HMIs can be expected to be superior to the existing machine-centered HMI from the operator's point of view, in particular by reducing the operator's mental workload. An empirical assessment of this reduction would provide strong evidence for a more intuitive interaction of the operator with the crane. While this paper establishes the foundation for the suspected effect and deduces a corresponding framework, the empirical evaluation based on a VR study—including measuring the operator's workload using pupil dilatation—will be presented in a subsequent paper.

References

1. Deutsches Institu für Normung e.V.: Cranes; Vocabulary, classification according to type. DIN 15001-1, Berlin (1973)
2. Shapira, A., Lucko, G., Schexnayder, C.: Cranes for building construction projects. J. Constr. Eng. Manage. **133**(9), 690–700 (2007)
3. Miadlicki, K., Pajor, M.: Overview of user interfaces used in load lifting devices. Int. J. Sci. Eng. Res. **6**(9), 1215–1220 (2015)
4. Fang, Y., Cho, Y.K., Durso, F., Seo, J.: Assessment of operator's situation awareness for smart operation of mobile cranes. Autom. Constr. **85**, 65–75 (2018). https://doi.org/10.1016/j.autcon.2017.10.007
5. Löfgren, B.: Kinematic control of redundant knuckle booms with automatic path-following functions. Dissertation. Royal Institute of Technology, Stockholm (2009)
6. Manner, J., Gelin, O., Mörk, A., Englund, M.: Forwarder crane's boom tip control system and beginner-level operators. Silva Fennica **51**(2), 1717 (2017). https://doi.org/10.14214/sf.1717
7. Peng, K.: Interfaces and control systems for intuitive crane control. Master's thesis. Atlanta: Georgia Institute of Technology (2009)
8. Peng, K., Singhose, W.: Crane control using machine vision and wand following. In: 2009 IEEE International Conference on Mechatronics. IEEE Institute of Electrical and Electronics Engineers. IEEE, Piscataway (2009). http://ieeexplore.ieee.org/servlet/opac?punumber=4914928
9. Kazerooni, H., Fairbanks, D., Chen, A., Shin, G.: The magic glove. In: 2004 IEEE International Conference on Robotics and Automation. IEEE Institute of Electrical and Electronics Engineers, pp. 757–763. IEEE, Piscataway (2004)
10. Chi, H.-L., Chen, Y.-C., Kang, S.-C., Hsieh, S.-H.: Development of user interface for tele-operated cranes. Adv. Eng. Inf. **26**(3), 641–652 (2012). https://doi.org/10.1016/j.aei.2012.05.001

11. DeJong, B.P., Colgate, J.E., Peshkin, M.A.: Improving teleoperation: reducing mental rotations and translations. In: 2004 IEEE International Conference on Robotics and Automation. IEEE Institute of Electrical and Electronics Engineers, pp. 3708–3714. IEEE, Piscataway (2004). https://doi.org/10.1109/ROBOT.2004.1308838
12. Kivila, A., Singhose, W.: The effect of operator orientation in crane control. In: Berg, J.M. (ed.) 2014 Proceedings of the ASME 7th Annual Dynamic Systems and Control Conference, pp. 1–7. ASME, New York (2014)
13. Norman, D.A.: The Design of Everyday Things. Basic Books, New York (2013).Revised and expanded
14. Mohs, C., et al.: IUUI – intuitive use of user interfaces. In: Usability Professionals 2006, pp. 130–133 (2006)
15. Blackler, A., Hurtienne, J.: Towards a unified view of intuitive interaction: definitions, models and tools across the world. MMI Interak. **13**, 36–54 (2007)
16. Blackler, A., Popovic, V., Mahar, D.: Investigating users' intuitive interaction with complex artefacts. Appl. Ergon. **41**(1), 72–92 (2010). https://doi.org/10.1016/j.apergo.2009.04.010
17. Naumann, A., et al.: Intuitive use of user interfaces: defining a vague concept. In: Harris, D. (ed.) EPCE 2007. LNCS (LNAI), vol. 4562, pp. 128–136. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73331-7_14
18. Rasmussen, J.: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Trans. Syst. Man Cybern. SMC **13**(3), 257–266 (1983). https://doi.org/10.1109/TSMC.1983.6313160
19. Blackler, A., Popovic, V., Mahar, D.: Designing for intuitive use of products. In: Harada, A. (ed.) 6th Asian Design Conference on Journal of the Asian Design International Conference Tsukuba, pp. 1–16 (2003)
20. Vandenbosch, B., Higgins, C.: Information acquisition and mental models: an investigation into the relationship between behaviour and learning. Inf. Syst. Res. **7**(2), 198–214 (1996). https://doi.org/10.1287/isre.7.2.198
21. Carroll, J.M., Olson, J.R.: Mental Models in Human Computer Interaction. National Academy Press, Washington (1987)
22. Greca, I.M., Moreira, M.A.: Mental models, conceptual models, and modelling. Int. J. Sci. Educ. **22**(1), 1–11 (2000). https://doi.org/10.1080/095006900289976
23. Deutsches Institut für Normung e.V.: Ergonomic principles related to mental workload - Part 2: Design principles. DIN EN ISO 10075-2. Berlin (2000)
24. Kieras, D.E., Bovair, S.: The role of a mental model in learning to operate a device. Cogn. Sci. **8**, 255–273 (1984)
25. Norman, D.A.: Design rules based on analyses of human error. Commun. ACM **26**(4), 254–258 (1983)
26. Young, R.M.: Surrogates and mappings: two kinds of conceptual models for interactive devices. In: Gentner, D., Stevens, A.L. (eds.) Mental Models. Cognitive science. Psychology Press, New York and London (2014)
27. Rasmussen, J.: Mental models and the control of action in complex environments. IEEE Trans. Syst. Man Cybern. **13**(3), 257–266 (1983)
28. Schmidtke, H., Bernotat, R. (eds.): Ergonomie. 3., neubearb. und erw. Aufl. Hanser, München (1993)
29. Fitts, P.M., Seeger, C.M.: S-R compatibility: spatial characteristics of stimulus and response codes. J. Exp. Psychol. **46**(3), 199–210 (1953)
30. Fitts, P.M., Deininger, R.L.: S-R compatibility: correspondence among paired elements within stimulus and response coudes. J. Exp. Psychol. **48**(6), 483–492 (1954)

31. Kornblum, S., Hasbroucq, T., Osman, A.: Dimensional overlap: cognitive basis for stimulus-response compatibility - a model and taxonomy. Psychol. Rev. **97**(2), 253–270 (1990)
32. Proctor, R.W. (ed.): Stimulus Response Compatibility: An Integrated Perspective. Advances in Psychology, vol. 65. North Holland, Amsterdam (1990)
33. Proctor, R.W., Wang, H.: Differentiating types of set-level compatibility. In: Hommel, B., Prinz, W. (eds.) Theoretical Issues in Stimulus-response Compatibility. Advances in Psychology, pp. 11–37. Elsevier, Amsterdam and New York (1997)
34. Shin, Y.K., Proctor, R.W., Capaldi, E.J.: A review of contemporary ideomotor theory. Psychol. Bull. **136**(6), 943–974 (2010). https://doi.org/10.1037/a0020541
35. Kunde, W.: Response-effect compatibility in manual choice reaction tasks. J. Exp. Psychol. **27**(2), 387–394 (2001)
36. Földes, N., Philipp, A.M., Badets, A., Koch, I.: Exploring modality compatibility in the response-effect compatibility paradigm. Adv. Cogn. Psychol. **13**(1), 97–104 (2017). https://doi.org/10.5709/acp-0210-1
37. Worringham, C.J., Beringer, D.B.: Directional stimulus-response compatibility: a test of three alternative principles. Ergonomics **41**(6), 864–880 (1998). https://doi.org/10.1080/001401398186694
38. Endsley, M.R., Kaber, D.B.: Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics **42**(3), 462–492 (1999). https://doi.org/10.1080/001401399185595
39. Wickens, C.D., Hollands, J.G.: Engineering psychology and human performance, 3rd edn. Prentice-Hall, Upper Saddle River (2000)
40. Bainbridge, L.: Forgotten alternatives in skill and work-load. Ergonomics **21**(3), 169–185 (1978). https://doi.org/10.1080/00140137808931711
41. Bainbridge, L.: Ironies of automation. Automatica **19**(6), 775–779 (1983)
42. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**(1), 1–17 (2015). https://doi.org/10.1080/00140139.2014.956151
43. Billings, C.E.: Aviation Automation: The Search for a Human-Centered Approach. Human Factors in Transportation, Lawrence Erlbaum Associates Publishers, Mahwah (1997)
44. Fottner, J., Bengler, K., Top, F., Prasch, L.: Entwicklung eines intuitiven Steuerungskonzepts für Lasthebemaschinen: Forschungsbericht. Technische Universität München, München (2020)
45. Top, F., Wagner, M., Fottner, J.: How to increase crane control usability: an intuitive hmi for remotely operated cranes in industry and construction. In: Karwowski, W., Ahram, T. (eds.) IHSI 2019. AISC, vol. 903, pp. 293–299. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11051-2_45
46. Loveless, N.E.: Direction-of-motion stereotypes: a review. Ergonomics **5**, 357–383 (1962). https://doi.org/10.1080/00140136208930601
47. Top, F., Krottenthaler, J., Fottner, J.: Evaluation of remote crane operation with an intuitive tablet interface and boom tip control. In: IEEE International Conference on Systems, Man and Cybernetics (SMC). Conference Proceedings. IEEE Institute of Electrical and Electronics Engineers, Piscataway, NJ (2020)

Author Index